UCLV
Universidad Central
"Marta Abreu" de Las Villas

FQF
Facultad de
Química y Farmacia

Departamento de Licenciatura en Ciencias Farmacéuticas

TRABAJO DE DIPLOMA

Título: Diseño y aplicación de un flujo de modelación QSPR para la predicción de la biodisponibilidad oral de fármacos y moléculas afines en humanos

Autora: Gabriela Falcón Cano

Tutor: DrC. Miguel Ángel Cabrera Pérez

Santa Clara, junio 2019

UCLV
Universidad Central
"Marta Abreu" de Las Villas

FQF
Facultad de
Química y Farmacia

Departament of Pharmaceutical Sciences

DIPLOMA THESIS

Title: Design and application of a QSPR modelling workflow for prediction of human oral bioavailability of drugs and drug-like

Author: Gabriela Falcón Cano

Thesis Advisor: PhD. Miguel Ángel Cabrera Pérez

Santa Clara, June 2019

“Defender la alegría como una trinchera, defenderla del escándalo y la rutina, de la miseria y los miserables, de las ausencias transitorias y las definitivas”.

Mario Benedetti

Dedicatoria

A Rissette, a mis padres

Agradecimientos

A mis padres, combinación perfecta de sacrificio, sensatez, originalidad y amor.

A mi tutor, por su ayuda incondicional y su entrega a este proyecto.

A mi abuela, Rissette-Paciencia-Control hoy y siempre, la fuente de mi inspiración, mis exigencias y alegrías.

A mi abuelo José Ramón, porque su alegría no se acabe.

A Taya, por su dulzura y su paciencia.

A Pepe, por tener siempre la frase que me impulsa a hacer algo mejor.

A Elisa, por tener siempre una palabra de aliento y de dulzura.

A Jose, por tantas cosas…

A mi hermano Abelito.

A todos los profesores que he tenido durante estos cinco años en la universidad.

A Christophe Molina por su asesoramiento.

A Claudia, Zenia, Tania, Omar y Raidel del CBQ.

A Josue, mi hermanito.

A Irán y Andy por su compañía irremplazable.

A Dianela y Chely, por enseñarme que la familia no solo es "de sangre".

A Raquel e Iván, por tanto cariño.

A los amigos que se fueron por cualquier razón, los amigos de ahora.

Y a todos aquellos que me han apoyado de una forma u otra.

A todos, muchas gracias.

Resumen

Resumen

La determinación exacta de la biodisponibilidad oral es primordial durante el descubrimiento y desarrollo de fármacos. La complejidad y variabilidad de esta propiedad hacen de su predicción una tarea difícil. Los modelos de Relación Cuantitativa Estructura-Propiedad constituyen una estrategia prometedora para la predicción de la biodisponibilidad oral, sin embargo, la mayoría los modelos disponibles han sido desarrollados sobre pequeñas bases de datos públicas con calidad y diversidad estructural cuestionables, utilizando técnicas estadísticas aplicadas de forma individual y sin una proyección a la automatización del proceso de modelación, lo que limita su utilización por cualquier tipo de usuario. En el presente trabajo se desarrolló un flujo continuo de modelación, utilizando KNIME 3.7.1, para la predicción de la biodisponibilidad oral humana, a partir de una base de datos de 1063 fármacos y moléculas afines, utilizándose los algoritmos de clasificación: bosques aleatorios, máquina de soporte vectorial, regresión logística, redes neuronales artificiales, árbol de decisión para clasificación y J48graft, combinados en un voto mayoritario simple y utilizando como variables independientes 3847 descriptores calculados mediante la herramienta “alvaDesc”. Los resultados del voto mayoritario fueron sustancialmente mejores que los de los modelos individuales, obteniéndose una capacidad predictiva mayor que 77%. Para la estimación del dominio de aplicación se utilizaron dos métodos: (i) basado en el propio algoritmo bosques aleatorios y (ii) basado en el *leverage*. La extracción de las variables de ambos modelos permitió establecer relaciones entre los descriptores moleculares y la biodisponibilidad oral.

Abstract

Abstract

Accurate determination of oral bioavailability is paramount during drug discovery and development. The complexity and variability of oral bioavailability make the accuracy prediction a challenge. Quantitative structure-property relationship approaches are promising alternative to the early oral bioavailability prediction, nevertheless the majority of available *in silico* models had been developed on small databases with questionable quality and structural diversity; using individual statistics techniques and machine learning algorithms and without automatic scope; that limits its use by any user. The present work develop a continuous modelling workflow for human oral bioavailability prediction, using KNIME 3.7.1, based on 1063 structurally diverse drug and drug-like molecules. The algorithms random forest, support vector machine, logistic regression, artificial neural networks, decision tree and J48graft were combined in a global vote and using 3847 molecular descriptors from “alvaDesc” as independents variables. Global vote performance was better than individual models, getting predictive ability above 77%. For estimating applicability domain two methods were used: (i) based on random forest algorithm and (ii) leverage approach. The extraction of variables from both models allowed establishing relations between the molecular descriptors implicated and oral bioavailability.

Índice

Introducción

Introducción

La búsqueda de agentes terapéuticos exitosos se ha convertido en una selección ingenieril, donde deben optimizarse múltiples parámetros, entre ellos dos conjuntos de propiedades críticas: (i) la potencia y selectividad del fármaco y (ii) las propiedades farmacocinéticas (ADME) y toxicológicas (1).

En las etapas iniciales de desarrollo de fármacos, las desfavorables propiedades farmacocinéticas son responsables del 10% de los fracasos de compuestos líderes, sin embargo, durante la fase clínica esta cifra asciende a un 16%, lo que manifiesta la necesidad de utilizar métodos experimentales o computacionales que permitan una temprana identificación de compuestos problemáticos. Entre todas las propiedades farmacocinéticas, la biodisponibilidad, es un punto crítico en el diseño y desarrollo de fármacos para la administración oral (2,3).

La determinación exacta de la biodisponibilidad oral humana solo puede realizarse *in vivo,* durante los ensayos clínicos. Sin embargo, diferentes métodos alternativos han sido explorados, entre los que sobresalen: la estimación de la biodisponibilidad oral humana a partir de diversos modelos animales y la combinación de los resultados de ensayos *in vitro*, para la determinación de la permeabilidad (células Caco-2) y el metabolismo en microsomas hepáticos humanos (4).

Dada la alta inversión de tiempo y recursos que suponen los métodos experimentales, la predicción de la biodisponibilidad *in silico* es una alternativa eficaz y que además está en correspondencia con los últimos avances de la química combinatoria y el cribado de alto flujo, lo que supone una poderosa herramienta para utilizar en las etapas tempranas del diseño y desarrollo de fármacos (5,6).

Aunque la aproximación *in silico* presenta importantes ventajas en comparación con las metodologías experimentales, la predicción de la biodisponibilidad es un gran reto y una tarea más compleja en comparación con la predicción de otras propiedades farmacocinéticas. Múltiples factores físico-químicos, fisiológicos y patológicos influyen en esta propiedad como son: el tránsito gastrointestinal, la estabilidad química en el tracto gastrointestinal, la permeabilidad intestinal y el

efecto de primer paso intestinal y hepático. Adicionalmente, es una propiedad muy variable lo que limita la aplicabilidad y fiabilidad de los modelos predictivos (7,8).

La mayoría de los métodos computacionales para la predicción de la biodisponibilidad oral se han basado en los estudios de Relación Cuantitativa Estructura-Propiedad (QSPR), utilizándose para ello diferentes tipos de descriptores moleculares y diversas técnicas estadísticas y de inteligencia artificial (3).

La mayoría de estos modelos *in silico* han sido desarrollados sobre pequeñas bases de datos públicas con calidad y diversidad estructural cuestionables, utilizando técnicas estadísticas y de inteligencia artificial aplicadas de forma individual y sin una proyección a la automatización del proceso de modelación, lo que limita su utilización por cualquier tipo de usuario.

Por todo lo anterior se plantea el Problema Científico siguiente: la **mayoría** de los **modelos in silico** disponibles para la predicción de la biodisponibilidad oral presentan una **cuestionable capacidad predictiva** y **limitada aplicabilidad** en las etapas iniciales de descubrimiento y desarrollo de fármacos.

Como vía para solucionar el Problema Científico se formula la Hipótesis siguiente: la utilización de una amplia base de datos y de métodos estadísticos combinados, bajo un sistema de flujo semiautomatizado, podría mejorar sustancialmente la robustez, capacidad predictiva y aplicabilidad de los modelos QSPR de biodisponibilidad.

Con vistas a contrastar la hipótesis y dar respuesta al problema científico planteado se propone como **objetivo general,** desarrollar un sistema de modelación QSPR, semiautomatizado y abierto, para la predicción efectiva de la biodisponibilidad oral en las etapas iniciales de descubrimiento y desarrollo de fármacos.

Objetivos específicos:

1. Conformar una amplia y heterogénea base de datos de fármacos y moléculas afines con sus correspondientes valores de biodisponibilidad oral, medidos en humanos.

2. Obtener modelos QSPR semiautomatizados, combinando diversos métodos estadísticos y de inteligencia artificial, que permitan una clasificación efectiva de la biodisponibilidad oral.

Marco Teórico Referencial

Capítulo I. Marco Teórico Referencial

Este capítulo presenta los resultados de la revisión bibliográfica realizada, la cual permitió estudiar las bases teóricas y prácticas correspondientes al tema de investigación. Para la presentación de la información se sigue una secuencia lógica que se muestra en la Figura 1.1.

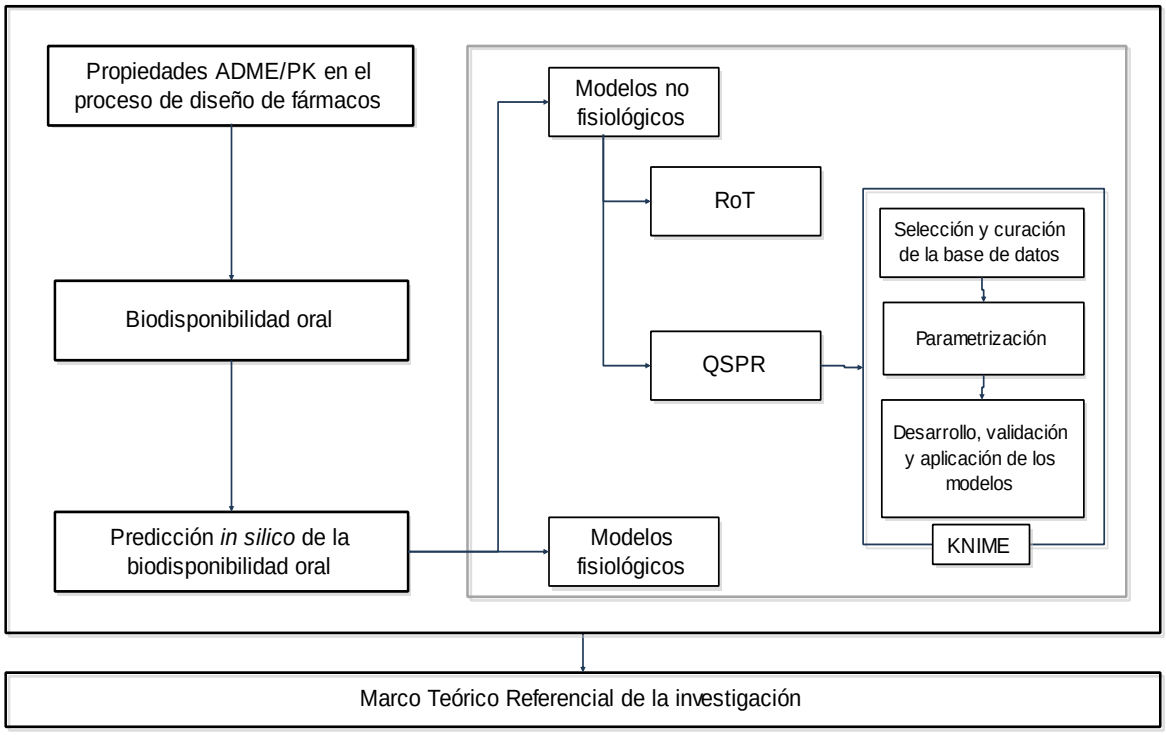


Figura 1.1. Hilo conductor del Marco Teórico Referencial de la investigación.

1.1. Propiedades farmacocinéticas de absorción, distribución, metabolismo y excreción (ADME) en el proceso de diseño de fármacos

El proceso de desarrollo de fármacos transita a través de dos etapas fundamentales: la fase preclínica y la fase clínica. Cada una de ellas agrupa un conjunto de ensayos que permiten la caracterización fisicoquímica, farmacocinética, farmacológica y toxicológica del fármaco. Los resultados de estos ensayos se convierten en criterios para la toma de decisiones relacionadas con el avance del fármaco en el proceso de desarrollo (9).

El cincuenta por ciento de los fracasos de compuestos líderes durante las fases iniciales de desarrollo en el año 2004 se debió a sus desfavorables propiedades ADME (2); y aunque el índice de compuestos fallidos se ha reducido a un diez por

ciento (3), el desarrollo de modelos *in silico* para la predicción de estas propiedades puede incrementar la efectividad, disminuir el costo, y por supuesto, reducir significativamente el número de fallos durante el proceso de desarrollo de fármacos (2). La obtención de modelos confiables permitirá la optimización paralela de la eficacia del compuesto y sus propiedades farmacocinéticas, aumentando su probabilidad de éxito (5,6,10).

1.2. Biodisponibilidad oral

La mayoría de los medicamentos disponibles en el mercado se administran por la vía oral, dada la conveniencia, correspondencia con los procesos fisiológicos y rentabilidad de esta vía de administración. La biodisponibilidad es una de las propiedades farmacocinéticas más evaluadas durante las etapas de desarrollo para cualquier fármaco diseñado para la administración oral (4,8,11).

Un bajo valor de esta propiedad para una nueva molécula candidata a fármaco puede contribuir a una alta variabilidad terapéutica (8,12). El incremento de esta variabilidad puede derivar en concentraciones plasmáticas fuera del intervalo terapéutico (concentraciones tóxicas o subterapéuticas), y esto es particularmente importante para fármacos con estrecho margen terapéutico o con potencial para generar resistencia (8).

Según la Agencia de Medicamentos Europea (EMA, por sus siglas en inglés) la biodisponibilidad oral es definida como la velocidad y la extensión a la cual un ingrediente activo es absorbido a partir de la forma farmacéutica que lo contiene, y accede a la circulación sistémica (3).

La biodisponibilidad es una propiedad muy compleja y variable. Como se puede observar en la Figura 1.2 la solubilidad, la permeabilidad y el metabolismo son los principales factores que contribuyen al valor de esta propiedad. Por lo tanto se puede definir la misma como el resultado de tres componentes: la fracción oral absorbida (Fa), la biodisponibilidad intestinal (Fi) y la biodisponibilidad hepática (Fh) (7,13).

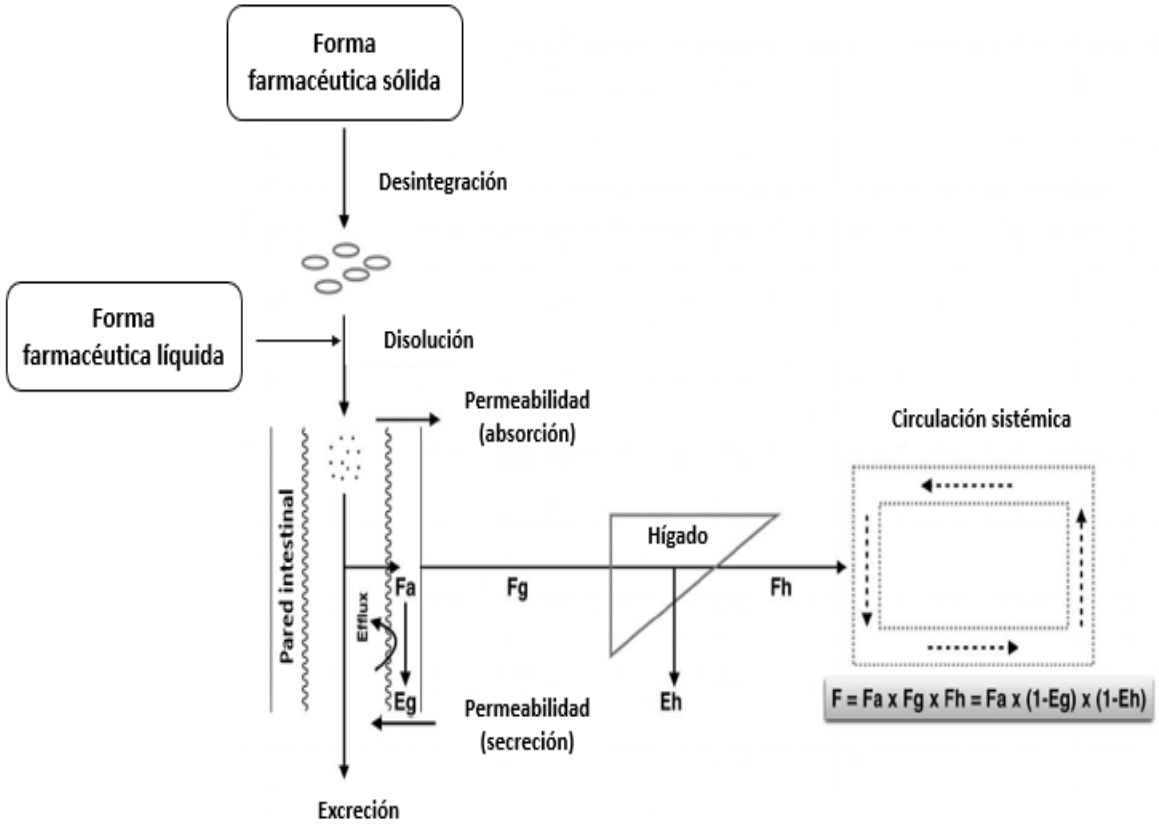


Figura 1.2. Representación general del proceso de acceso del fármaco a circulación sistémica. Principales procesos involucrados en la biodisponibilidad oral. Fa: fracción de la dosis oral absorbida en el tracto gastrointestinal; Fg: Biodisponibilidad intestinal; Eg: extracción intestinal del fármaco; Fh: Biodisponibilidad hepática; Eh: Extracción hepática del fármaco y F: Biodisponibilidad sistémica.

La fracción de la dosis oral administrada que es absorbida en el tracto gastrointestinal (Fa) está determinada por propiedades biológicas y fisicoquímicas como la velocidad de disolución, la solubilidad acuosa y la permeabilidad (Figura 1.2) (3).

La biodisponibilidad intestinal (Fg) es la fracción de la dosis oral que escapa al metabolismo de primer paso intestinal, los mecanismos de eflujo y la degradación luminal (4,11,13).

