UNIVERSIDAD NACIONAL DE LA PLATA

FACULTAD DE CIENCIAS EXACTAS

DEPARTAMENTO DE CIENCIAS BIOLÓGICAS

Trabajo de Tesis Doctoral:

DESARROLLO DE FILTROS IN SILICO ORIENTADOS A OPTIMIZAR EL DESARROLLO DE NUEVOS FÁRMACOS DESTINADOS AL SISTEMA NERVIOSO CENTRAL

Tesista: Farm. Juan Francisco Morales

Director/a: Dr. Alan Talevi

Codirector/a: Dra. María Esperanza Ruiz

Año: 2022

AGRADECIMIENTOS

Quisiera agradecer a las personas e instituciones que ayudaron directa o indirectamente a la realización de este trabajo de Tesis Doctoral:

A mi director, el Dr. Alan Talevi, y a mi co-directora, la Dra. María Esperanza Ruiz, por la paciencia y su acompañamiento incondicional durante todo el doctorado, por sus importantes aportes a mi formación con toda su experiencia, y por hacer de mí un mejor profesional y persona.

Al CONICET, por otorgarme la beca para el desarrollo de este trabajo de Tesis Doctoral.

Al Laboratorio de Investigación y Desarrollo de Bioactivos (LIDeB) y la Facultad de Ciencias Exactas de la UNLP por otorgar el lugar donde se pudieron llevar a cabo las tareas de investigación.

A todos mis compañeros del LIDeB: Sebas, Lucas, Giuli, Caro, Malala, Julián, Juani, Melisa, Sarita, Luciana, Manu, Andre, Chaco y a todas aquellas personas con las que he compartido y vivido momentos, por las charlas, anécdotas y tiempo compartido dentro y fuera del laboratorio.

A mis padres, mis hermanas y mi familia, que con su infinito esfuerzo e inagotable paciencia, me dieron todas las herramientas necesarias para poder desarrollarme en todos los aspectos de la vida.

Índice

1. Introducción 1

1.1. Enfermedades del sistema nervioso central 1

1.2. Investigación y desarrollo de nuevos fármacos 2

1.3. Descubrimiento de fármacos para del SNC 3

1.4. Hipótesis del fármaco libre y barrera hematoencefálica 5

1.5. Parámetros que describen la distribución de fármacos entre el cerebro y el plasma 7

1.6. Cuantificación del transporte de drogas a través de la BHE 10

Referencias 14

2. Modelado QSAR 25

2.1. Descubrimiento de fármacos: breve reseña histórica 25

2.2. Métodos computacionales en el descubrimiento de fármacos 29

2.3. Descriptores moleculares 32

2.4. Relación estructura-actividad cuantitativa (*Quantitative Structure–Activity Relationship*, QSAR) 35

2.5. Breve revisión de los modelos *in silico* de $K_{p,uu}$ desarrollados hasta el momento 40

2.6. Objetivos 44

Referencias 45

3. Metodologías de modelado 56

3.1. Bases de datos 57

3.1.1. Recopilación y curado de las bases de datos de la propiedad de interés 57

3.1.2. Partición de los conjuntos de datos en los conjuntos de entrenamiento y de prueba 61

3.1.3. Cálculo de descriptores 65

3.2. Modelado. Generación de modelos - Métodos de modelado 66

3.2.1. *k*-Vecinos más cercanos (kNN) 68

3.2.2. Cuadrados mínimos parciales clasificatorios (cPLS) 69

3.2.3. Máquinas de Soporte Vectorial (SVM) 71

3.2.4. Bosques aleatorios (RF) 77

3.2.5. Máquina de Potenciación por Gradiente Estocástico (sGBM) 79

3.2.6. Potenciación por Gradiente Extremo (XGBOOST)....83
3.2.7. Redes Neuronales Profundas (DNN)....84
3.2.8. Método interno (*in-house*) de modelado -aprendizaje por ensamblado- basado en subespacios aleatorios....87
3.3. Evaluación del poder explicativo de los modelos....88
3.4. Validación interna o computacional de los modelos....90
3.5. Medición de la importancia de las variables....90
3.6. Cálculo del dominio de aplicación de los modelos desarrollados....91
Referencias....93

4. Validación de modelos....102
4.1. Validación externa o experimental de los modelos desarrollados con el set de datos MSH....102
4.1.1. Determinación de la fracción libre en plasma....103
4.1.2. Fracción libre en cerebro por el método del homogenato....110
4.1.3. Cálculo del parámetro farmacocinético $K_{p,uu}$....113
4.2. Validación externa de los modelos desarrollados con el set de datos MS....116
4.3. Métodos analíticos por HPLC....116
Referencias....118

5. Resultados y discusión....122
5.1. Resultados en el conjunto de datos MSH....124
5.1.1. Conjunto de datos....124
5.1.2. Partición del conjunto de datos en los conjuntos de entrenamiento y de prueba....128
5.1.3. Cálculo de descriptores....129
5.1.4. Modelos....129
5.1.5. Dominio de aplicación....132
5.1.6. Validación experimental....132
5.2. Resultados en el conjunto de datos MS....134
5.2.1. Conjunto de datos y partición....134
5.2.2. Cálculo de descriptores....138
5.2.3. Modelos....138
5.2.4. Dominio de aplicación....141
5.2.5. Validación externa o experimental....141
5.2.6. Resultados en el conjunto de datos MS refinado....142
5.3. Discusión....146

Referencias 158

6. Conclusiones 162

7. Anexos 166

Publicaciones realizadas, becas obtenidas y presentaciones a congresos durante el período en el que se desarrolló la presente tesis doctoral 166

Detalle de la base de datos 170

Capítulo 1

Introducción

1.1. Enfermedades del sistema nervioso central

Debido a múltiples factores (entre ellos, por ejemplo, la mayor expectativa de vida y la adopción de estilos de vida que fomentan la ocurrencia de trastornos afectivos y del estado del ánimo), los desórdenes del sistema nervioso central (SNC) están teniendo cada vez mayor relevancia en la sociedad actual.

En 2012, Murray y colaboradores (Murray et al., 2012) estudiaron las enfermedades con mayor impacto en la población mundial en términos de los años de vida ajustados por discapacidad (AVaD). Esta medida se obtiene por la sumatoria de los años de vida perdidos debido a la mortalidad prematura (AVP), más los años vividos con discapacidad (AVD) (puede pensarse que un AVaD es un año de vida "saludable" perdido). En el mencionado estudio se compararon los AVaD del año 1990 contra los del año 2010, discriminando por enfermedad. Luego del análisis de los datos, se observó que el AVaD total ha disminuido un 23%, lo cual sugiere que en términos generales el estado de salud de la población mundial mejoró en el período considerado. Sin embargo, los AVaD debidos a los desórdenes neurológicos y a

trastornos mentales y del comportamiento han aumentado 16,7% y 5,9%, en ese orden, contrariamente a la tendencia general. De hecho, en el año 2010, el 3,0% de los AVaD globales surgían de trastornos neurológicos; si sumamos los trastornos mentales y del comportamiento, la proporción asciende a 7,4% de los AVaD totales. Si los tomamos en conjunto, el impacto de los desórdenes del SNC -hablando, estrictamente, en términos de AVaD- es comparable al de todas las neoplasias y similar al de las enfermedades cardiovasculares y circulatorias, que representan un 7,6% y 11,8%, de los AVaD globales, respectivamente.