La biodisponibilidad hepática (Fh) es la fracción del fármaco que escapa el metabolismo de primer paso hepático (Eh) y la secreción biliar (4).

La biodisponibilidad sistémica (F) es el producto de la fracción de la dosis oral absorbida y las fracciones de fármaco que escapan del metabolismo de primer paso intestinal y hepático. En términos matemáticos puede ser descrita según la ecuación 1.1 (11).

$$F = fa * fi * fh \quad (1.1.)$$

En términos farmacocinéticos la biodisponibilidad es descrita por dos parámetros: la fracción de la dosis administrada que accede a la circulación sistémica (F) y la velocidad a la cual ocurre dicho acceso (11).

La fracción de la dosis administrada que alcanza la circulación sistémica es calculada como se plantea en la ecuación 1.2 (11).

$$F = \frac{M}{Dosis} \quad (1.2.)$$

Donde, M es la cantidad de fármaco que accede a la circulación sistémica, y su valor es obtenido de la ecuación 1.3.

$$M = Cl * ABC \quad (1.3.)$$

En la ecuación 1.3, Cl es el aclaramiento total plasmático (la fracción del volumen de distribución que es depurado de fármaco por unidad de tiempo) y ABC es el Área Bajo la Curva.

Sustituyendo M en la ecuación 1.3 resulta la ecuación 1.4 (11):

$$F = \frac{Cl*ABC}{Dosis} \quad (1.4.)$$

El componente velocidad de la biodisponibilidad es estimado por dos parámetros: la concentración plasmática máxima (Cmáx) y el tiempo en el cual se alcanza esta concentración (Tmáx). Cmáx está relacionada con el aclaramiento plasmático, la fracción de dosis no metabolizada que alcanza la circulación sistémica, la velocidad de absorción y las velocidades de distribución y eliminación (4). Tmáx depende de la velocidad de absorción y las velocidades de distribución y eliminación (11).

La determinación de la biodisponibilidad oral absoluta en magnitud considera la administración del fármaco por dos vías: oral e intravenosa, esta última, tomada como referencia, dado que por esta vía se asume que el 100% de la dosis administrada accede inalterada a la circulación sistémica (14). Cuando no es posible el uso de la vía intravenosa, se debe utilizar otra administración extravasal como estándar de referencia, y en este caso se estaría determinando la biodisponibilidad relativa (4).

La determinación de la biodisponibilidad en velocidad es más compleja, debido a la dificultad que representa caracterizar y cuantificar el proceso de la absorción de los fármacos, determinado por variables fisiológicas y patológicas que hacen muy difícil su tratamiento matemático (7). Para ello se deben considerar parámetros puntuales (Tmáx y Cmáx), parámetros compartimentales, no compartimentales y el grado de absorción (11).

1.3. Modelos *in silico* para la predicción de la biodisponibilidad oral

La biodisponibilidad oral humana se determina frecuentemente a partir de las mediciones de biodisponibilidad oral en varias especies animales (rata, ratón, conejo, perro, primates no humanos, etc.). Sin embargo, las diferencias fisiológicas, metabólicas y de trasportadores existentes entre estos modelos, hacen que la medición directa de la biodisponibilidad en animales no sea un modelo sustitutivo a la estimación en humanos y los resultados alcanzados sean altamente variables (15,16).

Otro de los métodos descritos para la estimación de la biodisponibilidad oral han estado basados en la utilización de datos de permeabilidad *in vitro* en líneas celulares y en ensayos *in vitro* en microsomas hepáticos humanos para la estimación del metabolismo para la evaluación del metabolismo (14,15).

En comparación con otras propiedades ADME, la predicción *in silico* de la biodisponibilidad oral es una tarea compleja porque el comportamiento de esta propiedad farmacocinética depende de numerosos factores fisicoquímicos,

fisiológicos y patológicos (15).Un resumen de los mismos se muestran en la Tabla 1.1.

Tabla 1.1. Principales factores influyentes en la biodisponibilidad oral.

Factores que afectan la biodisponibilidad oral de fármacos			
Fisiológicos	**Fisicoquímicos**	**Tecnológicos**	**Otros**
pH gastrointestinal	Estabilidad del fármaco en el fluido GI	Velocidad de disolución	Interacciones farmacocinéticas
Vaciado estomacal	Constante de ionización	Mecanismos de liberación	Estados patológicos
Motilidad intestinal	Lipofilia	Efectos de excipientes	
Período de tránsito intestinal	Solubilidad en fluido GI		
Permeabilidad	Cristalización		
Microflora intestinal	Disolución		
Secreción intestinal	Tamaño de partícula		
Efflux	Polimorfismo		
Metabolismo luminal/hepático	Formación de complejos		
Edad	Adsorción		
Género	Carga		
Capacidad de transporte	Difusión		

La variabilidad interindividual e incluso intraindividual de la biodisponibilidad oral (si es determinada en diferentes tiempos) explica por qué las mediciones de esta propiedad presentan un índice tan elevado de error. Según Wang y colaboradores el error promedio no explicado de esta propiedad en un estudio realizado a 367 fármacos fue de 12,1% (15).

Las principales estrategias actualmente usadas para la modelación computacional de la biodisponibilidad oral son:

- los modelos de relación estructura-propiedad (QSPR),
- las reglas simples (*RoT*: *rule-of-thumb*) y,
- los modelos basados en características fisiológicas (PBPK) (3).

Dentro de los principales modelos no fisiológicos utilizados para la predicción de biodisponibilidad oral se encuentran los basados en QSPR y RoT. Los primeros

desarrollan modelos matemáticos que relacionan los descriptores moleculares (información de la estructura) y los valores experimentales de biodisponibilidad oral. Mientras que, los métodos de RoT relacionan las propiedades fisicoquímicas con los valores de biodisponibilidad mediante reglas simples (3,17).

Varias reglas han sido desarrolladas con vistas a definir valores límites de propiedades fisicoquímicas que permitan clasificar un compuesto según su valor de biodisponibilidad.

Veber y colaboradores analizaron más de 1100 compuestos candidatos a fármacos; y propusieron una regla basada en dos propiedades: para que una molécula mostrara un valor de biodisponibilidad oral en ratas mayor o igual que 20%, debía tener menos de diez enlaces simples y un área superficial polar (PSA, por sus siglas en inglés) menor que 140 $Å^2$ (o la suma de hidrógenos aceptores y donantes mayor o igual que 12) (18).

Por su parte, Martin y colaboradores desarrollaron un sistema de puntuación que asignaba la probabilidad de que un compuesto pudiera mostrar un valor de biodisponibilidad sistémica en ratas mayor o igual que 10% (19,20). Los autores plantearon que ni la regla-de-los-cinco, los coeficientes de partición y distribución, ni tampoco la combinación del número de enlaces simples y el área superficial polar clasificaban suficientemente los compuestos, de hecho, diferentes propiedades influían en la biodisponibilidad de los compuestos en dependencia de la carga molecular predominante a un determinado pH biológico (19).

La principal ventaja de estas reglas simples basadas en varias propiedades moleculares es su fácil interpretación y comprensión. Sin embargo, no son capaces de caracterizar procesos más complicados que influyen en la biodisponibilidad, como es el metabolismo mediado por enzimas de la familia del citocromo P450 por lo que la exactitud de la predicción es cuestionable (18).

En contraste con estas reglas basadas en propiedades moleculares, desde el año 2000 numerosos modelos QSPR han sido desarrollados para predecir la biodisponibilidad de fármacos en humanos y roedores, utilizando diversas técnicas

estadísticas y de inteligencia artificial (3,20,21). Un resumen de los principales modelos desarrollados se presenta en la Tabla 1.2.

Claramente, la baja exactitud y la elevada variabilidad de los modelos desarrollados están en concordancia con lo compleja que es la propiedad que se modela (3,15). Una de las alternativas que se han propuesto para mejorar la capacidad predictiva de los modelos es el uso de propiedades vinculadas a la biodisponibilidad como la absorción intestinal en humanos, el volumen de distribución y la velocidad de eliminación (determinados en animales) como descriptores del modelo (22,23).

Por ejemplo, Wang y colaboradores con el objetivo de incrementar la robustez y fiabilidad del modelo, particionaron los datos iniciales (776 compuestos) en cuatro subseries que agrupaban los compuestos según sus propiedades ADME, y luego construyeron modelos de regresión lineal para cada uno de ellos (23). Imawaka y colaboradores, propusieron el cálculo directo de la biodisponibilidad, utilizando como variables datos farmacocinéticos en animales. Sin embargo, la desviación promedio de las predicciones con respecto a los valores experimentales fue de un 20.9% (22).

En cuanto a los modelos fisiológicos, la modelación farmacocinética basada en la fisiología (PBPK: *physiology-based pharmacokinetic*, en inglés) integra el conocimiento de los procesos fisiológicos y las propiedades fisicoquímicas de los compuestos para la predicción y simulación de complejas propiedades biológicas y proporcionan descripciones más realistas de la absorción, la distribución, el metabolismo y la eliminación (24). Su funcionamiento se basa en la descripción matemática de los procesos físicos y fisiológicos y consiste básicamente en la subdivisión del organismo en órganos individuales, describiendo la disposición del fármaco en cada uno de los compartimentos (25,26).

En este sentido, los parámetros de entrada al modelo pueden ser datos *in silico*, *in vitro* e *in vivo* que se combinan para la predicción de las concentraciones plasmáticas y tisulares del fármaco con el objetivo de construir el perfil de concentración *vs* tiempo (3).

Tabla 1.2. Principales modelos QSPR desarrollados para la predicción de la biodisponibilidad**.**

#	Métodos	Variables	Resultados	Referencias
1	RLM	85 descriptores estructurales	N=591; Validación cruzada 80/20: R^2=0.58	(30)
2	Regresión SW		$N_{(entr)}$:159; R^2= 0.35; Q^2(LOO)=0.25	(31)
3	RBF/ANN-regresión	Análisis de sensibilidad:10-31-1(topológicos)	$N_{(entr)}$ =137; R^2= 0.736; N_{prueba}= 15 R^2=0.89; $N_{externa}$=15; R^2=0.68	(32)
4	GA/MLR-regresión	42 grupos funcionales-2 DMs: *Molecular refractivity* y *Rule-of-five*	N=577; R^2 =0.55, RMSE = 21,9%;Validación cruzada 90/10 R^2=0.42; RMSE =24.6%	(33)
5	HQSAR-regresión	8 componentes	$N_{(entr)}$=250, Q^2=0.70, R^2=0.93; Next=52; R^2=0.85; Q^2(LOO)=0.70	(34)
6	RLM/GFA	7 propiedades moleculares 10-130 DMs	$N_{(entr)}$=916, R=0.38; N_{prueba}=80; Q (leave-one-out)=0.68 (60 DMs);$R_{consenso}$=0.71	(20)
7	GFA/RLM	Propiedades moleculares y *fingerprints*	N=996, R=0.79, Q (leave-one-out) =0.72, SMRE=22.3%; R_{prueba}=0.71, $RMSE_{prueba}$=23.6%.	(18)
8	SVM-regresión	Descriptores Dragon	$N_{(entr)}$=197, R^2=0.69; SMRE=42%; N_{prueba}=43, RMSE=46% ;Q^2=0.68	(35)
9	GA/SW/AFP-clasificación	10 DMs	N=272; R^2=0.75; N=432; R^2=0.64	(36)
10	GA-CG-SVM-clasificación	25 descriptores Cerius2	N= 690;(5 validaciones cruzadas) Q^2=0.80; Clase de baja biodisponibilidad: Exactitud:0.25	(37)
11	ORMUCS-clasificación	logP, logD, $logD_{6.5} - logD_{7.4}$,15 descriptores estructurales	$N_{(entr)}$:232; R^2=0.71; Q^2 (LOO)=0.67; $N_{externa}$: 40; R^2=0.60	(38)
12	Clasificación Logística	47 descriptores	N= 969, Q^2= 0.71; Exactitud 0.71-0.75	(39)
13	RF, SVM, kNN, CASE ULTRA	1597 descriptores (Dragon) 186 (MOE)	N=995, $R^2_{consenso}$ =0.28, Q^2=0.76	(40)

RLM: Regresión Lineal Múltiple; SW: *StepWise*; RBF (*Radial Basis Function*): función de base radial; GFA: algoritmo de función genética; GA: algoritmo genético; HQSAR: holograma Relación-Cuantitativa-Estructura-Actividad; SVM (*Support Vector Machine*): máquina de soporte vectorial; AFP (*adaptive fuzzy partition*); CG (*conjugated gradient*): gradiente conjugado; ORMUCS (*Ordered Multicategorical Classification Method Using the Simplex Technique*); RF (Random Forest) : bosque aleatorio; kNN (*k-Nearest Neighbor*): k-vecino más cercano; DMs: Descriptores Moleculares; LOO(*Leave-One-Out*): dejando uno fuera.N(entr):serie de entrenamiento.

Los modelos PBPK han sido ampliamente aplicados durante las fases de descubrimiento y desarrollo de fármacos y recientemente se ha incrementado su uso en la predicción de la biodisponibilidad oral y sus componentes (27).

El primer intento para la predicción de la absorción oral usando un modelo PBPK fue descrito por Yu y Amidon, quienes desarrollaron un modelo compartimental del sitio de absorción y el tránsito gastrointestinal (CAT, por sus siglas en inglés) para la predicción de la fracción oral absorbida de diferentes fármacos (3,28).

Su versión más avanzada, el modelo ACAT considera la velocidad de disolución, la absorción en el estómago y el colon, el metabolismo de primer paso y la degradación del fármaco para la simulación de la fracción oral absorbida. Este modelo ha sido aplicado para la predicción de diferentes aspectos de la absorción de fármacos y la biodisponibilidad (28).

Recientemente, Paixao y colaboradores emplearon una variación del modelo ACAT, proponiendo la combinación de datos *in vitro* de permeabilidad y aclaramiento con datos *in silico*. El estudio demostró que los resultados mejoran cuando los datos *in vitro* eran incorporados como parámetros de entrada al modelo (29).

Dada la importancia que representan los modelos QSPR para el desarrollo de este trabajo, en los dos epígrafes siguientes se profundiza en sus principios regulatorios, y específicamente en el protocolo estándar seguido para la predicción de la biodisponibilidad oral mediante estos modelos.

1.4. Modelos no fisiológicos QSPR para la predicción de la biodisponibilidad oral

Los modelos QSPR han sido los más explorados para la predicción de la biodisponibilidad, no obstante, la confiabilidad de estos modelos es aún un problema sin resolver. Los modelos de regresión han sido más explorados que los de clasificación, pero la complejidad del parámetro y su variabilidad apuntan a que la implementación de modelos de clasificación puede brindar mejores resultados, porque implica el establecimiento de un valor límite que es usado para la división de los compuestos en alta/baja biodisponibilidad, lo que minimiza los errores del uso

de un valor concreto de la propiedad. Además, sería aconsejable la combinación de múltiples modelos en orden de mejorar la confiabilidad de las predicciones y en este sentido, varias técnicas robustas de clasificación pueden ser evaluadas y fusionadas para construir un modelo consenso o ensamblado.

1.4.1. Principios regulatorios de los modelos de Relación Cuantitativa Estructura- Propiedad

De acuerdo a los principios de la Organización para el Desarrollo y Cooperación Económica (OECD) (41), los modelos de relación estructura-propiedad deben cumplir con:

1. *Un punto de medición definido*: la selección de la propiedad físico-química objeto de predicción debe ser trasparente e idealmente homogénea, este último requisito raras veces puede cumplirse porque los datos experimentales con los cuales se trabaja provienen de diversas fuentes y diversos protocolos (41,42).
2. *Un algoritmo no ambiguo*: resulta necesario la descripción detallada de los procedimientos utilizados para la selección de la base de datos y de los valores experimentales, los métodos de cálculo de descriptores y fragmentos así como para la selección de variables, la construcción y evaluación de los modelos (41).
3. *Un dominio de aplicación definido*: este punto está relacionado con los alcances y limitaciones del modelo. El dominio de aplicación es aquella región teórica en el espacio químico, definida por los compuestos de la serie de entrenamiento y representada en cada modelo por descriptores moleculares específicos y por la respuesta modelada. Solamente las predicciones de los compuestos que están dentro del dominio de aplicación pueden ser consideradas como confiables (41,43).
4. *Medidas apropiadas de bondad de ajuste, robustez y predictibilidad*: este principio se asocia con la necesidad de una validación estadística para establecer el desempeño del modelo, en la práctica esto se aplica mediante las validaciones interna y externa del modelo (41,44,45).
5. *Una interpretación mecanicista, si es posible*: una interpretación mecanicista se asocia directamente con la cantidad y naturaleza de los descriptores y su relación con la propiedad predicha (41).

1.5. Protocolo estándar para la modelación QSPR de la biodisponibilidad oral

Los modelos no fisiológicos de la biodisponibilidad oral basados en estudios QSPR, generalmente siguen un protocolo estándar: (i) selección y curación de la base de datos experimental, (ii) parametrización de la estructura molecular y (iii) construcción y validación del modelo estadístico (3).

1.5.1. Selección y curación de la base de datos experimentales

El acceso a datos experimentales de biodisponibilidad oral es probablemente la mayor barrera en el desarrollo de modelos *in silico* robustos y confiables (3). Existen publicadas varias bases de datos de biodisponibilidad oral en humanos, pero algunas veces no poseen la calidad y diversidad estructural requeridas (15). La mayoría de ellas consisten en la recopilación de valores descritos en la literatura, y no de datos controlados experimentalmente. Usualmente, los valores de biodisponibilidad oral *in vivo* son colectados en los ensayos clínicos, pero estos datos son significativamente variables de una fuente a otra y se pueden encontrar parcializados hacia los compuestos de alta o moderada biodisponibilidad, lo que limita la capacidad predictiva del modelo *in silico* (15,21).

En el año 2007 Hu y colaboradores publicaron una base de datos de 768 compuestos, que recopilaba información experimental de 185 artículos. Posteriormente, la base de datos se extendió hasta 1013 compuestos, incluyendo fármacos estructuralmente diversos y moléculas similares. Esta es una de las bases de datos más citadas y constituye una buena fuente para el desarrollo de modelos predictivos más fiables (3).

Además de la selección de una base de datos experimental fiable, es preciso la curación de la misma. Este proceso debe organizarse en varios pasos, de manera que permita la identificación y corrección de errores estructurales, en ocasiones, a expensas de remover algunas estructuras incompletas o incongruentes (42).

El procedimiento general de curación debe incluir: remoción de compuestos inorgánicos, organometálicos, sales y mezclas para los cuales la mayoría de los

programas para la generación de descriptores no están bien diseñados. Además, limpieza de las estructuras (detección de valencias inusuales), aromatización de anillos, normalización de grupos funcionales específicos y estandarización de tautómeros (42). Finalmente deben ser removidos los compuestos duplicados, resultados de la curación, estandarización y normalización y debe realizarse una inspección manual de los casos complejos (42).

1.5.2. Parametrización de la estructura molecular

Los descriptores moleculares son el resultado de la transformación de la información química de una molécula en números, mediante procedimientos lógico-matemáticos (46–48). En la predicción de la biodisponibilidad los descriptores 1D, 2D y 3D han sido los más utilizados para la obtención de modelos (3).

Una de las propiedades más importantes de un modelo debe ser la interpretación mecanicista que comienza con el número y la naturaleza de los descriptores empleados (3). Los descriptores 1D permiten una fácil interpretación, proporcionando información acerca de átomos, enlaces, composición y peso molecular. La mayoría de los modelos construidos con este tipo de descriptores ha tenido una limitada robustez y capacidad predictiva. No obstante, han sido usados para la estimación de la biodisponibilidad a través de reglas basadas en propiedades moleculares (20,21).

La representación 2D de una molécula, comúnmente referida como su representación topológica, define la conectividad de los átomos en términos de la presencia y naturaleza de los enlaces químicos. Tal representación permite la definición de los descriptores moleculares 2D. Las principales ventajas de estos parámetros son que: (i) contienen información simple y útil acerca de la estructura molecular, (ii) son invariantes en cuanto a la rotación de la molécula, y (iii) pueden ser calculados sin previa optimización de la geometría molecular (42). En general, los descriptores 2D son una buena alternativa para el desarrollo de modelos predictivos de la biodisponibilidad oral debido fundamentalmente a que pueden ser calculados rápidamente para grandes bases de datos, y a que permiten generalmente una interpretación viable (17).

Los descriptores 3D proporcionan más información estructural, pero el principal problema para su determinación es la elección de la conformación de la molécula (49). Varios descriptores 3D han sido utilizados para la predicción de la biodisponibilidad, pero su selección está influenciada por la complejidad y el tiempo de cálculo (20).

1.5.3. Métodos estadísticos y de inteligencia artificial para el desarrollo de modelos predictivos QSPR

Existen un gran número de técnicas estadísticas y de inteligencia artificial disponibles para establecer una relación entre las variables independientes (descriptores) calculadas para los objetos (estructuras) y la variable dependiente (valores experimentales de biodisponibilidad oral) (50). Estas técnicas abarcan desde la regresión lineal múltiple hasta complejos algoritmos de aprendizaje automático (machine learning) de clasificación o regresión que permiten la construcción de modelos individuales o múltiples (por consenso y modelos ensamblados) (14,51).

Como se planteó en el epígrafe 1.4 los algoritmos de clasificación suponen una buena alternativa para el desarrollo de modelos predictivos de biodisponibilidad oral porque permiten la asignación de una clase determinada a cada uno de los objetos, mediante la determinación de reglas de clasificación, límites de clase y probabilidades de pertenencia a una clase. Para la construcción de las reglas de pertenencia a clases se utilizan los compuestos de la serie de entrenamiento, y en función de estas es que ocurre la clasificación de los compuestos desconocidos (aquellos compuestos que no fueron usados para la construcción del modelo) (52).

A continuación, se presentan cuatro de las técnicas que pueden y han sido usadas para la solución de problemas de clasificación:

1. **Regresión logística.** La regresión logística se emplea para modelar la probabilidad de la ocurrencia de algún evento como una función lineal de una serie de variables predictoras. Por ejemplo, en un problema de clasificación binaria, la relación entre una propiedad (que puede asumir dos valores: alto/bajo) y una serie de descriptores moleculares. Dado un compuesto

desconocido, el método de regresión logística calcula la probabilidad de que un compuesto pertenezca a una de las dos clases (53).

2. **Máquina de soporte vectorial.** La máquina de soporte vectorial (SVM, por sus siglas en inglés) es probablemente uno de los más conocidos métodos de Kernel para el desarrollo de modelos (51). El algoritmo implementa el principio inductivo de minimización de riesgos estructurales, de la teoría del aprendizaje estadístico, para obtener una buena generalización sobre un número limitado de patrones de aprendizaje. La minimización del riesgo estructural implica un intento simultáneo de minimizar el riesgo empírico y la dimensión Vapnik-Chervonenkis (54). En un problema de clasificación binaria, SVM trata de construir un hiperplano capaz de discriminar entre las dos clases de los objetos en análisis (50).

   Es el clasificador menos afectado por datos duplicados (*overlapping*) y el de menor riesgo de sobre-aprendizaje, además puede trabajar con datos complejos y de alta dimensión y es capaz de obtener resultados robustos, incluso con datos redundantes. Otras ventajas son su relativa simplicidad de uso, porque el usuario solo necesita definir algunos parámetros y la capacidad de obtener resultados estables y reproducibles. Sus desventajas son su difícil interpretación, y la dificultad para extraer las variables implicadas en los modelos (55,56).

3. **Árboles de decisión y bosques aleatorios.** Los árboles de decisión y su extensión bosques aleatorios (*random forest*, en inglés) son algoritmos de aprendizaje automático robustos y fáciles de interpretar. Su principal ventaja es la facilidad con la que se pueden ver las variables que contribuyen a la clasificación y su importancia relativa en función de su ubicación en el árbol ya sea en tareas de regresión o clasificación (49,57).

   Un árbol de decisión es una estructura con disposición jerárquica de nodos y ramas, en la cual se ubican tres tipos de nodos: un nodo raíz, nodos internos, y nodos hojas. La base de la clasificación de un compuesto desconocido es un nodo hoja, luego que se verifiquen una serie de preguntas (nodos) y respuestas empezando por el nodo raíz (49).

Los árboles de decisión son fácilmente interpretables si son pequeños, y presentan la ventaja de que sus resultados no son afectados por la presencia de descriptores innecesarios. Su principal inconveniente es que son susceptibles al sobre-aprendizaje debido a la carencia de datos o la presencia de objetos mal clasificados en la serie de entrenamiento (49).

Para la solución de este problema métodos como la poda del árbol (pruning, en inglés), la validación cruzada y los bosques aleatorios (*random forest*) pueden ser usados (49).

El método de poda (*pruning*) previene la construcción de árboles excesivamente complicados mientras que el random forest utiliza una clasificación consenso para reducir el problema del sobre-aprendizaje y al mismo tiempo incrementa la precisión y exactitud del modelo (49).

El algoritmo random forest implementa muchos árboles de decisión que son colectivamente llamados como "bosque" (*forest*) y realiza la predicción final basado en un voto mayoritario que utiliza las predicciones realizadas por cada árbol (57,58).

Random forest puede manipular un elevado número de descriptores y objetos. Junto a la clasificación de un compuesto desconocido, el método puede ser extendido para realizar un análisis de conglomerados no supervisado y para la detección de compuestos outliers (58). También puede ser usado para inferir la influencia de los descriptores en una tarea de clasificación y también permite la estimación de valores perdidos (58,59).

El algoritmo es menos afectado por datos "ruidosos" y poco convincentes, y no es muy propenso al sobre-aprendizaje. Sin embargo, sus resultados pueden estar influenciados por datos desbalanceados, muestras pequeñas, el número de árboles y las características seleccionadas (58).

4. **Redes neuronales artificiales.** Tras la descripción inicial, en 1943, de una neurona formal, las primeras redes neuronales aparecieron en 1958 con el "perceptrón" de Rosenblatt. Se desarrollaron rápidamente en la década de 1980 y se han utilizado ampliamente en la industria desde la década de 1990 (60). Una red neuronal tiene una arquitectura basada en la del cerebro, organizada

en neuronas y sinapsis, y toma la forma de un conjunto de unidades interconectadas (o neuronas formales), donde a cada variable de entrada continua corresponde una unidad de primer nivel, llamada capa de entrada, y cada categoría de una variable cualitativa también corresponde a una unidad de la capa de entrada. Una unidad recibe valores en su entrada y devuelve de 0 a n valores en la salida.

Una función de combinación calcula un primer valor a partir de las unidades conectadas a la entrada y el peso de las conexiones (49). Para determinar un valor de salida, se aplica una segunda función, llamada función de transferencia (o función de activación), a este valor. Las unidades de la capa de entrada son simples, en el sentido de que no crean ninguna combinación sino que sólo transmiten los valores de las variables que les corresponden (61).

El aprendizaje de la red neuronal se realiza a partir de una muestra de la población estudiada; utiliza a los individuos de la muestra para ajustar el peso de las conexiones entre las unidades. En el curso del aprendizaje, el valor entregado por la unidad de salida se compara con el valor real, y los pesos p_i de todas las unidades se ajustan para mejorar la predicción, mediante un mecanismo que depende del tipo de red neuronal (62)(60).

Las redes multicapas basadas en el algoritmo de retropropagación de errores, conocidas por sus siglas en inglés BNN-MLP (*Backpropagation Neural Network Multilayer Perceptron*), constituyen una de las redes neuronales artificiales más exploradas. Su principal ventaja es su capacidad genérica de mapeo de patrones y su mayor desventaja es la dificultad de su interpretación.

1.6. Konstanz Information Miner (KNIME) para el desarrollo, validación y aplicación de modelos QSPR

La minería de datos (*data mining*) integra todos los procedimientos de análisis que son requeridos para la preparación de los datos, la modelación, la evaluación de los modelos y su interpretación de la manera más eficiente (55,63). Desde hace más de veinte años se han desarrollado herramientas de software libres y públicas con vistas a facilitar el proceso de análisis de datos y ofrecer a los investigadores una alternativa a las plataformas comerciales disponibles (64).

KNIME se ha convertido en una de las plataformas analíticas líderes para la innovación, el descubrimiento de la naturaleza oculta de los datos y la predicción de nuevas características (65). Está codificado en Java, y basado en la plataforma Eclipse e integra varios componentes de *machine learning* (algoritmos para problemas de clasificación regresión y *clustering*) y minería de datos (lectura, procesamiento, análisis y visualización). Todo lo anterior se hace posible mediante la implementación de nodos que están interconectados y que presentan puertos de entrada y salida específicos, en dependencia de su funcionalidad (66). La secuencia de nodos interconectados crea un flujo de trabajo a través del cual fluye la información (proceso conocido como programación visual) (67).

KNIME permite la integración de nodos especiales que han sido desarrollados para ser usados en la química, la biología y el diseño de fármacos (2). A continuación, se mencionan algunos de estos paquetes de nodos y se adjunta una breve descripción:

1. "RD-kit" (67): proveen robustas funcionalidades que permiten la conversión de moléculas en distintos formatos (SMILES, RD-kit o SDF), la generación de SMILES canónicos, el filtrado de sales, la generación de coordenadas 2D y 3D, el cálculo de descriptores y *fingerprints*.
2. "alvaDesc": permite el cálculo de un amplio rango de descriptores moleculares y *fingerprints*. Además facilita la primera exploración de los datos mediante los servicios de PubChem y herramientas para el Análisis de Componentes Principales.
3. "Enalos" (68): permite la determinación del dominio de aplicación basado en la Distancia Euclidiana y basado en el *Leverage*, el criterio de aceptabilidad de un modelo de regresión QSAR/QSPR y además el cálculo de descriptores con el nodo Enalos Mol2.
4. "Waikato Environment for Knowledge Analysis"(WEKA) (67): ofrece varios nodos adicionales para el desarrollo de algoritmos de clasificación, regresión y *clustering*.

Recientemente se han reportado estudios que utilizan KNIME para la obtención de modelos predictivos ADME:

1. Zakharov y colaboradores propusieron modelos QSPR para la predicción de la estabilidad metabólica de fármacos a partir de datos *in vitro* del tiempo de vida media medidos en microsomas hepáticos humanos. Para la obtención de los modelos se desarrollaron diferentes técnicas: *K*-vecinos más cercanos, regresión logística, perceptrón multicapa y redes bayesianas, todas ellas implementadas en KNIME (69).
2. En el 2016, Legehar y colaboradores desarrollaron una base de datos (IDAAPM) que integraba información regulatoria, efectos adversos, datos de propiedades ADME, toxicológicos, las estructuras químicas, las dianas terapéuticas y los descriptores moleculares de más de 30 000 productos aprobados por la Administración de Alimentos y Fármacos de los Estados Unidos (FDA: *Food and Drug Administration,* FDA). La base de datos fue desarrollada en KNIME, plataforma que permitió el acceso a las fuentes de la FDA, la transformación de los datos, el análisis y la construcción de modelos predictivos (70).
3. Trapotsi, en el año 2017 publicó un estudio que consistía en la evaluación y desarrollo de modelos predictivos ADME, que incluía la determinación del dominio de aplicación de todos los modelos con el uso de diferentes métricas. Todo esto integrado en KNIME (2).

Por lo tanto, KNIME ofrece la gran ventaja de facilitar la integración de todos los pasos de los estudios QSPR en un único flujo de trabajo (71), por lo que sería una alternativa eficiente para la curación de la base de datos, la parametrización de las estructuras moleculares, el desarrollo, validación y aplicación de modelos predictivos de biodisponibilidad oral.

1.7. Conclusiones parciales

1. La biodisponibilidad oral es uno de los puntos críticos durante el diseño y optimización de fármacos previstos para la administración oral.
2. La predicción de la biodisponibilidad oral mediante modelos *in silico* es una tarea compleja porque el comportamiento de esta propiedad fisiológica está en

función de numerosos factores biológicos, fisicoquímicos y tecnológicos. Adicionalmente también es afectada por estados patológicos del individuo y por la ocurrencia de interacciones farmacocinéticas con otros fármacos o con alimentos.

3. Las principales estrategias actualmente usadas para la modelación computacional de la biodisponibilidad oral en humanos son: los estudios QSPR, las reglas simples (*rule-of-thumb*) y la modelación farmacocinética basada en la fisiología (PBPK).
4. La confiabilidad de los modelos computacionales para la predicción de la biodisponibilidad es aún un problema sin resolver.
5. El acceso a datos experimentales de biodisponibilidad oral es probablemente la mayor barrera en el desarrollo de modelos *in silico* robustos y confiables. También influyen la selección de las variables del modelo (descriptores moleculares y/o fragmentos) y los métodos estadísticos y de inteligencia artificial usados para la modelación.
6. La combinación de varias técnicas de clasificación para lograr la clasificación final en base a un modelo consenso o ensamblado, pudiera ser una alternativa para incrementar la fiabilidad de los modelos predictivos.
7. La consideración de todos los parámetros (y sus componentes) que influyen en la biodisponibilidad oral mejoraría también los resultados de los modelos. En este sentido, pudieran combinarse datos de solubilidad, permeabilidad, unión a proteínas plasmáticas y estabilidad metabólica. La utilización de estos datos obtenidos de modelos *in silico* pueden combinarse con datos de ensayos *in vitro* e *in vivo* para la implementación de modelos farmacocinéticos basados en características fisiológicas.
8. KNIME permite la integración en un único flujo de trabajo de todos los procesos que componen el desarrollo de modelos predictivos, por lo que constituye una alternativa eficiente para la construcción de modelos *in silico* para la predicción de la biodisponibilidad oral de fármacos.

Materiales y métodos

Capítulo II. Materiales y métodos

Este capítulo describe los métodos implementados para la selección y curación de la base de datos, la construcción y validación de los modelos de clasificación, y para la determinación de los dominios de aplicación de los mismos. Las herramientas informáticas que se utilizaron para todos estos procesos se mencionan a continuación.

2.1. Herramientas informáticas empleadas

Las herramientas informáticas empleadas fueron Microsoft Excel 2016 para la selección de la base de datos y Konstanz Information Mining (KNIME), para la curación de las estructuras moleculares, el desarrollo y la evaluación de los modelos predictivos.

2.1.1. KNIME

La herramienta para la minería de datos KNIME, fue usada para el desarrollo de un flujo semiautomatizado que tiene como entrada los datos iniciales para la modelación (estructuras moleculares y valores experimentales de la propiedad) y como salida los resultados estadísticos de los modelos, los resultados de las validaciones interna y externa, las moléculas fuera del dominio de aplicación y varias salidas gráficas. Todos los trabajos se realizaron con KNIME 3.7.1, utilizándose nodos auxiliares que fueron descargados gratuitamente de http://www.knime.com. La interfaz del software se muestra en la Figura 2.1.

Los nodos *"RD-kit", "Indigo", "CDK", y "Chemistry*" fueron utilizados para la curación de la base de datos inicial y la conversión pertinente de las estructuras en varios formatos, según exigía el flujo de trabajo.

Los nodos de los paquetes *"Analytics" y "Weka 3.7"* fueron usados para la obtención de modelos de clasificación y el nodo *"Domain Leverage"* de *"Enalos"* para la determinación de los dominios de aplicación.

Otros nodos, pertenecientes a los paquetes *"Analytics"* y *"Manipulation"* fueron utilizados oportunamente para las operaciones de trasformación y procesamiento de los datos, de acuerdo a las exigencias del flujo semiautomatizado.

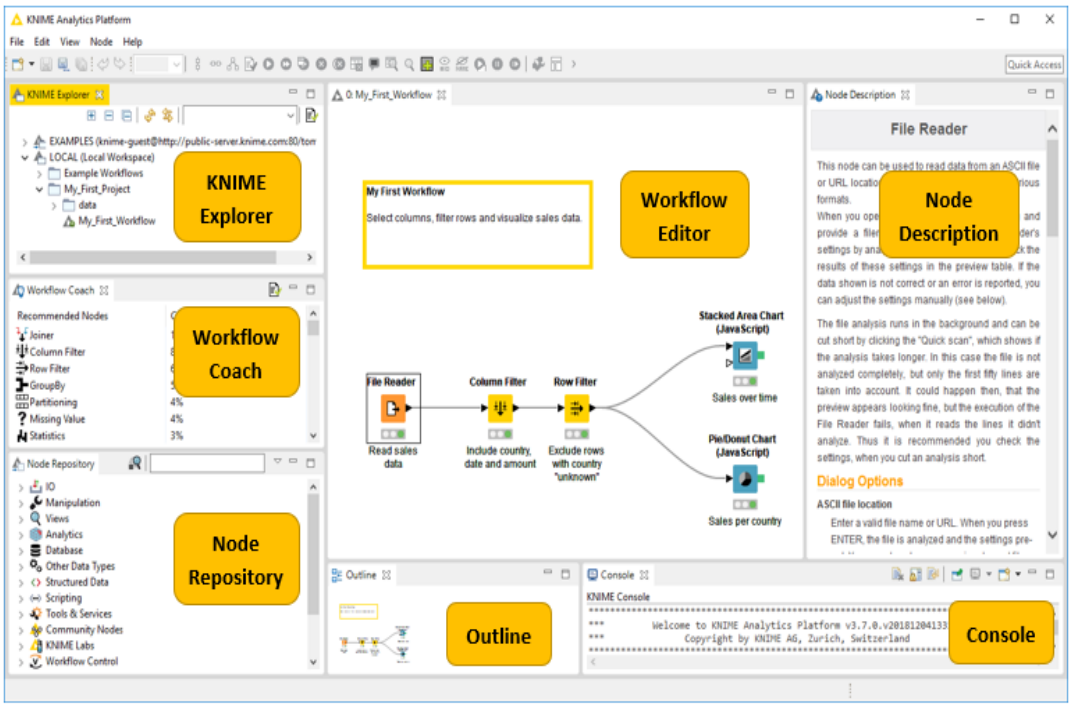


Figura 2.1. Interfaz de KNIME 3.7.1.