En otro trabajo del mismo año, Vos *et al.* (Vos et al., 2012) realizaron un análisis similar pero solo con los datos de los AVD. Cuando efectuaron un análisis en función de la edad, observaron que a partir de los 10 años y extendiéndose hasta los 65 años, los trastornos mentales y de conducta fueron la causa más importante de discapacidad, contribuyendo hasta un 36% de los AVD totales en el rango etario de 20 a 29 años. Ocho trastornos del SNC se posicionan entre las 25 principales causas mundiales de AVD. Estos fueron, ordenado de mayor a menor importancia: trastorno depresivo mayor, trastorno de ansiedad, migraña, esquizofrenia, desorden bipolar, distimia, epilepsia y enfermedad de Alzheimer. Los autores concluyeron que los mayores aportantes a los AVD globales fueron los trastornos mentales y de conducta. Dicha tendencia se corrobora en trabajos más recientes (James et al., 2018; Kyu et al., 2018). Estos resultados implican que las enfermedades que afectan al SNC deben ser especialmente tenidas en cuenta por los sistemas de salud de los diferentes países, para así tomar medidas acordes y lograr disminuir su prevalencia, pudiendo evitar las consecuencias sociales, clínicas y económicas que estas conllevan.

1.2. Investigación y desarrollo de nuevos fármacos

La importancia de la Investigación y Desarrollo (I&D) para la industria farmacéutica se evidencia en la inversión que el sector acumula en I&D. El gasto total mundial en I&D de las compañías farmacéuticas y de biotecnología aumentó de USD 108 mil millones en 2006 a USD 141 mil millones en 2015 (Schuhmacher et al., 2016). Entre las 50 empresas más importantes del mundo por la inversión total en I&D en el año fiscal 2014/2015 se encuentran 16 compañías farmacéuticas. Novartis, Roche,

Johnson & Johnson y Pfizer se ubicaron en el top 10 de las principales empresas de inversión en I&D a nivel mundial (Schuhmacher et al., 2016). Sin embargo, la baja tasa de éxito persiste como problemática en el área del descubrimiento de fármacos. Diferentes trabajos han estimado que solo 1 de cada 10 nuevos fármacos que ingresan a fase clínica alcanza el mercado farmacéutico (Kola et al., 2004; Smietana et al., 2016; Wong et al., 2019). Esto último se contrapone al alto nivel de inversión del sector. El número de nuevos fármacos aprobados por cada mil millones de dólares gastados en I&D se ha reducido a la mitad cada 9 años desde 1950, cayendo 80 veces si las cifras se ajustan por la inflación (Scannell et al., 2012). A esta tendencia los autores la han llamado "Ley de Eroom", en contraste con la Ley de Moore, la cual habla del aumento exponencial en el número de transistores que se pueden colocar a un costo razonable en un circuito integrado. El término se usa generalmente para tecnologías que mejoran exponencialmente con el tiempo, ya que el número de transistores se duplicó cada 2 años desde la década de 1970 hasta 2010. Los comportamientos opuestos descritos por la Ley de Moore y la Ley de Eroom podrían relacionarse con la complejidad y la limitada comprensión actual de los sistemas biológicos, versus la simplicidad relativa y un mayor nivel de comprensión de la física del estado sólido. La tendencia negativa en la rentabilidad de las compañías farmacéuticas ha llevado a cuestionar sus estrategias generales y la sustentabilidad del modelo comercial de las mismas (Schnorrenberg, 2018).

1.3. Descubrimiento de fármacos para del SNC

En el caso de las enfermedades que afectan al SNC, la baja tasa de éxitos en la fase clínica empeora con respecto a la situación presentada en el apartado anterior, siendo la misma inferior al promedio si se compara con otras áreas de la terapéutica (Kesselheim et al., 2015; Kola et al., 2004). La tasa de éxito (entendido como obtener la aprobación del ente regulatorio correspondiente) es menos de la mitad que aquella observada para fármacos no relacionados al SNC en el período 1995-2007 (6,2% frente a 13,3%, respectivamente) (Gribkoff et al., 2017). Sumado a este mayor riesgo de fracaso, también nos encontramos con tiempos de desarrollo más largos. En Estados Unidos, el tiempo del ensayo clínico total más el tiempo de revisión por parte de la autoridad sanitar (FDA, *Food and Drug Administration*) para los fármacos

destinados al SNC aprobados entre 1996 y 2010 fue, en promedio, 32 meses más que para los fármacos de acción periférica. Este tiempo adicional consume parte de la vida útil de la patente, lo que resulta en períodos más cortos de protección antes de la entrada de la competencia de medicamentos genéricos (Choi et al., 2014). Todas estas cuestiones que hemos remarcado han generado una disminución de la inversión en proyectos para fármacos destinados al SNC (Kesselheim et al., 2015). De hecho, en el período entre 2009 y 2014, hubo una disminución del 52% en el número total de programas de I&D de fármacos del SNC en las grandes empresas farmacéuticas (Yokley et al., 2017).

Recapitulando, nos encontramos en un escenario en el que los desórdenes del SNC se encuentran en ascenso a nivel global pero en el que, paradójicamente, la inversión en I&D de fármacos destinados a su tratamiento se encuentra a la baja.

Las causas que explican la baja la tasa de éxito en fase clínica de un fármaco en desarrollo son diversas. Dentro de estas encontramos, como las más relevantes, la falta de eficacia, la falta de seguridad, razones de estrategia comercial y cuestiones farmacocinéticas. Ante estas problemáticas, la industria farmacéutica ha reaccionado tomando diferentes medidas, que en algunos casos resultaron exitosas. Por ejemplo, la tasa de fracaso debido a problemas farmacocinéticos se redujo de aproximadamente 40% a tan solo un 10% entre 1991 y 2000 (Kola et al., 2004), lo que puede atribuirse a la incorporación de ensayos y modelos *in vitro* e *in silico* de absorción, distribución, metabolismo y excreción (ADME) en las etapas tempranas del descubrimiento y desarrollo de fármacos (Kassel, 2004; Shih et al., 2017; J. Wang, 2009; Wishart, 2007). En trabajos más recientes se sigue observando la tendencia en la disminución de la tasa de fracaso debido a problemas farmacocinéticos (Morgan et al., 2018; Waring et al., 2015).

Del mismo modo, las compañías farmacéuticas han trasladado los ensayos de evaluación toxicológica a las etapas tempranas de I&D, y hay señales de que esto condujo a una disminución de la tasa de fracaso asociada a toxicidad en ensayos clínicos (Morgan et al., 2018; Shih et al., 2017). Por ende, puede hipotetizarse que la aplicación de estas mismas estrategias al campo de I&D de fármacos destinados al SNC podría contribuir a mejorar las tasas de éxito en fase clínica de los nuevos fármacos destinados al tratamiento de desórdenes del cerebro.

Por esta razón, el presente trabajo de tesis se centrará en el desarrollo de filtros *in silico* de predicción ADME (particularmente relacionados con la biodisponibilidad a nivel central) aplicables al comienzo del ciclo de desarrollo de principios activos destinados al SNC.

1.4. Hipótesis del fármaco libre y barrera hematoencefálica

In vivo, las moléculas de fármaco pueden unirse reversiblemente a proteínas, lípidos y otros elementos tisulares en plasma y en tejidos, o pueden circular libres, es decir no unidas, y de esta forma difundir a través de las biomembranas para interactuar con el objetivo terapéutico deseado, o con otras biomoléculas (por ejemplo, enzimas, transportadores o receptores). Lo antedicho configura la **hipótesis del fármaco libre** (D. A. Smith et al., 2010), que consta de dos proposiciones:

» La concentración de fármaco libre es la misma en ambos lados de cualquier biomembrana en estado estacionario.

(existen sin embargo algunas excepciones: cuando un fármaco tiene baja permeabilidad por difusión pasiva; cuando el fármaco es sustrato de transportadores; cuando el fármaco se distribuye hacia tejidos con flujo sanguíneo reducido y discontinuo).

» La concentración de fármaco libre en el sitio de acción, o biofase, se relaciona de manera directa con la intensidad de la respuesta farmacológica.

(aquí también aparecen excepciones: cuando la acción del fármaco produce la inactivación irreversible del blanco molecular o cuando la acción del fármaco involucra múltiples mecanismos).