Para la descripción de los métodos empleados y con el objetivo de facilitar la presentación de la información, en los siguientes epígrafes el flujo semiautomatizado desarrollado será dividido en varias secciones.

2.2. Selección de la base de datos experimental

Para la selección de la base de datos con valores experimentales de biodisponibilidad, fueron consultadas tres bases de datos públicas:

1. Base de datos de 1013 compuestos con valores experimentales de biodisponibilidad oral, recolectados de 250 fuentes bibliográficas, disponible en: modem.ucsd.edu/adme/databases_bioavailability.htm.
2. Base de datos publicada por Varma y colaboradores, que integra 309 moléculas con valores de biodisponibilidad oral en humanos (72).
3. Base de datos de 995 compuestos, obtenida de la información suplementaria del artículo: *Critical Evaluation of Human Oral Bioavailability for Pharmaceutical Drugs by Using Various Cheminformatics Approaches* (40).

Las tres bases de datos fueron cotejadas con vistas a comparar la información entre ellas. Para obtener un criterio de decisión sobre cuál información mantener para los

casos donde los identificadores (SMILES canónicos y número CAS) y los valores experimentales de biodisponibilidad fueran incongruentes, se trazaron las siguientes estrategias:

1. Corrección manual de los números CAS y los SMILES canónicos tomando como referencia la base de datos especializada de PubChem, sistema operado y mantenido por el Centro Nacional para Información Biotecnológica (NCBI, por sus siglas en inglés): http://www.ncbi.nlm.nih.gov
2. Rectificación de los valores de biodisponibilidad que diferían en más de un cinco por ciento entre una fuente y otra para un mismo compuesto, tomando como referencia los artículos: Use of In Vitro and In Vivo Data to Estimate the Likelihood of Metabolic Pharmacokinetic Interactions (73) y The absolute oral bioavailability of selected drugs (74) y el libro The Pharmacological Bases of Therapeutics (75).
3. Determinación del promedio entre los valores de biodisponibilidad que diferían en menos de un cinco por ciento.

2.3. Curación de la base de datos

Luego de la selección y revisión de la base de datos experimental, el primer proceso realizado fue la curación de las estructuras moleculares, según el procedimiento descrito por Fourches y colaboradores (76). Para aplicar este procedimiento se creó la primera secuencia de nodos en KNIME.

El primer paso consistió en la revisión de las valencias de los átomos, con vistas a descartar estructuras con valencias inusuales. Seguidamente, se aplicaron filtros para: moléculas inorgánicas y moléculas que presentaran elementos diferentes a: Hidrógeno, Boro, Carbono, Nitrógeno, Oxígeno, Flúor, Cloro, Bromo y Yodo. A continuación se filtraron las sales, las estructuras desconectadas (manteniendo el mayor fragmento) y las moléculas con masa molar mayor que 1200 g/mol.

Con el objetivo de chequear la estereoquímica de los compuestos las estructuras fueron normalizadas y posteriormente se convirtieron las representaciones moleculares a una forma estándar. Los compuestos fueron aromatizados, siguiendo un mismo patrón de aromatización, y finalmente se realizó la optimización de la

geometría molecular utilizando el campo de fuerza UFF (*Universal Force Field*).Para el tratamiento de las estructuras duplicadas se generaron los códigos InChi, únicos para cada molécula y se filtraron los casos duplicados.

2.3. Parametrización de las estructuras moleculares y reducción de la dimensionalidad

Para la parametrización de las estructuras se utilizó el nodo *"alvaDesc"*, el cual es una integración en KNIME del software del mismo nombre. Este nodo permite el cálculo de un amplio rango de descriptores moleculares y *fingerprints*. Además, la determinación de más de 4000 descriptores independientes de la información tridimensional como son los constitucionales, topológicos y farmacóforos. Incluye los índices ETA y *Atom-type E-state*, el conteo de grupos funcionales y fragmentos y el cálculo de varias propiedades fisicoquímicas, entre ellas: refractividad molecular, área superficial topológica, y dos modelos de coeficiente de partición (Moriguchi y el coeficiente de partición octanol-agua de Ghose-Chippen).

Después del cálculo de los descriptores se aplicó el nodo *"Normalizer"*, y mediante la elección del parámetro *"Z-score-normalisation"* los descriptores asumieron una distribución con media aritmética 0 y desviación estándar 1 (Normalización Gaussiana).

Seguidamente, se aplicaron tres de los métodos disponibles para reducción de la dimensionalidad:

1. la eliminación de los descriptores con varianza menor a 0.001, con el nodo *"Low Variance Filter"*,
2. un modelo de correlación, seguido del nodo "*Correlation Filter*", que filtra aquellos descriptores con un índice de correlación mayor que 0.95. Durante el análisis "sobreviven" las variables con mayor número de columnas correlacionadas, mientras que las que están correlacionadas con las primeras son excluidas, y
3. el filtro de los descriptores con más del 90% de valores perdidos, mediante el nodo *"Missing Value Column Filter"*.

2.4. Clasificación inicial de los compuestos

La fortaleza de un modelo de clasificación depende de cómo define la clase alta o baja en el caso de una clasificación binaria, o las diferentes clases, en un problema de clasificación múltiple. En esta investigación se utilizó 50% como valor de corte, como alternativa para lograr un mejor balance de ambas clases (alta y baja) (16). Para ello se utilizó el nodo *"Rule Engine"*, y se creó la siguiente regla: *Si el valor de biodisponibilidad oral es mayor o igual que 50, entonces se clasificará al compuesto como de alta biodisponibilidad (1), de lo contrario será clasificado como de baja biodisponibilidad (0)*.

2.5. Selección de las series de entrenamiento, predicción y validación externa

Con el objetivo de evaluar la confiabilidad y estabilidad de los modelos que fueran construidos, a la serie original se le realizaron dos particiones sucesivas utilizando el nodo *"Partitioner"*. La primera partición se realizó para separar el 20% de los compuestos y fue totalmente aleatoria, resultando en la serie de validación externa.

El 80% restante se fraccionó, pero esta vez se utilizó un muestreo estratificado según la columna "Clase" (que diferencia los compuestos en 0 y 1 según su valor de biodisponibilidad oral) con el objetivo de tener una representación significativa de cada clase en las series de entrenamiento (para la construcción de los modelos) y predicción (para la selección de los mejores modelos con vistas al posterior desarrollo de la validación interna y externa). Los resultados de esta última división fueron: un 80% de los compuestos destinados a conformar la serie de entrenamiento y un 20% a la serie de predicción interna.

2.6. Selección y optimización de los parámetros para la construcción de los modelos

En este estudio se utilizaron seis algoritmos para la construcción de modelos de clasificación: (i) *random forest* (ii) selección de variables hacia adelante (*forward*) y luego el algoritmo J48graft*,* (iii) máquina de soporte vectorial (optimización secuencial mínima) (iv) selección de variables hacia adelante (*forward*) y luego el algoritmo *multilayer perceptron*, (v) árbol de decisión, y (vi) regresión logística.

Los modelos *random forest*, máquina de soporte vectorial (optimización secuencial mínima), árbol de decisión y regresión logística fueron optimizados, utilizando para ello la serie de predicción interna. Los valores de exactitud y *Cohen´s Kappa* fueron seleccionados como criterios para la elección de los mejores modelos.

2.6.1. Bosques aleatorios

Para la implementación del algoritmo bosques aleatorios (*random forest*) los parámetros "número de árboles" (modelos), "profundidad de árbol", y "mínimo tamaño de nodo" fueron optimizados. Adicionalmente fue optimizado el "mínimo número de nodos hijos", con vistas a lograr una generalización del algoritmo.

En los bosques aleatorios el incremento del número de árboles no produce sobre-aprendizaje, sino un valor límite de generalización del error (58). En base a esto se decidió optimizar este parámetro, asignándole seis valores posibles: 50, 100, 200, 300, 400,500. Las instancias posibles para el parámetro "profundidad de árbol" fueron: 7, 8,9 y 10; y 1, 2,3 y 4 fueron los valores establecidos para los parámetros "mínimo tamaño de nodo" y "mínimo tamaño del nodo hijo" respectivamente.

En cuanto a los otros parámetros de la configuración se mantuvo el índice de Gini en todos los casos, como medida para la elección del valor de corte del descriptor molecular, que separará las moléculas de la serie de entrenamiento en nuevas subseries que formarán los nodos hijos.

2.6.2. Máquina de soporte vectorial de clasificación

Para el desarrollo de la máquina de soporte vectorial se seleccionó el nodo "SMO" de Weka 3.7 que implementa el algoritmo de optimización secuencial mínima de John Platt para el entrenamiento de un clasificador de soporte vectorial. La función de Kernel usada para la transformación de los datos en la máquina de soporte vectorial fue *"NormalizedPolyKernel"*, con tamaño de cache 250007 y exponente 2.0.

Como parámetro de optimización se seleccionó *"filterType"* que indica si/cómo serán los datos transformados. La opción admite tres instancias: (0) para que los datos

sean normalizados, (1) para la estandarización de los datos y (2) para que no sean transformados.

2.6.3. Selección de variables hacia adelante y Red Neuronal Multicapa

El nodo *"AttributeSelectedClassifier*" de Weka 3.7 ofrece la ventaja de implementar un proceso de selección de variables antes de la implementación de cualquier algoritmo de clasificación. Se aplicó este nodo para ejecutar un proceso de selección de variables eligiendo la opción *"BestFirst"* y dentro de ella en el parámetro *"Direction"* la opción hacia adelante ("*Forward*") fijando la cantidad de pasos en 25, seguido de un algoritmo de clasificación por *"Multilayer Perceptron"*. Los parámetros de la configuración fueron: "t" como parámetro fijado para las capas ocultas, definido por la suma de las clases y los atributos; 0.2 como ritmo de aprendizaje; 0.3 como velocidad ("*Momentum"*) y 1500 como tiempo de aprendizaje.

2.6.4. Árbol de decisión

La primera variante de árboles de clasificación se implementó mediante el nodo *"DecisionTreeLearner"*. Se utilizó el índice de Gini como medida de calidad y como criterio para el corte. Se empleó MDL (*Minimal Description Lenght*) como método de poda (*pruning)* para prevenir la construcción de árboles excesivamente complicados y reducir el sobre-aprendizaje. Además se seleccionó la opción que permite reducir el error del *pruning*.

El número de hilos (*threads*) fue optimizado, estableciendo como valores posibles los valores naturales del siguiente intervalo: [1; 8].

El mínimo número de parámetros (descriptores) por nodo se corresponde con un criterio de parada *(preprunning)*, de manera que si el número de parámetros por nodo es menor o igual al valor fijado se detendrá el crecimiento del árbol. Los valores establecidos para la optimización de este parámetro fueron los mismos que para la optimización anterior.

2.6.5. Selección de variables hacia adelante y aplicación del algoritmo J48

El algoritmo C4.5 (J48) se utilizó para generar otro árbol de clasificación utilizando el nodo *"AtributteSelectClassifier"* de Weka 3.7. Una selección de variables hacia

adelante se desarrolló antes de la implementación del clasificador J48graft, fijando en 25 la cantidad de pasos. Esta variante de los Árboles de Decisión para Clasificación y Regresión presenta el mismo principio básico para la división, las diferencias radican en la estructura del árbol, el método de *pruning,* el criterio usado para el corte y la forma de manipulación de los valores perdidos. El parámetro optimizado durante este clasificador fue “mínimo número de parámetros por nodo”. El factor de confidencia se mantuvo por defecto, fijado en 0,25, al igual que los otros parámetros del modelo.

2.6.6. Regresión logística

Utilizando el nodo *“KernelLogisticRegression”* se desarrolló un modelo de clasificación mediante una regresión logística utilizando una función de Kernel. La función de Kernel seleccionada fue *“NormalizedPolyKernel”*. El parámetro clave para la optimización fue “lambda” que asumió los valores comprendidos en el intervalo entre 0,01 y 0,09 con saltos de 0,01. Alternativamente se seleccionó la opción que permite la optimización por gradiente conjugado.

2.7. Clasificación de los compuestos en base al voto mayoritario

En orden de mejorar la exactitud y fiabilidad de las predicciones, los seis modelos anteriores fueron combinados para obtener una única clasificación, basada en un voto mayoritario simple.

2.8. Evaluación de las predicciones en base a las probabilidades de pertenencia a clases

Los modelos de clasificación permiten la asignación de una clase determinada a cada objeto en base a reglas de clases, límites de clases y probabilidades de pertenencia a clases. Las probabilidades de pertenencia a clases indican la posibilidad de que un compuesto pertenezca a una u otra clase. Si la diferencia entre ambos valores es pequeña para un compuesto “X”, es lógico pensar que el modelo no puede discernir con confiabilidad la clase a la cual pertenece dicho compuesto.

En este sentido, se filtraron aquellos compuestos de las series de predicción y validación externa, para los cuales la diferencia modular entre las probabilidades fuera menor que cierto valor límite, fijado previamente con el objetivo de evaluar el comportamiento de los resultados estadísticos de los modelos cuando estos compuestos fueran removidos.

2.9. Evaluación de los modelos

Los valores de exactitud y *Cohen's Kappa* fueron los parámetros claves para la elección de los mejores modelos durante la optimización, que se realizó con los compuestos de la serie de predicción. Una vez seleccionados los mejores modelos, se desarrollaron las validaciones interna y externa.

La validación externa se realizó para cada modelo individual y para los resultados del voto mayoritario, utilizándose en todos los casos el 20% del total de compuestos (resultado de la primera partición).

Cada modelo de clasificación fue validado internamente utilizando la serie de entrenamiento mediante cinco validaciones cruzadas, utilizando el metanodo *"CrossValidation"*. Además, se realizaron cinco validaciones cruzadas, pero considerando la clasificación final resultado del voto mayoritario. Para la ejecución de esta validación interna se integró en el metanodo *"CrossValidation*" todos los clasificadores, y los nodos correspondientes para la concatenación de las predicciones y para la decisión final.

Los compuestos clasificados empleando un modelo de clasificación QSPR pueden ser divididos en cuatro categorías, en base a la comparación entre la clase predicha y observada de la propiedad en estudio: (i) verdaderos positivos, (ii) falsos positivos, (iii) verdaderos negativos y (iv) falsos negativos. Basado en esta clasificación y el número de moléculas de cada una de las series de prueba, se construye la matriz de confusión de 2x2, también referida como tabla de contingencia.

La calidad predictiva y la estabilidad de los modelos fueron evaluadas mediante el análisis de los valores de exactitud, sensibilidad, especificidad, precisión, y *Cohen's*

Kappa durante las validaciones externa e interna. Estos parámetros son definidos matemáticamente en la Tabla 2.1.

Tabla 2.1. Métricas para la evaluación de los modelos de clasificación QSPR

Métricas para la evaluación de los modelos de clasificación QSPR	
Sensibilidad= *Recall*	$\frac{TP}{TP+FN}$
Especificidad	$\frac{TN}{TN+FP}$
Exactitud	$\frac{TP+TN}{TP+FP+TN+FN}$
Precisión	$\frac{TP}{TP+FP}$
Cohen's Kappa	$\frac{\mathbf{Pr}(a)-\mathbf{Pr}(e)}{\mathbf{1}-\mathbf{Pr}(e)}$; donde: Pr(a)= $\frac{TP+TN}{TP+FP+FN+TN}$ y Pr(b)=$\frac{(TP+FP)X(TP+FN)+(TN+FP)X(TN+FN)}{(TO+FN+FP+TN)^2}$
*TP (True positives):*Verdaderos positivos; *FN (False negatives)*:Falsos negativos; *TN (True negatives)*:Verdaderos negativos; *FP (False positives)*:Falsos positivos	

2.10. Dominio de aplicación

La fiabilidad de un modelo QSPR depende de su capacidad para alcanzar predicciones confiables para nuevos compuestos que no fueron considerados en la construcción del modelo, y está definida por su dominio de aplicación.

Entre las aproximaciones para determinar el dominio de aplicación de un modelo se encuentran los métodos basados en la distancia, que calculan la distancia de los compuestos desde un punto definido dentro del espacio de descriptores de la serie de entrenamiento. La idea general es comparar las mediciones de las distancias entre el punto definido y la serie de datos con un límite definido.

Aproximaciones similares basadas en el *leverage* son ampliamente recomendadas en los estudios QSAR/QSPR. El *leverage* de cada compuesto es proporcional a la distancia de Mahalanobis medida desde el centro de la serie de entrenamiento. Los *leverages* son calculados para una serie de datos X, a partir de la matriz de *Leverage* (H) según la siguiente ecuación:

$$hi = X(X^T X)^{-1} X^T \qquad (2.1)$$

Donde: X es la matriz del modelo y X^T es la matriz transpuesta.

Los valores diagonales de la matriz H representan los valores de *leverage* para diferentes puntos de la serie de datos. Un valor mayor que 3k/n (donde k: número de descriptores y n: número de compuestos de la serie de entrenamiento) es considerado elevado y significa que la respuesta predicha es resultado de una sustancial extrapolación del modelo y puede no ser confiable.

Para la implementación de este método se utilizó el nodo *"Domain-Leverage"*, de Enalos. Se determinó el dominio de aplicación para cada uno de los seis modelos, luego de implementar una secuencia de nodos para extraer las variables comprendidas en cada modelo.

Otra aproximación que fue empleada para la determinación del dominio de aplicación está basada en el algoritmo *random forest.* El método consiste en el desarrollo de un modelo *random forest*, con parámetros prefijados, seguido de un proceso de extracción de variables, integrado por múltiples iteraciones. Durante cada iteración son definidas las reglas de cada árbol de clasificación construido durante el algoritmo, y son creadas adicionalmente reglas que permiten determinar si un compuesto se encuentra dentro o fuera del dominio de aplicación.

2.11. Interpretación de las variables de los modelos

Para la interpretación de las variables implicadas en los modelos, primeramente se creó una secuencia de nodos que permitió la extracción de las variables de cada modelo de manera individual, siempre que fuera posible. Posteriormente se creó una tabla que relacionaba las variables extraídas con su frecuencia de aparición. La interpretación de las variables se realizó mediante la revisión de la literatura especializada.

Resultados y discusión

Capítulo III. Resultados y discusión

Este capítulo presenta los resultados obtenidos después de la aplicación de todos los procedimientos descritos en el capítulo II. Para la mejor comprensión de la información que se presenta, además de las salidas gráficas y estadísticas del software KNIME, la información se organiza en tablas y gráficos adicionales. El capítulo se divide en cuatro secciones fundamentales: (i) selección y curación de la base de datos; (ii) parametrización de las estructuras, reducción de la dimensionalidad y selección de las series de entrenamiento, predicción y validación externa;(iii) construcción, validación y aplicabilidad de los modelos; e (iv) interpretación. Cada uno de los epígrafes dedicados a estas secciones presenta la parte del flujo semiautomatizado correspondiente.

3.1. Desarrollo del flujo de trabajo

El flujo de trabajo desarrollado se encuentra dividido en dos partes fundamentales: (i) toda la secuencia de nodos creada para la curación, parametrización de la estructura molecular, pre-procesamiento de los datos, partición de los mismos, obtención, validación y determinación del dominio de aplicación de los modelos y (ii) la secuencia de pasos para la clasificación de cualquier nuevo compuesto en virtud de los modelos construidos.

3.2. Selección y curación de la base de datos

Luego de la concatenación de las tres bases de datos consultadas, la rectificación de los identificadores y la corrección de los valores de biodisponibilidad oral, la base de datos lista para el proceso de curación quedó conformada por 1063 fármacos y moléculas afines, con sus correspondientes SMILES canónicos, números CAS y valores experimentales de biodisponibilidad.

Tomando como base el tipo de propiedad en estudio se puede plantear que es una base de datos grande y con una amplia diversidad estructural, lo cual es un requerimiento para un adecuado proceso de modelación.

El proceso de curación excluyó seis compuestos. Cinco compuestos se eliminaron debido a su masa molar, mayor que 1200 g/mol y el otro compuesto se excluyó después del análisis de estructuras duplicadas. Las estructuras de los compuestos removidos se presentan en el Anexo 1.

3.3. Parametrización de las estructuras, reducción de la dimensionalidad y selección de las series de entrenamiento, predicción y de validación externa

Los descriptores 0,1 y 2D fueron calculados para las 1057 estructuras, mediante el nodo *"alvaDesc"*, implementado en KNIME para un total de 3847 descriptores moleculares, que fueron posteriormente normalizados. El filtro de los descriptores con varianza inferior a 0.001 excluyó 1136. El análisis de correlación mostró 1403 variables con un índice de correlación superior a 0.95, las cuales fueron eliminadas. La matriz de correlación se muestra en la Tabla 2 de los Anexos. Finalmente, quedaron disponibles 1306 descriptores para la modelación.

Mediante el nodo *"Rule Engine"* cada valor de biodisponibilidad fue asociado con el valor 0 o 1, de acuerdo a la regla descrita en el epígrafe 2.5.De esta forma los datos quedaron inicialmente divididos en 477 compuestos con baja biodisponibilidad oral (<50%) y 580 compuestos con alta biodisponibilidad oral (≥50%).

Aunque todavía no existe un consenso sobre qué valor *cutoff* utilizar para el desarrollo de problemas de clasificación binaria para la predicción de la biodisponibilidad y valores como 20% (37), 30% (23); y 50% (16) han sido empleados en la literatura, en esta investigación se decidió optar por el valor 50% como alternativa para balancear ambas clases. No obstante, en estudios posteriores podría utilizarse otro valor *cutoff* y emplearse consecutivamente alguna estrategia de re-muestreo para el balance de las clases como son: el sobre-muestreo (*oversampling*) con reemplazo, *oversampling* aleatorio, sub-muestreo (*undersampling)* o *Synthetic Minority Oversampling Technique (SMOTE)* (77).

Para la selección de las series de entrenamiento, predicción interna y externa se decidió realizar dos divisiones. La primera, completamente aleatoria, separó 212 compuestos (serie de validación externa) que no estarían implicados en la

construcción de los modelos. Esta primera partición decidió realizarse de manera completamente aleatoria, para tener una medida de la calidad predictiva del modelo para la clasificación de compuestos externos.

Los 845 compuestos restantes fueron divididos nuevamente para conformar la serie de entrenamiento (676) y la serie de predicción interna (169); esta vez el muestreo se realizó estratificado según la columna “Clase” para lograr una representación significativa de ambas clases en las series de entrenamiento y predicción. El objetivo de obtener esta serie de predicción (o prueba) fue la selección de los mejores modelos durante el proceso de optimización.

La serie de entrenamiento fue utilizada para la validación cruzada para la evaluación de los modelos individuales y para la evaluación de las predicciones de acuerdo al voto global o mayoritario.

3.3. Optimización, obtención y evaluación de los modelos individuales

El proceso de optimización se realizó utilizando una secuencia de nodos comunes a todos los modelos, donde solo variaban los parámetros establecidos en el nodo *“Table Creator”*, los cuales estaban en correspondencia con los modelos individuales y, por supuesto con las variables seleccionadas para la optimización, estas últimas, detalladas en el capítulo II.

Para la evaluación de los modelos individuales se utilizó el metanodo *“CrossValidation”*. Para la evaluación de las predicciones realizadas por el voto global o mayoritario en el metanodo “CrossValidation” anterior se integraron los nodos “_Learner” y “_Predictor” correspondientes a cada uno de los modelos seguido de la secuencia de nodos creada para aplicar el voto mayoritario.

3.3.1. Bosques aleatorios

A partir del análisis de los resultados generados durante la optimización del algoritmo *random forest*, se determinó que los valores 9, 1,4 y 50 fueron los óptimos para la "profundidad de árbol", "mínimo tamaño de nodo", "mínimo tamaño de nodo hijo" y "número de árboles". La ejecución del modelo con estos parámetros tuvo como resultados los descritos en la Tabla 3.1.

Tabla 3.1. Resultados del *random forest* con los parámetros optimizados.

Serie	Clase	VP	FP	*Recall*	Precisión	Sensibilidad	Especificidad	Exactitud	Cohen´s Kappa
	1	370	11	1,00	1,00	0,96	0,98		
	0	295	0	0,96	0,96	1,00	0,98		
SE								**0,98**	**0,97**
	1	71	18	0,80	0,76	0,76	0,78		
	0	58	22	0,73	0,76	0,76	0,74		
SP								**0,76**	**0,52**
	1	99	43	0,85	0,897	0,85	0,55		
	0	52	18	0,55	0,743	0,55	0,85		
SVE								**0,71**	**0,40**
	1	299	89	0,77	0,67	0,77	0,55		
	0	161	127	0,56	0,67	0,55	0,77		
VC								**0,67**	**0,32**

SE: serie de entrenamiento; SP: serie de predicción; SVE: serie de validación externa; VC: validación cruzada; VP: verdaderos positivos; FP: falsos positivos; 0 (baja biodisponibilidad); 1 (alta biodisponibilidad)

3.3.2. Máquina de soporte vectorial por optimización secuencial mínima

El mejor modelo obtenido a partir de la utilización del algoritmo de optimización secuencial mínima para el entrenamiento del clasificador de soporte vectorial se alcanzó cuando los datos iniciales fueron estandarizados. La función de Kernel que se utilizó fue *"NormalizedPolyKernel"*. Los resultados del mejor modelo se muestran en la Tabla 3.2.

3.3.3. Selección de variables por *Forward* y Red Neuronal Multicapa

Los resultados del modelo obtenido después de la implementación del nodo *"AtributteSelectClassifier"* para el entrenamiento de una Red Neuronal Multicapa precedida de una selección de variables hacia adelante se muestran en la Tabla 3.4. Este nodo permitió la selección de variables como método para reducir la

dimensionalidad y la implementación de la Red Neuronal Multicapa, utilizando esta última el algoritmo de retropropagación de errores.

Tabla 3.2. Resultados de la máquina de soporte vectorial por optimización secuencial mínima con los parámetros optimizados.

Serie	Clase	VP	FP	Recall	Precisión	Sensibilidad	Especificidad	Exactitud	Cohen´s Kappa
	1	347	41	0,94	0,89	0,94	0,87		
	0	265	23	0,87	0,92	0,87	0,94		
SE								0,91	0,81
	1	78	30	0,84	0,72	0,84	0,61		
	0	46	15	0,61	0,75	0,61	0,84		
SP								0,73	0,45
	1	101	38	0,86	0,73	0,86	0,60		
	0	57	16	0,60	0,78	0,60	0,86		
SVE								0,75	0,47
	1	268	120	0,80	0,68	0,55	0,76		
	0	158	130	0,55	0,69	0,80	0,61		
VC								0,69	0,36

SE: serie de entrenamiento; SP: serie de predicción; SVE: serie de validación externa; VC: validación cruzada; VP: verdaderos positivos; FP: falsos positivos; 0 (baja biodisponibilidad); 1 (alta biodisponibilidad)

Tabla 3.3. Resultados de la red neuronal multicapa implementada luego de la selección de variables hacia adelante.

Serie	Clase	VP	FP	Recall	Precisión	Sensibilidad	Especificidad	Exactitud	Cohen's Kappa
	1	353	17	0,954	0,95	0,95	0,94		
	0	289	17	0,944	0,94	0,94	0,95		
SE								0,95	0,90
	1	69	28	0,742	0,71	0,74	0,63		
	0	48	24	0,632	0,67	0,63	0,74		
SP								0,69	0,38
	1	94	38	0,803	0,71	0,80	0,60		
	0	57	23	0,600	0,71	0,60	0,80		
SVE								0,71	0,41
	1	275	113	0,71	0,67	0,71	0,56		
	0	161	127	0,56	0,61	0,56	0,71		
VC								0,64	0,34

SE: serie de entrenamiento; SP: serie de predicción; SVE: serie de validación externa; VC: validación cruzada; VP: verdaderos positivos; FP: falsos positivos; 0 (baja biodisponibilidad); 1 (alta biodisponibilidad)

3.3.4. Árbol de decisión

El Árbol de Decisión implementado mediante el nodo “DecisionTreeLearner”, que presentó los mejores resultados durante la optimización tuvo la siguiente configuración: 4 como número de “hilos” (*threads*) y 4 como “mínimo tamaño de nodo”. Para la construcción del árbol se fijó la variable TPSA (Área Superficial Polar Topológica) (Total) como nodo raíz. La selección de esta variable se basó en los resultados de la revisión bibliográfica. El área superficial polar topológica (total, porque considera las contribuciones polares de los átomos de O, N, S y P) se ha reportado como variable implicada en varios modelos publicados (17,18).

Los resultados del mejor modelo obtenido mediante este algoritmo se muestra en la Tabla 3.4.

Tabla 3.4. Resultados del modelo árbol de decisión con los parámetros optimizados

Serie	Clase	VP	FP	*Recall*	Precisión	Sensibilidad	Especificidad	Exactitud	Cohen´s Kappa
	1	341	52	0,92	0,87	0,92	0,83		
	0	254	29	0,83	0,90	0,83	0,92		
SE								**0,88**	**0,76**
	1	66	29	0,71	0,70	0,71	0,62		
	0	47	27	0,62	0,64	0,62	0,71		
SP								**0,67**	**0,33**
	1	93	41	0,80	0,69	0,80	0,57		
	0	54	24	0,57	0,69	0,57	0,80		
SVE								**0,69**	**0,37**
	1	272	116	0,70	0,66	0,70	0,68		
	0	161	127	0,56	0,60	0,56	0,58		
VC								**0,66**	**0,23**

SE: serie de entrenamiento; SP: serie de predicción; SVE: serie de validación externa; VC: validación cruzada; VP: verdaderos positivos; FP: falsos positivos; 0 (baja biodisponibilidad); 1 (alta biodisponibilidad)

3.3.5. Regresión logística

La optimización del parámetro “lambda” de la regresión logística concluyó en que el mejor modelo se obtenía cuando este parámetro se fijaba en 0.06, usando la función de Kernel *“NormalizedPolyKernel”* y el gradiente de descenso conjugado. Los resultados se resumen en la Tabla 3.5.

3.3.6. Selección de variables hacia adelante y aplicación del algoritmo C4.5 (J48graft)

Los resultados obtenidos con este algoritmo fueron mejores que los obtenidos con la aplicación del nodo *"DecisionTree"*, sobre todo el parámetro sensibilidad, donde se supera el valor 0,6 para ambas clases de ambas series de datos. El "mínimo tamaño de nodo" óptimo fue 1. Los resultados se muestran en la Tabla 3.6.

Tabla 3.5. Resultados de la regresión logística con los parámetros optimizados.

Serie	Clase	VP	FP	*Recall*	Precisión	Sensibilidad	Especificidad	Exactitud	Cohen´s Kappa
	1	368	8	0,99	0,97	0,99	0,97		
	0	298	2	0,97	0,993	0,97	0,99		
SE								***0,98***	***0,97***
	1	72	29	0,77	0,713	0,77	0,61		
	0	47	21	0,61	0,691	0,61	0,77		
SP								***0,70***	***0,396***
	1	100	41	0,86	0,709	0,86	0,57		
	0	54	17	0,57	0,761	0,57	0,86		
SVE								***0,73***	***0,43***
	1	287	101	0,74	0,68	0,74	0,57		
	0	164	124	0.57	0,65	0,57	0,74		
VC								***0,67***	***0,32***

SE: serie de entrenamiento; SP: serie de predicción; SVE: serie de validación externa; VC: validación cruzada; VP: verdaderos positivos; FP: falsos positivos; 0 (baja biodisponibilidad); 1 (alta biodisponibilidad)

Tabla 3.6. Resultados del algoritmo J48graft implementado luego de la selección de variables hacia adelante.

Serie	Clase	VP	FP	*Recall*	Precisión	Sensibilidad	Especificidad	Exactitud	Cohen´s Kappa
	1	365	27	0,98	0,93	0,98	0,91		
	0	279	5	0,91	0,98	0,91	0,98		
SE								***0,95***	***0,90***
	1	66	23	0,71	0,74	0,71	0,70		
	0	53	27	0,70	0,67	0,70	0,71		
SP								***0,70***	***0,41***
	1	91	37	0,78	0,71	0,78	0,61		
	0	58	26	0,61	0,69	0,61	0,78		
SVE								***0,70***	***0,39***
	1	260	128	0,67	0,64	0,67	0,54		
	0	156	132	0.54	0,54	0,54	0,67		
VC								***0,61***	***0,20***

SE: serie de entrenamiento; SP: serie de predicción; SVE: serie de validación externa; VC: validación cruzada; VP: verdaderos positivos; FP: falsos positivos; 0 (baja biodisponibilidad); 1 (alta biodisponibilidad)

3.4. Clasificación de los compuestos en base al voto mayoritario

Los resultados de la clasificación final después de la aplicación del voto mayoritario simple se muestran en la Tabla 3.7.

Tabla 3.7. Resultados de la clasificación de los compuestos en base al voto mayoritario.

Serie	VP	FP	*Recall*	Precisión	Sensibilidad	Especificidad	Exactitud	Cohen's kappa
	74	17	0,80	0,81	0,80	0,78		
	59	19	0,78	0,76	0,78	0,80		
SP							**0,79**	**0,57**
	100	31	0,86	0,76	0,86	0,67		
	64	17	0,67	0,79	0,67	0,86		
SE							**0,77**	**0,54**
	291	97	0,75	0,72	0,75	0,64		
	184	104	0,64	0,68	0,64	0,75		
VC							**0,70**	**0,40**

SP: serie de predicción; SVE: serie de validación externa; VC: validación cruzada; VP: verdaderos positivos; FP: falsos positivos; 0 (baja biodisponibilidad); 1 (alta biodisponibilidad)

3.4.1. Comparación entre los modelos individuales y los resultados del voto mayoritario

Los gráficos 3.1, 3.2 y 3.3 comparan los valores de exactitud, precisión y sensibilidad obtenidos para cada modelo con los resultados del voto mayoritario, para las series interna y externa.

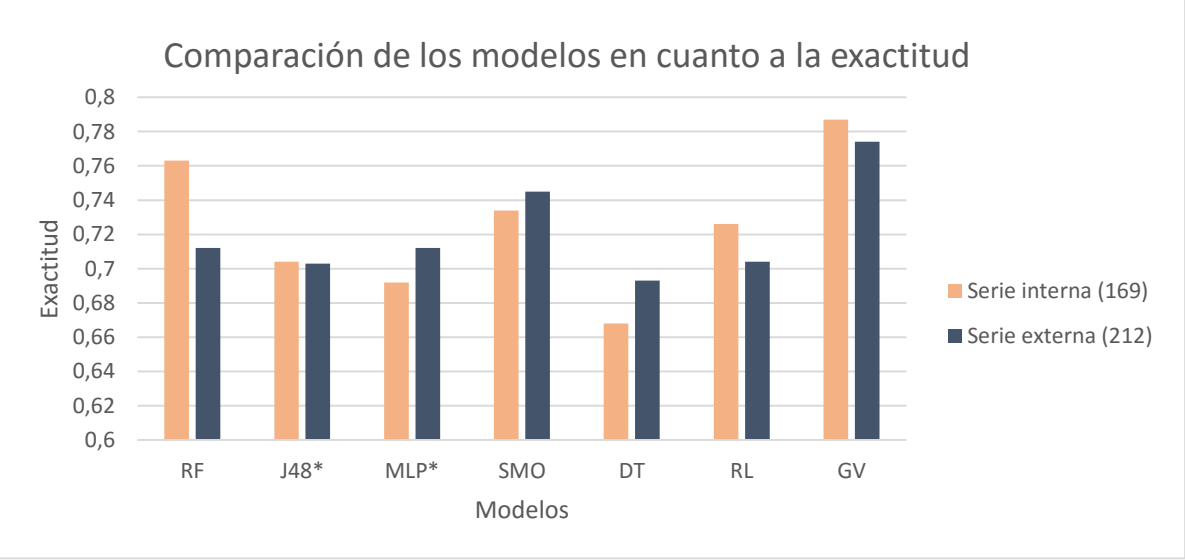


Figura 3.1. Valores de exactitud de los modelos y el voto mayoritario para la serie externa (en azul) y la serie interna (en rojo). RF: *random forest*; J48: árbol de Clasificación J48; MLP: perceptrón multicapa ("multilayer perceptron"); "SMO": máquina de soporte vectorial implementada según el algoritmo de optimización secuencial mínima; DT: árbol de decisión, RL: regresión logística, *(que fueron implementados con el nodo de Weka 3.7 *"AtributteSelectClassifier"*, utilizando previamente selección de variables hacia adelante), GV: voto mayoritario o global.