La hipótesis del fármaco libre se aplica ampliamente en el descubrimiento y desarrollo de fármacos para establecer relaciones farmacocinéticas-farmacodinámicas (PK/PD), y predecir la dosis terapéuticamente relevante. Ha sido confirmada por numerosos estudios en diversas áreas terapéuticas y para diferentes tipos de blancos moleculares (D. A. Smith et al., 2010).

Dejando de lado las excepciones, la hipótesis del fármaco libre establece que, para poder lograr el efecto farmacológico, el fármaco libre debe alcanzar concentraciones adecuadas en el sitio de acción, lo que en el caso de los desórdenes del SNC implica poder atravesar la barrera hematoencefálica (BHE).

Dicha barrera, posiblemente la más selectiva y estrictamente regulada de todas las barreras biológicas, está conformada por el endotelio de los capilares cerebrales, siendo regulada por las células de la glía adyacentes. Es una estructura multicelular, dinámica, que separa el cerebro de la circulación sistémica (Kadry et al., 2020; Lochhead et al., 2020). Entre las células que forman la BHE hay conexiones laterales muy estrechas (uniones estrechas) que limitan la permeabilidad paracelular de las drogas y otros compuestos, los cuales se ven obligados a cruzar la BHE a través de la vía transcelular (Hawkins et al., 2005). Esta es la forma de entrada de la mayoría de los fármacos actualmente utilizados para el tratamiento de enfermedades del SNC, generalmente pequeñas moléculas lipofílicas capaces de atravesar la BHE pasivamente, como benzodiazepinas o barbitúricos, entre muchos otros. Otros dos mecanismos disponibles para transponer la BHE son el transporte activo y la vía endocítica (mediada por receptores y/o por adsorción, única forma de acceso para moléculas grandes, nanopartículas y complejos de alto peso molecular) (Begley et al., 2003; Pardridge, 2003). Los transportadores de eflujo dependientes de ATP o transportadores ABC (por sus siglas en inglés, *ATP-Binding Cassette*) merecen una mención especial ya que son responsables de la remoción de sus sustratos desde el cerebro hacia el espacio intravascular, lo que a su vez reduce la biodisponibilidad de los mismos en el SNC (Chen et al., 2012; Vasiliou et al., 2008; Vlieghe et al., 2013). La glicoproteína-P (Pgp), varios miembros de las proteínas asociadas a la resistencia a múltiples fármacos (MRPs, *multidrug resistance proteins*) y la proteína de resistencia al cáncer de mama (BCRP, *breast cancer resistance protein*) son los miembros más estudiados de esta superfamilia de transportadores, y su capacidad para exportar sustratos fuera de las células ha demostrado ser un obstáculo importante en la administración de fármacos de acción central (Ejendal et al., 2005; Marquez et al., 2011; Talevi et al., 2012).

Teniendo en cuenta la hipótesis del fármaco libre y lo mencionado acerca de la BHE, se podría pensar entonces que las concentraciones de fármaco libre en plasma y

cerebro serían los parámetros de elección para modelar el equilibrio de distribución establecido a través de la BHE.

A pesar de que las concentraciones de fármaco en plasma generalmente se informan como concentraciones totales (la mayoría de los datos de niveles terapéuticos o efectivos de fármacos que se encuentran en la literatura se refieren a la concentración total), en la actualidad se acepta que los niveles libres conducen a relaciones PK/PD más informativas (Di et al., 2013; Freeman et al., 2019; Hammarlund-Udenaes, 2010; Kalvass et al., 2007; Liu et al., 2008; Nirogi et al., 2020; Watson et al., 2009). Sin embargo, vale la pena destacar que, si la difusión pasiva ocurre muy lentamente, o si el principio activo es sustrato de transportadores, no se alcanzará un equilibrio con niveles libres iguales a ambos lados de la BH. En otras palabras, se logrará un pseudo-equilibrio con concentraciones libres diferentes entre el plasma y el SNC.

1.5. Parámetros que describen la distribución de fármacos entre el cerebro y el plasma

Por todo lo expuesto hasta aquí, resulta claro que el desarrollo de modelos *in silico* que permitan predecir la biodisponibilidad a nivel del SNC de compuestos tipo fármaco sería de gran ayuda durante las etapas iniciales del desarrollo de principios activos para el tratamiento de desórdenes del SNC. Por lo tanto, resulta necesario definir un parámetro farmacocinético que describa la disposición de los fármacos en el cerebro y que sería, por lo tanto, de interés para el modelado y predicción computacional. La Tabla 1.1 presenta un resumen de los principales parámetros que se han propuesto para describir la distribución de fármacos entre el cerebro y el plasma (Hammarlund-Udenaes et al., 2008; Jeffrey et al., 2008; Lanevskij et al., 2013).

Entre los parámetros más informativos se encuentran la concentración de fármaco libre en el cerebro ($C_{u,cerebro}$), la relación entre la concentración de fármaco libre en cerebro y la concentración de fármaco libre en plasma ($K_{p,uu}$) y la relación entre la concentración total de fármaco en cerebro y la concentración libre en plasma ($K_{p,u}$, el cual sólo elimina el efecto de la unión en el plasma). Basado en la suposición del

transporte pasivo, $K_{p,u}$ permite estimar el porcentaje de unión a elementos tisulares en el tejido cerebral (Lanevskij et al., 2013).

Tabla 1.1. Principales parámetros propuestos para describir la distribución de fármaco entre cerebro y plasma.

Parámetro	Expresión	Definición
C_{plasma} y $C_{cerebro}$		Concentraciones totales en plasma y cerebro (suma de las concentraciones de fármaco libre y unido).
$f_{u,plasma}$ y $f_{u,cerebro}$		Fracción de fármaco no unido en plasma y cerebro, respectivamente.
$C_{u,plasma}$ y $C_{u,cerebro}$	$C_{u,plasma} = f_{u,plasma} * C_{plasma}$ $C_{u,cerebro} = f_{u,cerebro} * C_{cerebro}$	Concentración libre, difusible y terapéuticamente activa del fármaco en plasma y cerebro, respectivamente.
K_p	$K_p = C_{cerebro}/C_{plasma}$	Relación de la concentración total en cerebro con respecto a la total en plasma, en estado estacionario.
Log BB	$Log\ BB = log(K_p)$	Expresión logarítmica de la relación de las concentraciones totales en cerebro con respecto a plasma.
$K_{p,u}$	$K_{p,u} = C_{cerebro}/C_{u,plasma}$	Relación entre la concentración total en cerebro con respecto a la libre en plasma, en estado estacionario.
$K_{p,uu}$	$K_{p,uu} = C_{u,cerebro}/C_{u,plasma}$	Relación entre la concentración libre en cerebro con respecto a la libre en plasma, en estado estacionario.

Otros parámetros como la relación entre las concentraciones totales en cerebro y plasma (K_p) y su expresión logarítmica (*Log BB*), así como las fracciones de fármaco libres en plasma y cerebro ($f_{u,plasma}$ y $f_{u,cerebro}$, respectivamente) son menos informativos para comprender o predecir la eficacia *in vivo* de los fármacos en el SNC (Di et al., 2013). A pesar de eso, *Log BB* es claramente el parámetro cuantitativo más popular, tradicionalmente utilizado por la industria farmacéutica para determinar la eficiencia de la distribución de fármacos en el SNC. Sin embargo, desde hace algunos años su utilidad comenzó a cuestionarse (de Lange et al., 2015; Lanevskij et al., 2013; Summerfield et al., 2013), debido a que *Log BB* no brinda

información acerca de la concentración de fármaco libre cerebro, la cual, de acuerdo con la hipótesis del fármaco libre discutida anteriormente, se relaciona de manera directa con la intensidad del efecto farmacológico (Reichel, 2009; Summerfield et al., 2013). Un ejemplo claro de la utilidad limitada de *Log BB* para describir los efectos farmacológicos de un fármaco en el SNC se puede encontrar en un trabajo de Watson *et al.* del 2009, donde los autores demostraron claramente que la ocupación de receptores D2 en ratas se correlacionaba mejor con $C_{u,cerebro}$ que con K_p. Se observó en ese estudio que, en términos de K_p, la Risperidona poseía una baja penetración al SNC, sin embargo, esto se vio compensado por el hecho de que poseía una alta concentración libre y una alta potencia D2 (Summerfield et al., 2013; Watson et al., 2009).

Inicialmente introducido por Gupta *et al.* (A. Gupta et al., 2006) el parámetro $K_{p,uu}$ permite evaluar la distribución de fármaco libre entre el plasma y el cerebro en estado estacionario. Como lo mostraron primero Gupta y luego Boström *et al.* (Boström et al., 2006), este parámetro también proporciona información sobre el mecanismo de transporte a través de la BHE, y es sensible a la afinidad del compuesto por los transportadores de eflujo/influjo que se encuentran en las células endoteliales. Los compuestos con buena permeabilidad por procesos pasivos y que no son sustratos de transportadores de eflujo (P-gp, BCRP) normalmente presentan valores de $K_{p,uu}$ cercanos a la unidad. Valores de $K_{p,uu}$ menores a 1, en cambio, podrían indicar que un compuesto es sustrato de un transportador de eflujo y/o que tiene penetración cerebral limitada debido a una baja capacidad de difusión pasiva a través de la BHE, mientras que, por el contrario, valores de $K_{p,uu}$ superiores a 1 sugieren captación activa hacia el cerebro mediada por transportadores de influjo (de Lange et al., 2015; Di et al., 2013; Summerfield et al., 2013).

Posiblemente debido a la menor dificultad que reviste la determinación experimental de *Log BB* (en comparación a parámetros que involucran concentraciones libres), y su uso histórico, existe una mayor disponibilidad de datos experimentales para este parámetro, y los modelos *in silico* que utilizan *Log BB* son abundantes en la literatura, incluso en los últimos años (Brito-Sánchez et al., 2015; Bujak et al., 2015; Golmohammadi et al., 2012; S. Gupta et al., 2015; W. Wang et al., 2015; Zhu et al., 2018), cuando se han reconocido las desventajas y limitaciones del

uso de este parámetro (Di et al., 2015; Loryan et al., 2015; Varadharajan et al., 2015). Loryan *et al.* (Loryan et al., 2014) han estudiado la relación entre K_p y $K_{p,uu,}$ encontrando una pobre correlación entre ambos. Existen excelentes revisiones de los modelos computacionales de *Log BB* (Bujak et al., 2015; Di et al., 2015; Lanevskij et al., 2013; Mehdipour et al., 2009; Mensch et al., 2009; Raevsky et al., 2013; Subramanian et al., 2003).

Por lo antedicho, en este trabajo nos enfocaremos en el parámetro farmacocinético de mayor biorrelevancia para evaluar la distribución de fármacos en el SNC, $K_{p,uu.}$ Dicho parámetro se utilizará para el desarrollo de modelos computacionales de predicción de biodisponibilidad a nivel del SNC, que luego podrían ser utilizados en las etapas tempranas de I&D de nuevos fármacos cuyo blanco molecular se localice en el SNC.

1.6. Cuantificación del transporte de drogas a través de la BHE

La determinación del $K_{p,uu}$ de un compuesto implica la cuantificación experimental de $C_{u,cerebro}$ y $C_{u,plasma}$, ambas concentraciones medidas en estado estacionario. Acorde a la expresión presentada en la Tabla 1.1, es posible obtener $C_{u,plasma}$ como el producto $f_{u,plasma} * C_{plasma.}$

Las técnicas de referencia para medir la $f_{u,plasma}$ son la diálisis de equilibrio y la ultrafiltración (Metsu et al., 2020); la $C_{u,plasma}$ luego es calculada conociendo la C_{plasma} (generalmente determinada por métodos cromatográficos o inmunológicos). Si bien se han propuesto otros enfoques para evaluar la $C_{u,plasma}$ [los cuales se basan en distintas técnicas experimentales como espectroscopía, calorimetría, HPLC y electroforesis capilar (Vuignier et al., 2010; Yang et al., 2012), e incluso en el uso de saliva, como ultrafiltrado natural de plasma, para la cuantificación directa de la concentración de fármaco libre (Ibarra et al., 2010; Fagiolino et al., 2013)], la diálisis de equilibrio y la ultrafiltración siguen siendo las metodologías más utilizadas. Cabe destacar, sin embargo, que estas técnicas también poseen inconvenientes, ya que requieren de mucho tiempo y grandes volúmenes de muestra, entre otros (Vuignier et al., 2013).

Por otro lado, la cuantificación directa de la $C_{u,cerebro}$ sólo es posible a través de la microdiálisis (MD) intracerebral *in vivo* (Spreafico et al., 2013; Ungerstedt, 1991), técnica que implica la implantación de una sonda en el cerebro mediante cirugía estereotáxica. Este método invasivo generalmente se realiza en roedores, y más comúnmente en ratas. La sonda está formada por una membrana de diálisis tubular en contacto con el medio cerebral. Dentro de la sonda, cánulas de entrada y salida permiten la perfusión y recolección de muestras. Por la cánula de entrada se perfunde líquido extracelular cerebral artificial (Zapata et al., 2009), para asegurar que la difusión de los compuestos se deba únicamente al gradiente de concentración a través de la membrana semipermeable, y por la cánula de salida se recoge el dializado. La cuantificación del fármaco se realiza luego en el dializado muestreado. Un aspecto importante de la técnica es que requiere un paso adicional para la determinación de la recuperación de la sonda, porque generalmente no se alcanza el equilibrio de concentración en ambos lados de la membrana y, por lo tanto, la concentración en el dializado es inferior a la verdadera concentración extracelular en el cerebro (Chefer et al., 2009; Di Giovanni et al., 2013; Di et al., 2015; Müller, 2013).

A pesar de ser la técnica de referencia, la MD en roedores tiene varias desventajas que evitan que se convierta en una técnica *high throughput*: es una metodología costosa, requiere mucho tiempo y sólo puede ser realizada por personal altamente capacitado. Estos inconvenientes son probablemente la causa de la escasez de datos de $K_{p,uu}$ disponibles en bibliografía, y también la razón por la cual se han propuesto otras técnicas experimentales para determinar $C_{u,cerebro}$ (Read et al., 2010).

Es importante subrayar que, aunque se ha propuesto el uso de concentraciones en el líquido cefalorraquídeo como sustituto de la $C_{u,cerebro}$, se demostró que ambas concentraciones no siempre se correlacionan bien (Fridén et al., 2011, 2007; A. Gupta et al., 2006; Q. R. Smith, 2003).

La técnica de homogeneización, implementada por primera vez por Kalvass y Maurer en 2002, consiste en colocar una pequeña muestra de homogenato de cerebro (tejido cerebral diluido con solución buffer y homogeneizado mecánicamente o por ultrasonido) enriquecido con el fármaco en estudio, en un aparato de diálisis de equilibrio de 96 pocillos, y dializarlo contra una solución

buffer fresca (Kalvass et al., 2002; Mano et al., 2002). Conociendo la $C_{cerebro}$ en estado estacionario y determinando la $f_{u,cerebro}$ mediante el experimento de diálisis descripto anteriormente, se puede obtener la $C_{u,cerebro}$ como $f_{u,cerebro}$* $C_{cerebro.}$ Se trata de una técnica de alto rendimiento, compatible con las etapas iniciales del descubrimiento de fármacos en el SNC (Wan et al., 2007), pero con una gran debilidad: el proceso de homogeneización puede modificar las propiedades de unión del tejido, exponiendo sitios adicionales de unión a fármacos, potencialmente subestimando la fracción libre (Becker et al., 2006; Read et al., 2010).