El modelo *random forest* mostró el mayor valor de exactitud para la serie interna entre todos los modelos individuales, presentando además valores balanceados de sensibilidad para ambas clases. Sin embargo la exactitud de la predicción para la serie externa es más baja, esta diferencia podría explicarse por el desbalance entre los valores de sensibilidad entre las clases alta y baja. El valor de exactitud más alto para la serie externa se obtuvo con el modelo de máquina de soporte vectorial basada en optimización secuencial mínima. Este modelo fue el que generó resultados más balanceados para ambas series y para ambas clases.

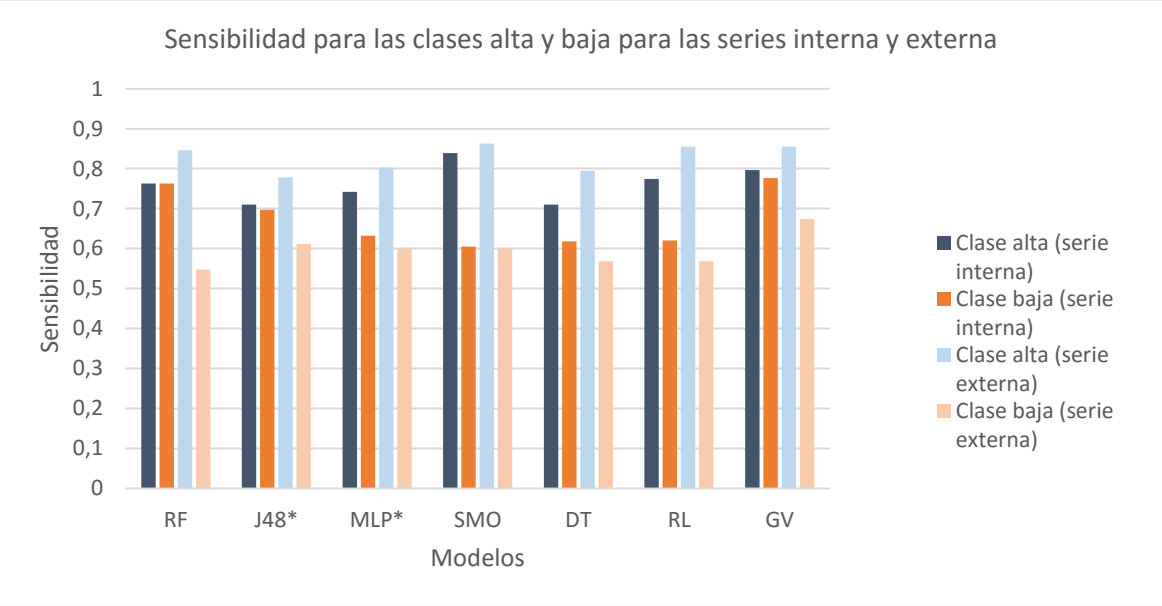


Figura 3.2. Comparación entre los valores de sensibilidad para las clases alta y baja de las series de predicción interna y validación externa de los modelos individuales y el voto global. RF: *random forest*; J48: árbol de Clasificación J48; MLP: perceptrón multicapa ("multilayer perceptron"); "SMO": máquina de soporte vectorial

implementada según el algoritmo de optimización secuencial mínima; DT: árbol de decisión, RL: regresión logística, *(que fueron implementados con el nodo de Weka 3.7 *"AtributteSelectClassifier"*, utilizando previamente selección de variables hacia adelante), GV: voto mayoritario o global.

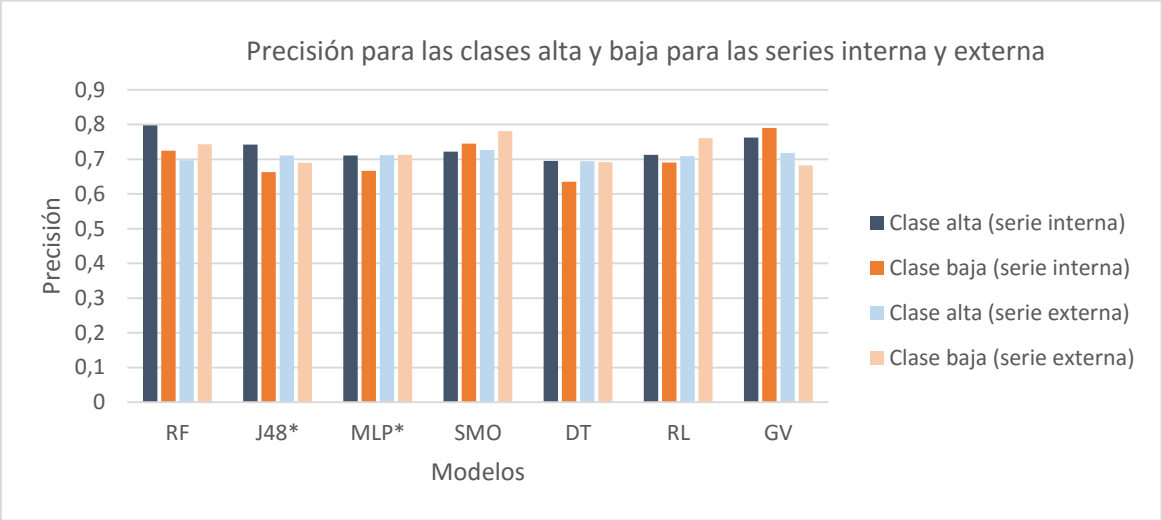


Figura 3.3. Valores precisión para las clases alta y baja de las series interna y externa de los modelos individuales y el voto global. RF: *random forest*; J48: árbol de Clasificación J48; MLP: perceptrón multicapa ("multilayer perceptron"); "SMO": máquina de soporte vectorial implementada según el algoritmo de optimización secuencial mínima; DT: árbol de decisión, RL: regresión logística, *(que fueron implementados con el nodo de Weka 3.7 *"AtributteSelectClassifier"*, utilizando previamente selección de variables hacia adelante), GV: voto mayoritario o global.

En general, el principal inconveniente de todos los modelos es la baja sensibilidad para la clase de baja biodisponibilidad.

La combinación de los seis algoritmos de clasificación mejoró los resultados estadísticos y consecuentemente la confiabilidad de las predicciones. Con la aplicación del voto mayoritario los valores de sensibilidad para la clase baja de ambas series de datos superaron al valor 0,67; resultado que no se había logrado con ningún modelo, y que influye directamente en la exactitud de los mismos. La Tabla 3.8 muestra una comparación entre los modelos disponibles en la literatura y el modelo consenso construido.

Tabla 3.8. Comparación entre modelos descritos en la literatura y el modelo consenso construido.

Métodos	Número de compuestos	Parámetros estadísticos	Ref.
GA/SW/AFP	N_1=272; N_2=432	R^2=0.75; R^2=0.64	(36)
GA-CG-SVM-clasificación	N=690	Q^2 (5 *cross-fold*)=0.80; Clase de baja biodisponibilidad: Exactitud:0.25	(37)
ORMUCS-clasificación	N (Entr.)=232 N (Ext.) =40	R^2= 0.71;Q^2 (LOO)=0.67) R^2=0.60	(38)
Clasificación Logística	N=969	Q^2= 0.71. Exactitud 0.71-0.75	(39)
RF, SVM, kNN, CASE ULTRA	N=995	$R^2_{consenso}$ =0.28, Q^2=0.76	(40)
RF,J48*,MLP*,SMO,DT,RL	Total:1057 N (Entr.)=676 N (P.interna)= 169 N (Ext.) = 212	Consenso: Exactitud (N (p.interna)=0,787 Exactitud (N (externa))=0,774 Q^2 consenso (5-folds) = 0,70	

AFP (*Adaptive Fuzzy Partition*); CG (*Conjugated gradient*): Gradiente conjugado; ORMUCS (*Ordered multicategorical classification method using the simplex technique*); RF (random forest): Bosque aleatorio; kNN (*k-Nearest Neighbor*): k-vecino más cercanos; LOO (*leave-one-out*): dejando uno fuera. Ref: Referencias bibliográficas.
N (Entr): serie de entrenamiento. N (P. interna): serie de predicción interna. N (Ext):serie de validación externa

3.5. Evaluación de las predicciones en base a las probabilidades de pertenencia a clases

Los resultados obtenidos al aplicar este procedimiento se presentan en la Tabla 3.9.

3.6. Dominio de aplicación

El dominio de aplicación del modelo árbol de decisión se determinó según el método basado en el *leverage*. Para esto fue necesario previamente la extracción de las principales variables del modelo.

.

Tabla 3.9. Resultados de la evaluación de las predicciones en base a las probabilidades de pertenencia a clases.

Clase	VP	FP	*Recall*	Precisión	Sensibilidad	Especificidad	Exactitud	Cohen's kappa	Valores límites
Serie de predicción interna									
1	70	18	0,83	0,80	0,83	0,75			
0	53	14	0,75	0,79	0,75	0,83			
							0,79	**0,58**	0.01
1	63	17	0,83	0,79	0,83	0,75			
0	51	13	0,75	0,80	0,75	0,83			
							0,79	**0,58**	0.025
1	53	12	0,79	0,82	0,79	0,81			
0	51	14	0,81	0,79	0,81	0,79			
							0,80	**0,60**	0.05
1	44	9	0,82	0,83	0,82	0,82			
0	42	10	0,82	0,81	0,82	0,81			
							0,82	**0,64**	0.1
1	40	7	0,85	0,85	0,85	0,82			
0	32	7	0,82	0,82	0,82	0,85			
							0,84	**0,67**	0.175
1	37	7	0,90	0,84	0,90	0,81			
0	29	4	0,81	0,88	0,81	0,90			
							0,86	**0,71**	0.2
Serie de validación externa									
1	99	35	0,89	0,74	0,89	0,6			
0	50	13	0,59	0,79	0,59	0,89			
							0,76	**0,49**	0.01
1	98	30	0,89	0,77	0,89	0,6			
0	44	12	0,60	0,79	0,60	0,9			
							0,77	**0,51**	0.025
1	84	17	0,84	0,83	0,84	0,72			
0	44	16	0,72	0,73	0,72	0,84			
							0,80	**0,56**	0.05
1	77	14	0,87	0,85	0,87	0,71			
0	35	12	0,71	0,75	0,71	0,87			
							0,81	**0,59**	0.1
1	69	12	0,91	0,85	0,90	0,67			
0	24	7	0,67	0,77	0,67	0,91			
							0,83	**0,6**	0.175
1	67	11	0,93	0,86	0,93	0,69			
0	24	5	0,69	0,83	0,69	0,93			
							0,85	**0,65**	0.2

VP: verdaderos positivos; FP: falsos positivos; 0 (baja biodisponibilidad); 1 (alta biodisponibilidad); Valores límites: los seis valores establecidos para el filtro de la diferencia modular de las probabilidades de pertenencia a clases.

El valor límite de *leverage* establecido fue 0,115 definido por el triplo del cociente entre el número de variables (descriptores) y el número de compuestos de la serie de entrenamiento. En base a este número se consideraron 27 compuestos fuera del dominio de aplicación. El listado de estos compuestos aparece en el Anexo 2. Los resultados de graficar el valor de *leverage* de cada compuesto se muestran en la Figura 3.4.

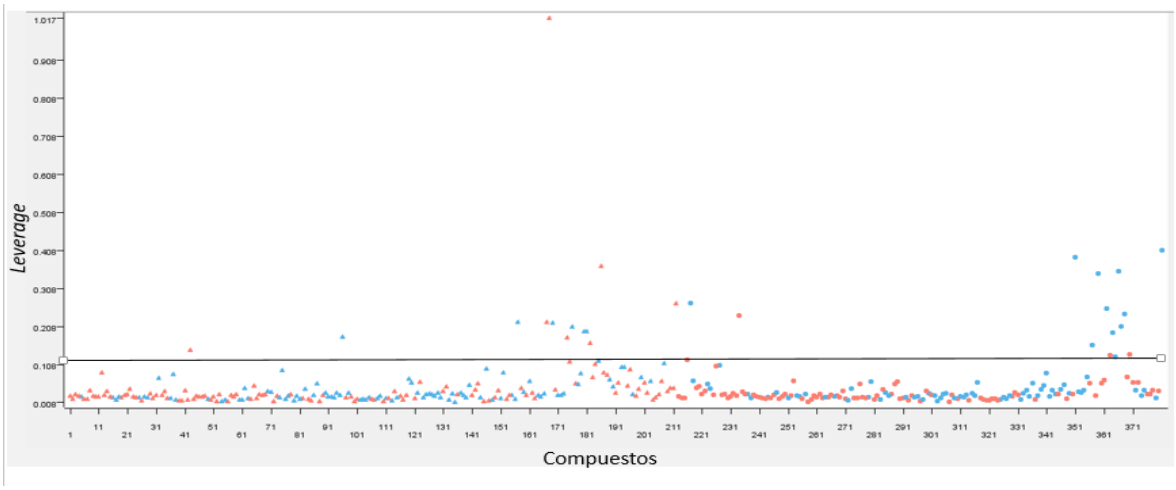


Figura 3.4. *Leverage vs* compuesto. Modelo: árbol de decisión.

Para la estimación del dominio de aplicación del random *forest* se utilizaron dos aproximaciones, la primera basada en el *leverage* y la segunda en múltiples iteraciones del propio algoritmo *random forest*.

El valor de límite de *leverage* para este modelo fue 0,74. Como puede observarse en la Figura 3.15 solo tres compuestos se alejan marcadamente del valor límite, estos son: etanol (con un valor de *leverage* de 324,8), tirosina (71,59) y diatrizoato (23,08), señalados con un círculo en el gráfico.

En el cuadro ampliado de la Figura 3.5 la línea trazada representa el valor límite de *leverage*, todos los compuestos que superan esta línea, incluyendo, por supuesto los tres compuestos señalados con un círculo, se encuentran fuera del dominio de

aplicación del modelo *random forest.* Son en total 44 moléculas de las series interna (17) y externa (27) las que se encuentran fuera del dominio de aplicación del modelo, sugiriendo un comportamiento anómalo en cuanto a la propiedad predicha por lo que sus clasificaciones no deben considerarse confiables. El listado de estos compuestos se muestra en el Anexo 3.

Los compuestos con los valores de *leverage* mayores 1 que resultaron fuera del dominio de aplicación del *random forest* fueron removidos y detectados como *outliers*, excluyéndose 34 compuestos, los que representan solo el 3,2% de los datos iniciales. Los resultados de la aplicación del voto mayoritario que se realizó excluyendo estos 34 compuestos se muestran en la Tabla 3.11.

Tabla 3.10. Resultados de la aplicación del voto mayoritario después de la remoción de las estructuras *outliers*.

Serie	Clase	VP	FP	*Recall*	Precisión	Sensibilidad	Especificidad	Exactitud	Cohen's kappa
	1	72	13	0,79	0,85	0,79	0,80		
	0	52	19	0,80	0,73	0,80	0,79		
SP								**0,80**	**0,58**
	1								
		88	26	0,85	0,78	0,85	0,70		
	0	61	16	0,70	0,79	0,70	0,85		
SVE								**0,78**	**0,532**

SP: serie de predicción; SVE: serie de validación externa; VP: verdaderos positivos; FP: falsos positivos; 0 (baja biodisponibilidad); 1 (alta biodisponibilidad)

Con la remoción de estos 34 compuestos mejoraron ligeramente las estadísticas de todos los modelos individuales y del voto mayoritario, sobre todo los valores de sensibilidad para la clase de baja biodisponibilidad.

La otra aproximación utilizada, basada en el propio algoritmo *random forest* estableció 34 moléculas fuera del dominio de aplicación. El listado de estos compuestos se muestra en el Anexo 4.

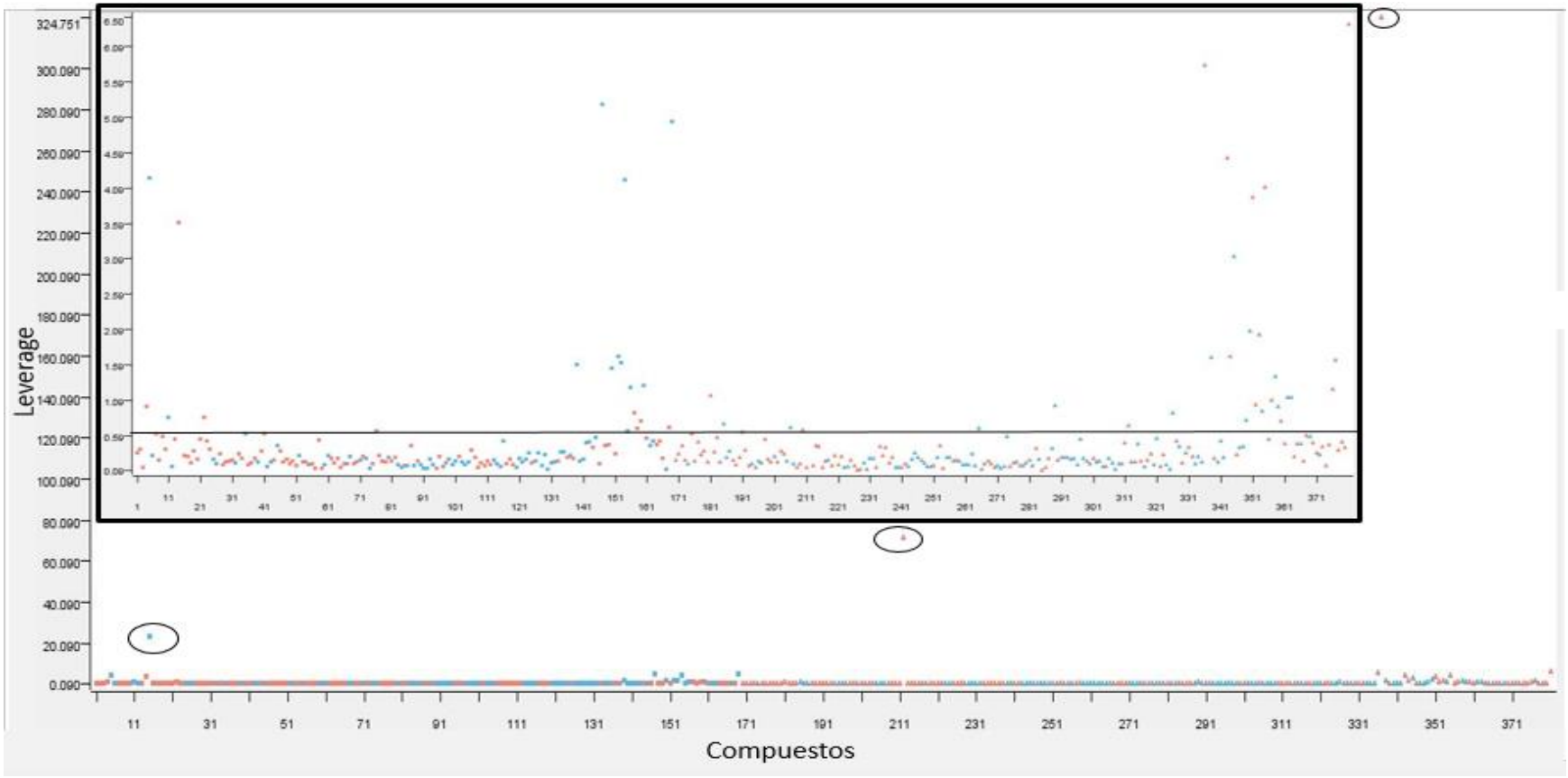


Figura 3.5. *Leverage* vs compuesto. En el cuadro ampliado se muestran todos los puntos excepto los tres puntos señalados con un círculo, que representan los compuestos con los valores de leverage más alejados del límite. Los puntos en rojo indican que son moléculas con alta biodisponibilidad y los puntos en azul que presentan baja biodisponibilidad.