Otro método para estimar la $K_{p,uu}$ es el método de *slice* (corte o rebanada)(Becker et al., 2006; Fridén et al., 2007), que consiste en la incubación a 37 °C de cortes de cerebro (de alrededor de 400 µm de ancho) de rata o ratón en solución buffer con una concentración conocida del compuesto de prueba. Mediante esta técnica se determina el volumen de distribución libre en cerebro en estado estacionario ($V_{u,cerebro}$). A los tiempos designados, se extraen alícuotas del buffer y el compuesto de prueba se cuantifica mediante una técnica adecuada. Al igual que la técnica de homogeneización, el método de *slice* también se ha desarrollado como método de alto rendimiento, con la ventaja adicional de que se conserva la estructura celular (Fridén et al., 2009). Finalmente, para poder obtener el valor de $K_{p,uu}$ mediante la técnica de *slice*, se necesita corregir el K_p del compuesto por su $V_{u,cerebro}$ y la $f_{u,plasma}$ mediante la ecuación 1.1:

$$K_{p,uu} = \frac{K_p}{V_{u,cerebro} \cdot f_{u,plasma}} \qquad (1.1)$$

Fridén y colaboradores compararon el desempeño de los tres métodos mencionados anteriormente en un conjunto de 15 compuestos diversos desde el punto de vista químico y farmacológico. El método de *slice* consiguió estimar el valor de $K_{p,uu}$ dentro de un rango de +/- tres veces el valor obtenido por MD, para 14 de los 15 compuestos ensayados. En cambio, el método del homogenato proveyó valores por fuera de ese rango para 5 de los 15 compuestos. Los mejores resultados del método de *slice* se deben, probablemente, a la mayor viabilidad del tejido (demostrada por la preservación de los niveles de ATP durante el experimento) y a la mayor preservación de las estructuras cerebrales, en comparación con el método que requiere homogenización del tejido (Fridén et al., 2007). El método del homogenato

proporcionó resultados particularmente pobres para compuestos que muestran una distribución muy heterogénea entre el fluido intersticial y el fluido intracelular (por ejemplo, el glucurónido de morfina) y para compuestos sujetos a transporte activo a nivel del parénquima cerebral (por ejemplo, la gabapentina).

En conclusión, la técnica MD debería ser la primera opción para medir el $K_{p,uu}$. Sin embargo, si se pretende una evaluación de numerosos candidatos a fármacos, los métodos de *slice* y *homogenato* podrían ser opciones adecuadas para obtener buenas estimaciones cerebrales dentro de tiempos optimizados y a bajo costo, teniendo en cuenta las limitaciones antes mencionadas (Fridén et al., 2011).

Referencias

Becker, S., & Liu, X. (2006). Evaluation of the utility of brain slice methods to study brain penetration. *Drug Metabolism and Disposition: The Biological Fate of Chemicals*, *34*(5), 855–861. https://doi.org/10.1124/dmd.105.007914

Begley, D. J., & Brightman, M. W. (2003). Structural and functional aspects of the blood-brain barrier. In *Peptide Transport and Delivery into the Central Nervous System,* 61, 39–78. https://doi.org/10.1007/978-3-0348-8049-7_2

Boström, E., Simonsson, U. S. H., & Hammarlund-Udenaes, M. (2006). In vivo blood-brain barrier transport of oxycodone in the rat: indications for active influx and implications for pharmacokinetics/pharmacodynamics. *Drug Metabolism and Disposition: The Biological Fate of Chemicals*, *34*(9), 1624–1631. https://doi.org/10.1124/dmd.106.009746

Brito-Sánchez, Y., Marrero-Ponce, Y., Barigye, S. J., Yaber-Goenaga, I., Morell Pérez, C., Le-Thi-Thu, H. et al. (2015). Towards better BBB passage prediction using an extensive and curated data set. *Molecular Informatics*, *34*(5), 308–330. https://doi.org/10.1002/minf.201400118

Bujak, R., Struck-Lewicka, W., Kaliszan, M., Kaliszan, R., & Markuszewski, M. J. (2015). Blood-brain barrier permeability mechanisms in view of quantitative structure-activity relationships (QSAR). *Journal of Pharmaceutical and Biomedical Analysis*, *108*, 29–37. https://doi.org/10.1016/j.jpba.2015.01.046

Chefer, V. I., Thompson, A. C., Zapata, A., & Shippenberg, T. S. (2009). Overview of brain microdialysis. *Current Protocols in Neuroscience*, *47*(1), 1-27. https://doi.org/10.1002/0471142301.ns0701s47

Chen, Y., & Liu, L. (2012). Modern methods for delivery of drugs across the blood-brain barrier. *Advanced Drug Delivery Reviews*, *64*(7), 640–665. https://doi.org/10.1016/j.addr.2011.11.010

Choi, D. W., Armitage, R., Brady, L. S., Coetzee, T., Fisher, W., Hyman, S. et al. (2014) Medicines for the mind: Policy-based “Pull” incentives for creating breakthrough CNS drugs. *Neuron*, 84, 554–563. https://doi.org/10.1016/j.neuron.2014.10.027

de Lange, E. C. M., & Hammarlund-Udenaes, M. (2015). Translational aspects of blood-brain barrier transport and central nervous system effects of drugs: from discovery to patients. *Clinical Pharmacology and Therapeutics*, *97*(4), 380–394. https://doi.org/10.1002/cpt.76

Di Giovanni, G., & Di Matteo, V. (Eds.). (2013). *Microdialysis Techniques in Neuroscience*. Humana Press, Totowa, NJ. https://doi.org/10.1007/978-1-62703-173-8

Di, L., & Kerns, E. H. (Eds.). (2015). *Blood-Brain Barrier in Drug Discovery*. John Wiley & Sons, Inc. Hoboken, NJ. https://doi.org/10.1002/9781118788523

Di, L., Rong, H., & Feng, B. (2013). Demystifying brain penetration in central nervous system drug discovery. Miniperspective. *Journal of Medicinal Chemistry*, *56*(1), 2–12. https://doi.org/10.1021/jm301297f

Ejendal, K. F., & Hrycyna, C. A. (2005). Multidrug resistance and cancer: The role of the human ABC transporter ABCG2. *Current Protein & Peptide Science*, *3*(5), 503–511. https://doi.org/10.2174/1389203023380521

Fagiolino, P., Vázquez, M., Maldonado, C., Ruiz, M. E., Volonté, M.G, Orozco-Suárez, S. et al. (2013). Usefulness of salivary drug monitoring for detecting efflux transporter overexpression. *Current Pharmaceutical Design*, *19*(38), 6701–6708. https://doi.org/10.2174/13816128113199990368

Freeman, B. B., Yang, L., & Rankovic, Z. (2019). Practical approaches to evaluating and optimizing brain exposure in early drug discovery. *European Journal of Medicinal Chemistry*, *182*, 111643. https://doi.org/10.1016/j.ejmech.2019.111643

Fridén, M., Bergström, F., Wan, H., Rehngren, M., Ahlin, G., Hammarlund-Udenaes, M. et al. (2011). Measurement of unbound drug exposure in brain: modeling of pH partitioning explains diverging results between the brain slice and brain homogenate methods. *Drug Metabolism and Disposition: The Biological Fate of Chemicals*, *39*(3), 353–362. https://doi.org/10.1124/dmd.110.035998

Fridén, M., Ducrozet, F., Middleton, B., Antonsson, M., Bredberg, U., & Hammarlund-Udenaes, M. (2009). Development of a high-throughput brain slice method for studying drug distribution in the central nervous system. *Drug Metabolism*

and Disposition: The Biological Fate of Chemicals, *37*(6), 1226–1233. https://doi.org/10.1124/dmd.108.026377