3.7. Predicción de nuevas series externas

Con el objetivo de automatizar la predicción cualquier serie interna se creó una secuencia de nodos que permite: (i) lectura del fichero Excel que contiene las estructuras, preferiblemente en formato SMILE (ii) parametrización de las estructuras moleculares mediante el nodo “alvaDesc”, (iii) normalización de los descriptores moleculares según el modelo de normalización usado para la construcción de los modelos, (iv) predicción de la clase de cada compuesto en función de los modelos individuales, (v) clasificación final de los compuestos en base al voto mayoritario simple, (vi) representación de la distribución de los compuestos según su clase utilizando un histograma.

3.8. Relación entre las variables implicadas en los modelos y la biodisponibilidad

La parametrización de las estructuras moleculares, o en otras palabras, la codificación de la estructura molecular en descriptores matemáticos que “describan” las características de la molécula que influyen en la propiedad en estudio es una parte integral de los estudios QSPR. Una de las finalidades de la modelación QSPR es la interpretación lógica de las variables incluidas, con el objetivo de contribuir a la caracterización y comprensión de la propiedad estudiada.

Para la interpretación de los descriptores moleculares se hace necesario lógicamente determinar cuáles de ellos estuvieron implicados en los modelos. Los algoritmos basados en árboles de clasificación facilitan esta interpretación. Por eso se utilizaron los modelos árbol de decisión y *random forest* para la extracción de las principales variables. Las secciones del flujo de trabajo aplicadas para la extracción de las variables implicadas en estos modelos se muestran en las Figuras 3.11 y 3.14 respectivamente. Como el procedimiento para la extracción de las variables consiste en múltiples iteraciones cada variable está acompañada de una columna definida por la frecuencia con la cual aparece.

En virtud de este valor de frecuencia se seleccionaron las variables **TPSA (total), MLOGP, ALOGP, Hy, JGI6, RBN** y **RBF** para intentar relacionarlas con la biodisponibilidad. La complejidad de esta propiedad es un factor que también dificulta la interpretación de las variables.

El modelo de Yoshida y Topliss define como variable la diferencia entre los valores de logD a los pH 6.5 y 7.4, en ese orden, lo que permite la diferenciación de los compuestos aniónicos de los compuestos neutros o con carga positiva (38). Claramente, el estado de la carga de las moléculas tiene un efecto crítico en la percepción de la molécula por parte de biomoléculas incluyendo membranas, enzimas y transportadores (19).

De acuerdo a Martin y colaboradores las propiedades físicas que determinan la biodisponibilidad y permeabilidad de los compuestos que están negativamente cargados a pH 6-7, difieren de aquellas que gobiernan la biodisponibilidad de los compuestos que son neutros o están positivamente cargados a ese pH. Para aniones la más importante propiedad es el área superficial polar (PSA) mientras que para los otros la regla de los cinco (*rule-of-five*) tiene bastante capacidad predictiva (19).

Por lo tanto, los descriptores que caracterizan la carga de las moléculas o algunas propiedades moleculares dependientes de la carga molecular influirían en alguna medida en la biodisponibilidad de las mismas. En este sentido, entre las variables implicadas en los modelos, el descriptor JG16 apareció en varios de ellos. Esta variable define el índice de carga media topológica de orden 6.Los índices de carga topológica fueron propuestos para evaluar la trasferencia de carga entre pares de átomos y además la carga global transferida en la molécula (78).

Tradicionalmente, el coeficiente de partición y el coeficiente de distribución han estado asociados con varias propiedades ADME. En los modelos construidos varias aproximaciones de estos parámetros, entre ellas: MLOGP (coeficiente de partición octanol-agua, Moriguchi) y el ALOGP aparecieron con frecuencia.

El MLOGP (Coeficiente de partición octanol-agua, Moriguchi) es calculado a partir del modelo de logP de Moriguchi, que consiste en una ecuación de regresión basado en trece parámetros estructurales. Los coeficientes de regresión fueron evaluados tomando como serie de entrenamiento 1230 moléculas orgánicas, incluyendo compuestos alifáticos, aromáticos y heterociclos, conteniendo los siguientes átomos: C, H, N, O, S, P, F, Cl, Br, I. Los parámetros estadísticos de los modelos fueron r^2=0,906 y s=0,422 (78).

El coeficiente de partición octanol-agua desarrollado por Ghose-Crippen-Viswanadhan (ALogP) es calculado a partir de un modelo de regresión basado en la contribución hidrófoba de 115 tipos de átomos. (78). Cada átomo en cada estructura es clasificado en uno de los 115 tipos de átomos (78).

Mediciones de la biodisponibilidad oral en ratas para más de 1100 candidatos a fármacos han permitido establecer relaciones entre las propiedades moleculares y la biodisponibilidad oral. Se ha sugerido que la flexibilidad molecular reducida, como medida del número de enlaces con libre rotación, y una pequeña área superficial topológica o el bajo conteo de enlaces de hidrógeno (puentes de hidrógeno) son importantes predictores de una buena biodisponibilidad oral, independientemente del peso molecular. Esta última observación es controversial, pues el peso molecular es considerado con frecuencia como un prerrequisito (ejemplo, < 500 g/mol como en la Regla de los Cinco de Lipinski) (23).

Como la biodisponibilidad es una combinación de la absorción oral y el metabolismo intestinal y hepático, la aproximación de la biodisponibilidad humana a partir de datos de biodisponibilidad en modelos animales (ejemplo, rata) no es completamente fiable. En particular, las diferencias metabólicas y de transportadores son muy importantes así como la interrelación entre diferentes propiedades moleculares. Las moléculas con alto peso molecular tienden a tener mayor cantidad de enlaces de hidrógeno y potencialmente son más flexibles. En adición a la flexibilidad, la forma total de la molécula debe ser considerada (4).

En relación con la flexibilidad molecular la variable RBN (*Rotatable Bond Number*) describe el número de enlaces con libre rotación. En otras palabras, la cantidad de Csp_3, excluyendo frecuentemente enlaces con libre rotación como -OH y $-CH_3$.También se ha sugerido excluir del conteo los enlaces tipo amido C-N (46).

Como fórmula general para el cálculo del número de enlaces con libre rotación se plantea:

$$RBN = N(nt) + \sum_r (nr - 4) \quad (3.2)$$

Donde, N (nt) es el número de los enlaces con libre rotación no terminales, y n(r) es el número de enlaces simples en un anillo no aromático (46).

La fracción de enlaces con libre rotación, denotada por RBF, es el cociente entre número de enlaces con libre rotación y la cantidad de enlaces de la molécula. Ambas variables RBN y RBF aparecieron en varios de los modelos (46).

Otra de las variables implicadas fue el Factor hidrofílico (Hy). Esto es razonable, dada la influencia de las propiedades lipófilas/hidrófilas de una molécula en su solubilidad y permeabilidad intestinal. El factor hidrofílico es un descriptor de hidrofilia definido como (78):

$$Hy = \frac{(1+N(Hy)*log2\ (1+N(Hy)+nC*\left(\frac{1}{nSK}*log2*\frac{1}{nSK}\right)+\sqrt{\frac{N(hy)}{nSK^2}}}{log2(1+nSK)} \quad (3.3)$$

Donde, N (Hy) es el número de grupos hidrofílos (-OH, -SH, -NH), nC, el número de átomos de carbono y nSK el número de átomos, excluyendo los hidrógenos (78).

Un estudio realizado por Tian y colaboradores mostró que el coeficiente de distribución a pH 5.5 presenta mayor correlación con la biodisponibilidad que los coeficientes de distribución a pH 2.0, 6.5, 7.4 y 10. En este estudio los autores dividieron los compuestos en alta o baja biodisponibilidad fijando como *cutoff* 20% y realizaron un análisis de correlación entre varias propiedades moleculares y la biodisponibilidad oral. Los compuestos con alta biodisponibilidad mostraron una media de logD (5.5) de 0.65, superior al valor medio de logD (5.5) de los compuestos

con baja biodisponibilidad. Un valor elevado de logD (5.5) indica mayor carácter lipófilo, relacionado directamente con la permeabilidad intestinal (18).

Una propiedad muy útil en la predicción de la absorción es el Área de Superficie Polar (PSA, por sus siglas en inglés). Usualmente es definida como aquellas partes de van der Waals o superficies accesibles al solvente de una molécula, que están asociadas con la capacidad de aceptar electrones para la formación de los enlaces de hidrógeno (ejemplo, átomos de nitrógeno u oxígeno) y la capacidad de donar electrones (ejemplo, grupos amido e hidroxilo) (4).

El Área Superficial Polar Topológica (TPSA) es calculada de acuerdo al modelo propuesto por Ertl (23,78). Ertl y colaboradores desarrollaron un método para generar valores de TPSA basados en valores 3D de PSA para 43 fragmentos resultados de un análisis de 34 810 compuestos tomados de la base de datos *World Drug Index* (WDI). La correlación entre los valores de PSA y TPSA es bastante alta (r^2=0,982). El TPSA de una molécula es determinado por la sumatoria de las contribuciones de las áreas superficiales polares de los átomos, y es definido matemáticamente según la ecuación (46,78):

$$TPSA = \sum_i Ni * Gi \qquad (3.4)$$

Donde, Ni es la frecuencia con la cual se encuentra el átomo en la molécula y Gi es la su contribución a la superficie polar. Las contribuciones de cada átomo se encuentran tabuladas (46,78).

Entre las variables seleccionadas implicadas en los modelos se encuentra TPSA (total) y TPSA (NO). La primera considera las contribuciones polares de los átomos de azufre, fósforo, nitrógeno y oxígeno; mientras que TPSA (NO) solo las contribuciones de los dos últimos (46).

En el año 2002 Veber y colaboradores analizaron más de 1100 candidatos a fármacos e identificaron características comunes de compuestos con biodisponibilidad en ratas mayor o igual al 20%. Entre estas características determinaron TPSA (total) ≤ 140 $Å^2$ (20,23).

En la predicción de la Absorción Intestinal Humana también ha sido incluida como variable TPSA. Ejemplo de ello es el estudio de Hou y colaboradores que desarrollaron un modelo de clasificación utilizando GFA (*function genetic algorithm*), MARS (multivariate adaptive regression splines) y RP (*recursive partitioning*) (17).

En un estudio realizado por Wang y colaboradores en el 2008, 74 descriptores fueron calculados usando el software ADRIANA.Code. Entre ellos MW (*molecular weight*), TPSA y MMP (*Mean Molecular Polarizability*). El índice de correlación entre la biodisponibilidad y TPSA para 772 fármacos mostró un valor de -0,236. La variable formó parte de dos de los modelos de regresión construidos con coeficientes de 0,20 y 0,12 (23).

En un artículo publicado en 2011 por Tian y colaboradores, los autores analizaron las relaciones entre la biodisponibilidad oral y 17 descriptores comúnmente usados en estudios ADME. Seleccionaron dos propiedades moleculares entre: MW, TPSA, RBN, Nrule-of-5, logP, logD (5.5) y las cantidades de hidrógenos aceptores y donantes para desarrollar 35 reglas de clasificación. Un resultado interesante de su estudio fue que cuando TPSA y RBN formaban parte de la misma regla, el valor *cutoff* para TPSA era 140 $Å^2$ el mismo que el reportado por Veber y Clark (18).

La mejor regla establecida en dicho estudio incluyó los descriptores TPSA < 160 $Å^2$ y logP > -2.2.En el mejor modelo de regresión lineal múltiple construido utilizando GFA, solo dos propiedades moleculares fueron incluidas logD (6.5) y TPSA como descriptores importantes. El modelo final estuvo conformado por 60 variables, indicativo de que las otras 58 eran *fingerprints* (18).

Conclusiones

Conclusiones

1. Se conformó una base de datos de biodisponibilidad oral amplia y con la diversidad estructural requerida, que puede ser usada por todos aquellos que modelen propiedades ADME sin necesidad previa de curación de las estructuras y valores experimentales.
2. Se obtuvo por primera vez un modelo QSPR semiautomatizado para la predicción de la biodisponibilidad oral en humanos, a través de la combinación de seis modelos de clasificación en un modelo consenso según un voto mayoritario simple, obteniéndose una capacidad predictiva mayor que 77% valor superior a los modelos reportados en la literatura.
3. Los modelos QSPR obtenidos, utilizando los descriptores moleculares del “alvaDesc” y combinando diversas técnicas estadísticas y de inteligencia artificial permitieron describir la biodisponibilidad oral en humanos.

Recomendaciones

Recomendaciones

1. Aplicar nuevas técnicas de modelación, implementado otras modalidades de modelos consenso y combinando más herramientas estadísticas y de inteligencia artificial.
2. Desarrollar modelos de clasificación de la biodisponibilidad a través de la predicción de propiedades relacionadas con los procesos de absorción, distribución, metabolismo y excreción del fármaco e integrarlos en la predicción de la biodisponibilidad utilizando los estudios QSPR y/o las reglas simples.
3. Desarrollar modelos fisiológicos (PBPK) para la simulación y predicción de la absorción intestinal, el metabolismo y la biodisponibilidad oral mediante la combinación de datos *in silico*, *in vitro* e *in vivo*.

Referencias bibliográficas

Referencias bibliográficas

1. Pritchard JF, Jurima-Romet M, Reimer MLJ, Mortimer E, Rolfe B, Cayen MN, et al. Making better drugs: Decision Gates in non-clinical drug development. Nat Rev Drug Discov. 2003;2:542-53.
2. Trapotsi M-A. Development and evaluation of ADME models using proprietary and opensource. University of Hertfordshire (Thesis); 2017.
3. Cabrera-Pérez MÁ, Pham-The H. Computational modeling of human oral bioavailability:what will be next? Expert Opin Drug Discov. 2018;1-13.
4. Waterbeemd H van de, Testa B. The Why and How of Drug Bioavilability Research. En: van de Waterbeemd H, Testa B, editores. Drug Bioavailability Estimation of Solubility, Permeability, Absorption and Bioavailability. Weinheim: Wiley-VCH; 2009. p. 1-3.
5. Igor V, Tudor IO. Early ADME/T Predictions Toy or Tool? Chemoinformatics Approaches to Virtual Screen. RSC Publishing; 2008;45:240-60.
6. Winiwarter S, Ahlberg E, Watson E, Oprisiu I, Noeske T, Greene N. In silico ADME in drug design, enhancing the impact. ADMET DMPK. 2018;6(1):15-33.
7. Hurst S, Loi C, Brodfuehrer J, El-kattan A. Impact of physiological , physicochemical and biopharmaceutical factors in absorption and metabolism mechanisms on the drug oral bioavailability of rats and humans. Expert Opin Drug Metab Toxicol. 2007;3(4):469-90.
8. El-Kattan AF. Oral bioavailability assessment. Basics and Strategies and Development. New Jersey: Wiley; 2017. 35-40 p.
9. Zaragoza F. The Drug Discovery Process. En: Lead Optimization for Medicinal Chemist. Weinheim: Wiley-VCH; 2012. p. 3-11.
10. Waterbeemd H van de, Gifford E. ADME/T in silico modeling:Towards prediction paradise? Nat Publ Gr. 2003;2:192-204.
11. Vergnaud J-M, Rosca I-D. Assessing Bioavailability of Drug Delivery Systems. Mathematical Modeling. Taylor & Francis; 2005.
12. Hellriegel ET, Bjornsson TD, Hauck WW. Interpatient variability in bioavailability is related to the extent of absorption: implications for bioavailability and bioequivalence studies. Clin Pharmacol Ther. 1996;60(6):601-7.
13. González I, Cabrera-Pérez MÁ, Bermejo M. Metodologías Biofarmacéuticas en el desarrollo de medicamentos. Hernández UM, editor. Universitas Miguel Hernández; 2015.
14. Norinder U. Computational Approaches to Drug Absorption and Bioavailability. En: Waterbeemd H van de, Testa B, editores. Drug Bioavailability Estimation of Solubility, Permeability, Absorption and Bioavailability. Weinheim: Elsevier; 2009.
15. Livingstone DJ, Waterbeemd H van de. In Silico Prediction of Human Bioavailability. En: van de Waterbeemd H, Testa B, editores. Drug Bioavailability Estimation of Solubility, Permeability, Absorption and Bioavailability. Wiley-VCH; 2009. p. 437-9.
16. Olivares-Morales A, Hatley OJD, Turner D, Galetin A, Aarons L, Rostami-hodjegan A. The Use of ROC Analysis for the Qualitative Prediction of Human Oral Bioavailability from Animal Data. Pharm Res. 2014;31(3):720-30.