Fridén, M., Gupta, A., Antonsson, M., Bredberg, U., & Hammarlund-Udenaes, M. (2007). In vitro methods for estimating unbound drug concentrations in the brain interstitial and intracellular fluids. *Drug Metabolism and Disposition*, *35*(9), 1711–1719. https://doi.org/10.1124/dmd.107.015222

Golmohammadi, H., Dashtbozorgi, Z., & Acree, W. E. (2012). Quantitative structure-activity relationship prediction of blood-to-brain partitioning behavior using support vector machine. *European Journal of Pharmaceutical Sciences : Official Journal of the European Federation for Pharmaceutical Sciences*, *47*(2), 421–429. https://doi.org/10.1016/j.ejps.2012.06.021

Gribkoff, V. K., & Kaczmarek, L. K. (2017). The need for new approaches in CNS drug discovery: Why drugs have failed, and what can be done to improve outcomes. *Neuropharmacology*, *120*, 11–19. https://doi.org/10.1016/j.neuropharm.2016.03.021

Gupta, A., Chatelain, P., Massingham, R., Jonsson, E. N., & Hammarlund-Udenaes, M. (2006). Brain distribution of cetirizine enantiomers: comparison of three different tissue-to-plasma partition coefficients: K(p), K(p,u), and K(p,uu). *Drug Metabolism and Disposition: The Biological Fate of Chemicals*, *34*(2), 318–323. https://doi.org/10.1124/dmd.105.007211

Gupta, S., Basant, N., & Singh, K. P. (2015). Qualitative and quantitative structure-activity relationship modelling for predicting blood-brain barrier permeability of structurally diverse chemicals. *SAR and QSAR in Environmental Research*, *26*(2), 95–124. https://doi.org/10.1080/1062936X.2014.994562

Hammarlund-Udenaes, M. (2010). Active-site concentrations of chemicals - are they a better predictor of effect than plasma/organ/tissue concentrations? *Basic & Clinical Pharmacology & Toxicology*, *106*(3), 215–220. https://doi.org/10.1111/j.1742-7843.2009.00517.x

Hammarlund-Udenaes, M., Fridén, M., Syvänen, S., & Gupta, A. (2008). On the rate and extent of drug delivery to the brain. *Pharmaceutical Research*, *25*(8), 1737–1750. https://doi.org/10.1007/s11095-007-9502-2

Hawkins, B. T., & Davis, T. P. (2005). The blood-brain barrier/neurovascular unit in health and disease. *Pharmacological Reviews*, *57*(2), 173–185. https://doi.org/10.1124/pr.57.2.4

Ibarra, M., Vázquez, M., Fagiolino, P., Mutilva, F., & Canale, A. (2010). Total, unbound plasma and salivary phenytoin levels in critically ill patients. *Journal of Epilepsy and Clinical Neurophysiology*, *16*(2), 69–73. https://doi.org/10.1590/S1676-26492010000200006

James, S. L., Abate, D., Abate, K. H., Abay, S. M., Abbafati, C., Abbasi, N. et al. (2018). Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017. *The Lancet*, *392*(10159), 1789–1858. https://doi.org/10.1016/S0140-6736(18)32279-7

Jeffrey, P., & Summerfield, S. G. (2008). Challenges for blood-brain barrier (BBB) screening. *Xenobiotica; the Fate of Foreign Compounds in Biological Systems*, *37*(10–11), 1135–1151. https://doi.org/10.1080/00498250701570285

Kadry, H., Noorani, B., & Cucullo, L. (2020). A blood–brain barrier overview on structure, function, impairment, and biomarkers of integrity. *Fluids and Barriers of the CNS*, *17*(1), 69. https://doi.org/10.1186/s12987-020-00230-3

Kalvass, J. C., & Maurer, T. S. (2002). Influence of nonspecific brain and plasma binding on CNS exposure: implications for rational drug discovery. *Biopharmaceutics & Drug Disposition*, *23*(8), 327–338. https://doi.org/10.1002/bdd.325

Kalvass, J. C., Olson, E. R., Cassidy, M. P., Selley, D. E., & Pollack, G. M. (2007). Pharmacokinetics and Pharmacodynamics of Seven Opioids in P-Glycoprotein-Competent Mice: Assessment of Unbound Brain EC 50,u and Correlation of in Vitro, Preclinical, and Clinical Data. *Journal of Pharmacology and Experimental Therapeutics*, *323*(1), 346–355. https://doi.org/10.1124/jpet.107.119560

Kassel, D. B. (2004). Applications of high-throughput ADME in drug discovery. *Current Opinion in Chemical Biology*, *8*(3), 339–345. https://doi.org/10.1016/J.CBPA.2004.04.015

Kesselheim, A. S., Hwang, T. J., & Franklin, J. M. (2015). Two decades of new drug development for central nervous system disorders. *Nature Reviews Drug Discovery*, 14, 815–816. https://doi.org/10.1038/nrd4793

Kola, I., & Landis, J. (2004). Can the pharmaceutical industry reduce attrition rates? *Nature Reviews Drug Discovery*, *3*(8), 711–716. https://doi.org/10.1038/nrd1470

Kyu, H. H., Abate, D., Abate, K. H., Abay, S. M., Abbafati, C., Abbasi, N. et al. (2018). Global, regional, and national disability-adjusted life-years (DALYs) for 359 diseases and injuries and healthy life expectancy (HALE) for 195 countries and territories, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017. *The Lancet*, *392*(10159), 1859–1922. https://doi.org/10.1016/S0140-6736(18)32335-3

Lanevskij, K., Japertas, P., & Didziapetris, R. (2013). Improving the prediction of drug disposition in the brain. *Expert Opinion on Drug Metabolism & Toxicology*, *9*(4), 473–486. https://doi.org/10.1517/17425255.2013.754423

Liu, Z., Sall, A., & Yang, D. (2008). MicroRNA: An emerging therapeutic target and intervention tool. *Int J Mol Sci*, *9*(6), 978–999. https://doi.org/10.3390/ijms9060978

Lochhead, J. J., Yang, J., Ronaldson, P. T., & Davis, T. P. (2020). Structure, function, and regulation of the blood-brain barrier tight junction in central nervous system disorders. *Frontiers in Physiology*, *11*. https://doi.org/10.3389/fphys.2020.00914

Loryan, I., Sinha, V., Mackie, C., Van Peer, A., Drinkenburg, W. H., Vermeulen, A. et al. (2015). Molecular properties determining unbound intracellular and extracellular brain exposure of CNS drug candidates. *Molecular Pharmaceutics*, *12*(2), 520–532. https://doi.org/10.1021/mp5005965

Loryan, I., Sinha, V., Mackie, C., Van Peer, A., Drinkenburg, W., Vermeulen, A. et al. (2014). Mechanistic understanding of brain drug disposition to optimize the selection of potential neurotherapeutics in drug discovery. *Pharmaceutical Research*, *31*(8), 2203–2219. https://doi.org/10.1007/s11095-014-1319-1

Mano, Y., Higuchi, S., & Kamimura, H. (2002). Investigation of the high partition of

YM992, a novel antidepressant, in rat brain - in vitro and in vivo evidence for the high binding in brain and the high permeability at the BBB. *Biopharmaceutics & Drug Disposition*, *23*(9), 351–360. https://doi.org/10.1002/bdd.328

Marquez, B., & Van Bambeke, F. (2011). ABC multidrug transporters: target for modulation of drug pharmacokinetics and drug-drug interactions. *Current Drug Targets*, *12*(5), 600–620. https://doi.org/10.2174/138945011795378504

Mehdipour, A. R., & Hamidi, M. (2009). Brain drug targeting: a computational approach for overcoming blood-brain barrier. *Drug Discovery Today*, *14*(21–22), 1030–1036. https://doi.org/10.1016/j.drudis.2009.07.009

Mensch, J., Oyarzabal, J., Mackie, C., & Augustijns, P. (2009). In vivo, in vitro and in silico methods for small molecule transfer across the BBB. *Journal of Pharmaceutical Sciences*, *98*(12), 4429–4468. https://doi.org/10.1002/jps.21745

Metsu, D., Lanot, T., Fraissinet, F., Concordet, D., Gayrard, V., Averseng, M. et al (2020). Comparing ultrafiltration and equilibrium dialysis to measure unbound plasma dolutegravir concentrations based on a design of experiment approach. *Scientific Reports*, *10*(1), 12265. https://doi.org/10.1038/s41598-020-69102-y

Morgan, P., Brown, D. G., Lennard, S., Anderton, M. J., Barrett, J. C., Eriksson, U. et al. (2018). Impact of a five-dimensional framework on R&D productivity at AstraZeneca. *Nature Reviews Drug Discovery*, 17, 167–181. https://doi.org/10.1038/nrd.2017.244

Müller, M. (Ed.). (2013). *Microdialysis in Drug Development*. Springer, NY. https://doi.org/10.1007/978-1-4614-4815-0

Murray, C. J. L., Vos, T., Lozano, R., Naghavi, M., Flaxman, A. D., Michaud, C. et al. (2012). Disability-adjusted life years (DALYs) for 291 diseases and injuries in 21 regions, 1990–2010: a systematic analysis for the Global Burden of Disease Study 2010. *The Lancet*, *380*(9859), 2197–2223. https://doi.org/10.1016/S0140-6736(12)61689-4

Nirogi, R., Molgara, P., Bhyrapuneni, G., Manoharan, A., Padala, N. P., & Palacharla, V. R. C. (2020). The use of inactivated brain homogenate to determine the in vitro fraction unbound in brain for unstable compounds. *Xenobiotica*, *50*(10), 1228–1235. https://doi.org/10.1080/00498254.2020.1771795

Pardridge, W. M. (2003). Blood-brain barrier drug targeting: the future of brain drug development. *Molecular Interventions*, *3*(2), 90–105, 51. https://doi.org/10.1124/mi.3.2.90

Raevsky, O. A., Solodova, S. L., Lagunin, A. A., & Poroikov, V. V. (2013). Computer modeling of blood brain barrier permeability for physiologically active compounds. *Biochemistry (Moscow) Supplement Series B: Biomedical Chemistry*, *7*(2), 95–107. https://doi.org/10.1134/S199075081302008X

Read, K. D., & Braggio, S. (2010). Assessing brain free fraction in early drug discovery. *Expert Opinion on Drug Metabolism & Toxicology*, *6*(3), 337–344. https://doi.org/10.1517/17425250903559873

Reichel, A. (2009). Addressing central nervous system (CNS) penetration in drug discovery: basics and implications of the evolving new concept. *Chemistry & Biodiversity*, *6*(11), 2030–2049. https://doi.org/10.1002/cbdv.200900103

Scannell, J. W., Blanckley, A., Boldon, H., & Warrington, B. (2012). Diagnosing the decline in pharmaceutical R&D efficiency. *Nature Reviews Drug Discovery*, 11, 191–200. https://doi.org/10.1038/nrd3681

Schnorrenberg, G. (2018). New trends in drug discovery. In J. Fischer, C. Klein & W. E. Childers (Eds.), *Successful Drug Discovery*, pp. 3–39). John Wiley & Sons, Inc, Hoboken, NJ. https://doi.org/10.1002/9783527808694.ch1

Schuhmacher, A., Gassmann, O., & Hinder, M. (2016). Changing R&D models in research-based pharmaceutical companies. *Journal of Translational Medicine*, 14, 105. https://doi.org/10.1186/s12967-016-0838-4

Shih, H.-P., Zhang, X., & Aronov, A. M. (2017). Drug discovery effectiveness from the standpoint of therapeutic mechanisms and indications. *Nature Reviews Drug Discovery*, *17*(1), 19–33. https://doi.org/10.1038/nrd.2017.194

Smietana, K., Siatkowski, M., & Møller, M. (2016). Trends in clinical success rates.

Nature Reviews Drug Discovery, 15, 379–380. https://doi.org/10.1038/nrd.2016.85

Smith, D. A., Di, L., & Kerns, E. H. (2010). The effect of plasma protein binding on in vivo efficacy: misconceptions in drug discovery. *Nature Reviews Drug Discovery*, *9*(12), 929–939. https://doi.org/10.1038/nrd3287

Smith, Q. R. (2003). A review of blood-brain barrier transport techniques. *Methods in Molecular Medicine*, *89*, 193–208. https://doi.org/10.1385/1-59259-419-0:193

Spreafico, M., & Jacobson, M. P. (2013). In silico prediction of brain exposure: Drug free fraction, unbound brain to plasma concentration ratio and equilibrium half-life. *Current Topics in Medicinal Chemistry*, *13*(7), 813–820. https://doi.org/10.2174/1568026611313070004

Subramanian, G., & Kitchen, D. B. (2003). Computational models to predict blood–brain barrier permeation and CNS activity. *Journal of Computer-Aided Molecular Design*, *17*(10), 643–664. https://doi.org/10.1023/B:JCAM.0000017372.32162.37

Summerfield, S. G., & Dong, K. C. (2013). In vitro, in vivo and in silico models of drug distribution into the brain. *Journal of Pharmacokinetics and Pharmacodynamics*, *40*(3), 301–314. https://doi.org/10.1007/s10928-013-9303-7

Talevi, A., & Bruno-Blanch, L. E. (2012). Efflux-transporters at the blood-brain barrier: Therapeutic opportunities. In P. A. Montenegro & S. M. Juárez (Eds.), *The Blood-Brain Barrier: New Research,* 1st ed., pp. 117–144. Nova Publishers, NY.

Ungerstedt, U. (1991). Microdialysis-principles and applications for studies in animals and man. *Journal of Internal Medicine*, *230*(4), 365–373. https://doi.org/10.1111/j.1365-2796.1991.tb00459.x

Varadharajan, S., Winiwarter, S., Carlsson, L., Engkvist, O., Anantha, A., Kogej, T. et al. (2015). Exploring in silico prediction of the unbound brain-to-plasma drug concentration ratio: model validation, renewal, and interpretation. *Journal of Pharmaceutical Sciences*, *104*(3), 1197–1206.

https://doi.org/10.1002/jps.24301

Vasiliou, V., Vasiliou, K., & Nebert, D. W. (2008). Human ATP-binding cassette (ABC) transporter family. *Human Genomics*, *3*(3), 281. https://doi.org/10.1186/1479-7364-3-3-281

Vlieghe, P., & Khrestchatisky, M. (2013). Medicinal chemistry based approaches and nanotechnology-based systems to improve CNS drug targeting and delivery. *Medicinal Research Reviews*, *33*(3), 457–516. https://doi.org/10.1002/med.21252

Vos, T., Flaxman, A. D., Naghavi, M., Lozano, R., Michaud, C., Ezzati, M. et al. (2012). Years lived with disability (YLDs) for 1160 sequelae of 289 diseases and injuries 1990–2010: a systematic analysis for the Global Burden of Disease Study. *The Lancet*, *380*(9859), 2163–2196. https://doi.org/10.1016/S0140-6736(12)61729-2

Vuignier, K., Schappler, J., Veuthey, J. L., Carrupt, P. A., & Martel, S. (2010). Drug-protein binding: a critical review of analytical tools. *Analytical and Bioanalytical Chemistry*, *398*(1), 53–66. https://doi.org/10.1007/s00216-010-3737-1

Vuignier, K., Veuthey, J. L., Carrupt, P. A., & Schappler, J. (2013). Global analytical strategy to measure drug-plasma protein interactions: from high-throughput to in-depth analysis. *Drug Discovery Today*, *18*(21–22), 1030–1034. https://doi.org/10.1016/j.drudis.2013.04.006