17. Bermejo M, Álvarez IG, Álvarez MG, Garrigues TM. QSPR in Oral Bioavailability:Specificity or Integrality? Mini-Reviews Med Chem. 2012;12(6):534-50.
18. Tian S, Li Y, Wang J. ADME evaluation in drug discovery. Prediction of oral bioavailability in humans based on molecular properties and structural fingerprints. Mol Pharm. 2011;8(3):841-51.
19. Martin YC. A Bioavailability Score. J Med Chem. American Chemical Society; 2005;48:3164-70.
20. Zhu J, Wang J, Yu H, Li Y, Hou T. Recent Developments of In Silico Predictions of Oral Bioavailability. Comb Chem High Throughput Screen. 2011;362-74.
21. Wang J, Hou T. Advances in computationally modeling human oral bioavailability. Adv Drug Deliv Rev. 2015;86:1-6.
22. Imawaka H, It K, Kitamura Y. Prediction of human bioavailability from human oral administration data and animal pharmacokinetic data without data from intravenous administration of drugs in humans. Pharm Res. 2009;26(8):1881-9.
23. Wang Z, Yan A, Yuan Q, Gasteiger J. Explorations into modeling human oral bioavailability. Eur J Med Chem. Elsevier Masson SAS; 2008;43(11):2442-52.
24. Suenderhauf C, Parrott N. A Physiologically Based Pharmacokinetic Model of the Minipig:Data Compilation and Model Implementation. Pharm Res. 2013;30:1-15.
25. Schmitt W, Willmann S. Physiology-based pharmacokinetic modeling: ready to be used. Drug Discov Today Technol. 2005;2(1):125-32.
26. Jamei M, Turner D, Yang J, Neuhoff S, Polak S, Rostami-hodjegan A, et al. Population-Based Mechanistic Prediction of Oral Drug Absorption. AAPS J. 2009;11(2):225-33.
27. Sager JE, Yu J, Ragueneau-Majlessi I, Isoherranen N. Minireview Physiologically Based Pharmacokinetic (PBPK) Modeling and Simulation Approaches : A Systematic Review of Published Models , Applications , and Model Verifications. Drug Met Dispos. 2015;43:1823-37.
28. Wang Y, Xing J, Xu Y, Zhou N, Peng J, Xiong Z, et al. In silico ADME / T modelling for rational drug design. Q Rev Biophys. 2015;48(4):488-515.
29. Paixão P, Gouveia LF, Morais JA-G. Prediction of the human oral bioavailability by using in vitro and in silico drug related parameters in a physiologically based absorption model. Int J Pharm. Elsevier B.V.; 2012;429(1-2):84-98.
30. Andrews CW, Bennett L, Yu LX. Predicting human oral bioavailability of a compound: development of a novel quantitative structure bioavailability relationship. Pharm Res. 2000;17(6):639-44.
31. Turner J V, Glass BD, Agatonovic-Kustrin S. Prediction of drug bioavailability based on molecular structure. Anal ChimActa. 2003;485:89-102.
32. JV T, DJ M. Bioavailability prediction based on molecular structure for a diverse series of drugs. Pharm Res. 2004;21(1):68-82.
33. Wang J, Krudy G, Xie XQ. Genetic algorithm-optimized QSPR models for bioavailability, protein binding, and urinary excretion. J Chem Inf Model. 2006;46(6):2674-83.
34. Moda TL, Montanari A, Andricopulo AD. Hologram QSAR model for the prediction of human oral bioavailability. BioorgMedChem. 2007;

35. Xu X, Zhang W, Huang C, Li Y, Yu H, Wang Y. A Novel Chemometric Method for the Prediction of Human Oral Bioavailability. Int J Mol Sci. 2012;13:6964-82.
36. Pintore M, Waterbeemd H van de, Piclin N, Chretien JR. Prediction of oral bioavailability by adaptive fuzzy partitioning. Eur J Med Chem. 2003;38:427-31.
37. Ma CY, Yang SY, Zhang H. Prediction models of human plasma protein binding rate and oral bioavailability derived by using GA-CG-SVM method. J Pharm Biomed Anal. 2008;47(4-5):677-82.
38. Yoshida F, Topliss JG. QSAR model for drug human oral bioavailability. J Med Chem. 2000;43(13):2575-85.
39. Ahmed S, Ramakrishna V. Systems biological approach of molecular descriptors connectivity: optimal descriptors for oral bioavailability prediction. PLoS One. 2012;7(7):1-10.
40. Kim MT, Sedykh A, Chakravarti SK, Saiakhov RD, Zhu H. Critical Evaluation of Human Oral Bioavailability for Pharmaceutical Drugs by Using Various Cheminformatics Approaches. Pharm Res. 2014;31:1002-14.
41. OECD Principles for the validation,for regulatory purposes of Quantitative-Structure-Activity-Relationship-Models. Paris: Environment Directorate Organization for Economic Coperation and Development; 2007.
42. Cherkasov A, Muratov EN, Fourches D, Varnek A, Baskin II, Cronin M, et al. QSAR Modeling .Where have you been? Where are you going to? J Med Chem. 2014;57(12):4977-5010.
43. Sahigara F, Mansouri K, Ballabio D, Mauri A, Consonni V, Todeschini R. Comparison of Different Approaches to Define the Applicability Domain of QSAR Models. Molecules. 2012;(December).
44. Gramatica P. Principles of QSAR models validation:internal and external. QSAR Comb Sci. 2007;26(5):694-701.
45. Validation of QSAR Models. En: Understanding the Basics of QSAR for Applications in Pharmaceutical Sciences and Risk Assessment experiments. Elsevier; 2015. p. 236-67.
46. Todeschini R, Consonni V. Molecular Descriptors for Chemoinformatics. Wiley-VCH; 2009.
47. Waterbeemd H. Chemometric methods in molecular design. New York: Wiley-VCH; 1995.
48. Roy K, Kar S. A Primer on QSAR/QSPR Modeling.Fundamental Concepts. Springer; 2015.
49. Dehmer M, Varmuza K. Statistical Modelling of molecular Descriptors in QSAR/QSPR. Weinheim: Wiley-VCH; 2012.
50. Roy K. Statistical Methods in QSAR/QSPR. En: A Primer on QSAR/QSPR Modeling. Springer; 2015.
51. Alpaydin E. Introduction to machine learning. London: MIT Press; 2004. 1-35 p.
52. Le H. Cribado Virtual para el Descubrimiento de Potentes Inhibidores de la Enzima Tirosinasa. UCLV (tesis); 2011.
53. Liew CY, Ma XH, Liu X, Yap CW. Logistic regression. J Chem Inf Model. 2009;49:877-85.
54. Basak D, Pal S, Patranabis DC. Support Vector Regression. Neural Inf Process – Lett Rev. 2007;11(10):203-24.

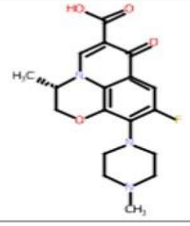
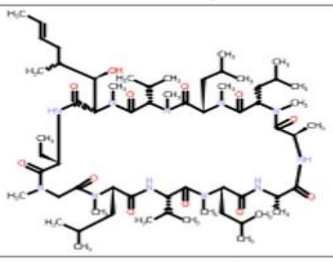
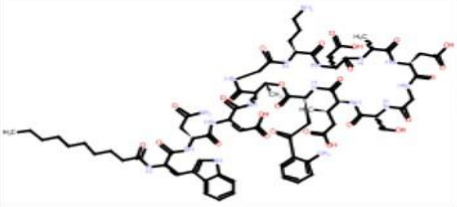
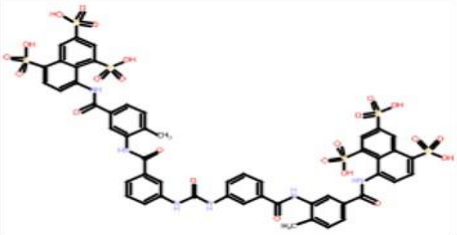
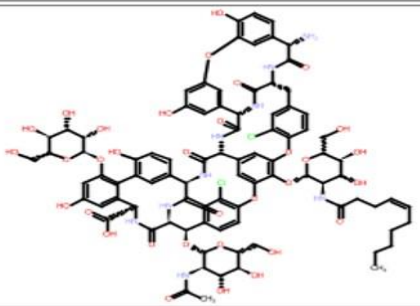
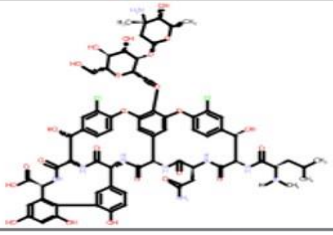
55. Witten I, Franj E. Data Mining.Practical Machine Learning Tools and Techniques. Elsevier; 2005.
56. Burbidge R, Trotter M, Buxton B, Holden S. Drug design by machine learning: support vector machines for pharmaceutical data analysis. Comput Chem. 2001;26:5-14.
57. Lariviere B, Van Den Poel D. Decision tree and random forest. Expert Syst Appl. 2005;29:472-84.
58. Breiman L. Random forest. California: Elsevier; 2001. p. 1-33.
59. Scornet E. A random forest guided tour. 2016;25(2):197-227.
60. Schmdhuber J. Deep learning in neural networks: An overview. Neural Networks. 2015;61:85-117.
61. Tsay RS, Chen R. Neural networks, deep learning and tree-based methods. En: Nonlinear Time Series Analysis. 2019. p. 173-216.
62. Tufféry S. Neural Networks. En: Data Mining and Statistics for Decision Making. John Wiley & Sons; 2011. p. 217-9.
63. Hand D, Mannila H, Smyth P. Principles of Data Mining. MIT Press; 2001. 45-48 p.
64. Jovi A, Brki K, Bogunovi N. An overview of free software tools for general data mining. Faculty of Electrical Engineering and Computing, University of Zagreb / Department of Electronics, Microelectronics, Computer and Intelligent Systems (report); 2015.
65. Bonthu S. Review of Leading Data Analytics Tools. Int J Eng Tehnol. 2017;7(3.31):10-5.
66. Chauhan N, Gautam N. Parametric Comparison of Data Mining Tools. Int J Adv Technol Eng Sci. 2015;3(11):291-8.
67. Mazanetz PM, Marmon JR, Reisser BTC, Morao I. Drug Discovery Applications for KNIME: An Open Source Data Mining Platform. Curr Top Med Chem. 2012;12(18):1965-79.
68. Tsoumanis A, Melagraki G, Afantitis A. Enalos KNIME nodes: exploring environmental durability , ageing resistance and corrosion inhibition. 2013;5:2013.
69. Filippov I V, Pugliese A. Computational tools and resources for metabolism-related property predictions.Application to prediction of half-life time in human liver microsomes. 2000;1933-44.
70. Legehar A, Xhaard H, Ghemtio L. IDAAPM: Integrated database of ADMET and adverse effects of predictive modeling based on FDA approved drug data. J Cheminform. Springer International Publishing; 2016;8(1):1-11.
71. Kausar S, Falcao AO. An automated framework for QSAR model building. J Cheminform. 2018;10(1):1-23.
72. Varma M, Obach R, Rotter C. Physicochemical space for optimum oral bioavailability: contribution of human intestinal absorption and first-pass elimination. J Med Chem. 2010;53:1098-108.
73. Bertz RJ, Granneman GR. Use of In Vitro and In Vivo Data to Estimate the Likelihood of Metabolic Pharmacokinetic Interactions. ClinPharmacokinetic. 1997;32(3):210-58.
74. Sietsema WK. The absolute oral bioavailability of selected drugs. Int J Clin

Pharmacol. 1989;27(4):179-211.
75. Goodman L., Gilman A. The Pharmacological Basis of Therapeutics. 9.ª ed. New York: McGraw-Hill Publishers; 1996.
76. Fourches D, Muratov E, Tropsha A, Hill C. Trust,But Verify II: A Practical Guide to Chemogenomics Data Curation. J Chem Inf Model. 2017;56(7):1243-52.
77. Chawla N V. Data mining for imbalanced datasets:an overview. En: Dai D, editor. Data Mining and Knowledge Discovery Handbook. Elsevier; 2005. p. 854-63.
78. Todeschini R, Consonni V. Handbook of Molecular Descriptors. Mannhold R, Kubinyi H, Timmerman H, editores. Weinheim: Wiley-VCH; 2000.

Anexos

Anexos

Anexo 1. Estructuras de los compuestos removidos durante la curación de las estructuras moleculares

Nombre	CAS	Smile canónico (CDK)	Motivo de la remoción
Ofloxacin	82419-36-1		Duplicado
Cyclosporine_A	59865-13-3		Masa molar mayor que 1200 g/mol
Daptomycin	103060-53-3		Masa molar mayor que 1200 g/mol
Suramin	145-63-1		Masa molar mayor que 1200 g/mol
Teicoplanin_A2-1	91032-34-7		Masa molar mayor que 1200 g/mol
Vancomycin	1404-90-6		Masa molar mayor que 1200 g/mol

Anexo 2. Compuestos fuera del dominio de aplicación del árbol de decisión según el método basado en el *leverage.*

Compuesto	CAS	*Leverage*	Valor límite de *leverage*	Serie
Amantadine	768-94-5	0,119	0,115	predicción interna
Pentosan Polysulfate	37300-21-3	0,270	0,115	predicción interna
Topiramate	97240-79-4	0,236	0,115	predicción interna
Hypericin	548-04-9	0,390	0,115	predicción interna
Pancuronium	16974-53-1	0,159	0,115	predicción interna
Dimercaprol	59-52-9	0,347	0,115	predicción interna
Alcuronium	23214-96-2	0,254	0,115	predicción interna
Rifapentine	61379-65-5	0,133	0,115	predicción interna
Nystatin	1400-61-9	0,192	0,115	predicción interna
Sodium_oxybate	502-85-2	0,128	0,115	predicción interna
Colistin	1066-17-7	0,353	0,115	predicción interna
Atracurium	64228-79-1	0,207	0,115	predicción interna
Mivacurium	106791-40-6	0,241	0,115	predicción interna
Allopurinol	315-30-0	0,134	0,115	predicción interna
Clodronate	10596-23-3	0,408	0,115	predicción interna
Thyroxine	51-48-9	0,147	0,115	externa
Raffinose	512-69-6	0,181	0,115	externa
Acarbose	56180-94-0	0,221	0,115	externa
Acetohydroxamic Acid	546-88-3	0,220	0,115	externa
Ethanol	64-17-5	1,018	0,115	externa
Probucol	23288-49-5	0,219	0,115	externa
Cycloserine	68-41-7	0,180	0,115	externa
Glyceryl-1-nitrate	624-43-1	0,116	0,115	externa
Fosfomycin	23155-02-4	0,208	0,115	externa
Desmopressin	16679-58-6	0,195	0,115	externa
Amphotericin_B	1397-89-3	0,197	0,115	externa
Methimazole	60-56-0	0,165	0,115	externa
Thioguanine	154-42-7	0,118	0,115	externa
Pentaerythritol_tetranitrate	78-11-5	0,367	0,115	externa
Hydroxyurea	127-07-1	0,268	0,115	externa

Anexo 3. Compuestos fuera del dominio de aplicación del *random forest* según el método basado en el *leverage*

Nombre	CAS	*Leverage*	Límite de leverage	Serie
Amantadine	768-94-5	*1,003*	0,746	predicción interna
Pentosan Polysulfate	37300-21-3	*4,223*	0,746	predicción interna
Bretylium	59-41-6	*0,846*	0,746	predicción interna
Thiocyanate	302-04-5	*3,597*	0,746	predicción interna
Diatrizoate	117-96-4	*23,089*	0,746	predicción interna
Topiramate	97240-79-4	*0,848*	0,746	predicción interna
Hypericin	548-04-9	*1,586*	0,746	predicción interna
Dimercaprol	59-52-9	*5,265*	0,746	predicción interna
Alcuronium	23214-96-2	*1,538*	0,746	predicción interna
Nystatin	1400-61-9	*1,706*	0,746	predicción interna
Sodium_oxybate	502-85-2	*1,610*	0,746	predicción interna
Colistin	1066-17-7	*4,200*	0,746	predicción interna
Mivacurium	106791-40-6	*1,271*	0,746	predicción interna
Alafosfalin	60668-24-8	*0,903*	0,746	predicción interna
Isoniazid	54-85-3	*0,789*	0,746	predicción interna
Clomethiazole	533-45-9	*1,293*	0,746	predicción interna
Clodronate	10596-23-3	*5,021*	0,746	predicción interna
Mecamylamine	60-40-2	*1,148*	0,746	externa
Rasagiline	136236-51-6	*0,759*	0,746	externa
Thyroxine	51-48-9	*71,454*	0,746	externa
Adefovir Dipivoxil	142340-99-6	*1,012*	0,746	externa
Acarbose	56180-94-0	*0,913*	0,746	externa
Acetohydroxamic Acid	546-88-3	*5,825*	0,746	externa
Ethanol	64-17-5	*324,751*	0,746	externa
Probucol	23288-49-5	*1,686*	0,746	externa
Cycloserine	68-41-7	*4,513*	0,746	externa
Glyceryl-1-nitrate	624-43-1	*1,709*	0,746	externa
Fosfomycin	23155-02-4	*3,121*	0,746	externa
Desmopressin	16679-58-6	*0,806*	0,746	externa
Amphotericin_B	1397-89-3	*2,062*	0,746	externa
Methimazole	60-56-0	*3,962*	0,746	externa
Acetazolamide	59-66-5	*1,018*	0,746	externa
Amifostine	20537-88-6	*2,015*	0,746	externa
Thioguanine	154-42-7	*0,932*	0,746	externa
Pentaerythritol_tetranitrate	78-11-5	*4,105*	0,746	externa
Busulfan	55-98-1	*1,087*	0,746	externa
Isosorbide_dinitrate	87-33-2	*1,423*	0,746	externa
Carbimazole	22232-54-8	*0,996*	0,746	externa

Ethosuximide	77-67-8	*0,789*	0,746	externa
Carnitine	461-06-3	*1,128*	0,746	externa
Levocarnitine	541-15-1	*1,128*	0,746	externa
Chloroxine	773-76-2	*1,244*	0,746	externa
Coumarin	91-64-5	*1,657*	0,746	externa
Hydroxyurea	127-07-1	*6,399*	0,746	externa

Anexo 4. Compuestos fuera del dominio de aplicación según el método basado en el *random forest.*

Nombre	CAS	Serie
Amantadine	768-94-5	predicción interna
Pentosan Polysulfate	37300-21-3	predicción interna
Thiocyanate	302-04-5	predicción interna
Diatrizoate	117-96-4	predicción interna
Topiramate	97240-79-4	predicción interna
Perhexiline	6621-47-2	predicción interna
Distigmine	15876-67-2	predicción interna
Hypericin	548-04-9	predicción interna
Dimercaprol	59-52-9	predicción interna
Alcuronium	23214-96-2	predicción interna
Nystatin	1400-61-9	predicción interna
Sodium_oxybate	502-85-2	predicción interna
Colistin	1066-17-7	predicción interna
Mivacurium	106791-40-6	predicción interna
Allopurinol	315-30-0	predicción interna
Clomethiazole	533-45-9	predicción interna
Clodronate	10596-23-3	predicción interna
Mecamylamine	60-40-2	externa
Adefovir Dipivoxil	142340-99-6	externa
Acarbose	56180-94-0	externa
Acetohydroxamic Acid	546-88-3	externa
Ethanol	64-17-5	externa
Probucol	23288-49-5	externa
Cycloserine	68-41-7	externa
Desmopressin	16679-58-6	externa
Amphotericin_B	1397-89-3	externa
Methimazole	60-56-0	externa
Pentaerythritol_tetranitrate	78-11-5	externa
Isosorbide_dinitrate	87-33-2	externa
Carnitine	461-06-3	externa
Levocarnitine	541-15-1	externa
Famotidine	76824-35-6	externa
Chloroxine	773-76-2	externa
Hydroxyurea	127-07-1	externa