Wan, H., Rehngren, M., Giordanetto, F., Bergström, F., & Tunek, A. (2007). High-throughput screening of drug-brain tissue binding and in silico prediction for assessment of central nervous system drug delivery. *Journal of Medicinal Chemistry*, *50*(19), 4606–4615. https://doi.org/10.1021/jm070375w

Wang, J. (2009). Comprehensive assessment of ADMET risks in drug discovery. *Current Pharmaceutical Design*, *15*(19), 2195–2219. https://doi.org/10.2174/138161209788682514

Wang, W., Kim, M. T., Sedykh, A., & Zhu, H. (2015). Developing enhanced blood–brain barrier permeability models: integrating external bio-assay data in QSAR modeling. *Pharmaceutical Research*, *32*(9), 3055–3065.

https://doi.org/10.1007/s11095-015-1687-1

Waring, M. J., Arrowsmith, J., Leach, A. R., Leeson, P. D., Mandrell, S., Owen, R. M. et al. (2015). An analysis of the attrition of drug candidates from four major pharmaceutical companies. *Nature Reviews Drug Discovery*, *14*(7), 475–486. https://doi.org/10.1038/nrd4609

Watson, J., Wright, S., Lucas, A., Clarke, K. L., Viggers, J., Cheetham, S., et al. (2009). Receptor occupancy and brain free fraction. *Drug Metabolism and Disposition: The Biological Fate of Chemicals*, *37*(4), 753–760. https://doi.org/10.1124/dmd.108.022814

Wishart, D. S. (2007). Improving early drug discovery through ADME modelling: An overview. *Drugs in R and D*, 8, 349–362. https://doi.org/10.2165/00126839-200708060-00003

Wong, C. H., Siah, K. W., & Lo, A. W. (2019). Estimation of clinical trial success rates and related parameters. *Biostatistics*, *20*(2), 273–286. https://doi.org/10.1093/biostatistics/kxx069

Yang, G. X., Li, X., & Snyder, M. (2012). Investigating metabolite-protein interactions: an overview of available techniques. *Methods*, *57*(4), 459–466. https://doi.org/10.1016/j.ymeth.2012.06.013

Yokley, B. H., Hartman, M., & Slusher, B. S. (2017). Role of academic drug discovery in the quest for new CNS therapeutics. *ACS Chemical Neuroscience*, *8*(3), 429–431. https://doi.org/10.1021/acschemneuro.7b00040

Zapata, A., Chefer, V. I., & Shippenberg, T. S. (2009). Microdialysis in rodents. *Current Protocols in Neuroscience*, *47*(1), 1-29. https://doi.org/10.1002/0471142301.ns0702s47

Zhu, L., Zhao, J., Zhang, Y., Zhou, W., Yin, L., Wang, Y. et al. (2018). ADME properties evaluation in drug discovery: in silico prediction of blood–brain partitioning. *Molecular Diversity*, *22*(4), 979–990. https://doi.org/10.1007/s11030-018-9866-8

Capítulo 2

Modelado QSAR

2.1. Descubrimiento de fármacos: breve reseña histórica

Las primeras etapas de desarrollo de la industria farmacéutica moderna se remontan al siglo XIX. En ese momento, además de las medicinas tradicionales y populares, el arsenal de fármacos que la comunidad médica tenía a disposición era sensiblemente menor a lo que conocemos hoy en día (Walsh, 2013). Los avances en los campos de la Biología y la Química Orgánica ayudaron a impulsar la innovación en la incipiente industria farmacéutica. Un ejemplo de esto podría ser la síntesis química de la aspirina en 1897 por el químico de Bayer Felix Hoffmann (Montinari et al., 2019).

En la década de 1930 la industria experimentó un punto de inflexión. El hito inicial fue probablemente el descubrimiento y la síntesis química de las sulfamidas, un grupo de moléculas derivadas del colorante rojo llamado *prontosil rubrum*, que demostraron ser efectivas en el tratamiento de una amplia variedad de infecciones bacterianas. Por otro lado, aunque había sido ya previamente aislada a partir de

páncreas caninos en 1921 y utilizada terapéuticamente, la producción industrial de insulina a gran escala también comenzó en los años 30 (Vecchio et al., 2018).

El éxito de estos medicamentos dio un gran impulso a la industria farmacéutica, reflejado en la fabricación de penicilina a escala industrial desde principios de los años 40. También por esa época se fundaron muchas de las principales compañías farmacéuticas actuales o sus precursoras, como Ciba Geigy, Eli Lilly, Glaxo y Roche, entre otras. En los años siguientes, estas compañías desarrollaron fármacos como las tetraciclinas, los corticosteroides, los anticonceptivos orales, los antidepresivos y muchos más (Aminov, 2017; Lopez-Munoz et al., 2009; Walsh, 2013).

En sus orígenes, buena parte de los hallazgos exitosos de la industria farmacéutica tuvieron su origen en el azar, lo que también se conoce como serendipia. De hecho, mucho antes de que tal industria existiera, los fármacos se descubrían por accidente y su uso se transmitía por tradición escrita y oral. Posteriormente, en las décadas de 1950 y 1960, la industria farmacéutica avanzó hacia aproximaciones más sistemáticas para el descubrimiento de fármacos. En particular, el cribado sistemático exhaustivo supuso el testeo o *screening* de colecciones de compuestos químicos disponibles en modelos fenotípicos, mayormente en modelos *in vivo.* Mediante esta metodología se lograron descubrir compuestos tales como la clorpromazina, el meprobamato y las benzodiazepinas, todos los cuales se han convertido en fármacos exitosos (Ratti et al., 2001). Los modelos fenotípicos funcionaban, en cierta medida, como una caja negra, en tanto no se elucidaba por qué un compuesto testeado poseía actividad en el modelo utilizado, ni cómo podría ser optimizado para aumentar su actividad. La aproximación se caracterizaba por una tasa de éxitos relativamente baja, e implicaba largos tiempos de trabajo y costos elevados (Reddy et al., 1999). Como ya se dijo, muchas moléculas que demostraban tener actividad en los modelos *in vivo* durante el screening fenotípico tenían un mecanismo de acción desconocido, por lo que en el caso de falla debido a problemas de toxicidad y/o farmacocinética, el desarrollo de derivados del hit se veía condicionado a ensayos a prueba y error.

La filosofía del **diseño racional de fármacos** comenzó a instalarse a fines de la década de 1960. A diferencia del screening fenotípico, el diseño racional implicaba una comprensión teórica de cuál sería la diana farmacológica, cómo actuaría el

fármaco sobre dicho blanco y/o qué mecanismos conducirían a los efectos terapéuticos deseados (Adam, 2005). En dicha década, Hansch y Fujita introdujeron el concepto de relación estructura-actividad cuantitativa (*Quantitative Structure–Activity Relationship*, QSAR) para la predicción de propiedades fisicoquímicas de compuestos con un núcleo químico común (Hansch et al., 1964). El enfoque fue trasladado, posteriormente, al campo del **descubrimiento de fármacos asistido por computadora** (DFAC).

A finales de los años 60, se postuló también el concepto de farmacóforo (van Drie, 2012). Dicho concepto es uno de los más duraderos en el DFAC y actualmente se define como *una descripción abstracta de las características moleculares que son necesarias para el reconocimiento molecular de un ligando por su diana molecular, es decir, la estructura molecular mínima fundamental para producir la respuesta biológica* (Talevi, 2016). Aproximadamente diez años después, diversos grupos científicos comenzaron a incorporar la biología estructural en las etapas iniciales del descubrimiento de fármacos. Esto generó un cambio de paradigma hacia una estrategia centrada en el blanco molecular. Este nuevo paradigma se centraba en una proteína (cuya participación en un proceso patológico justificase su consideración como blanco de fármacos) y en el diseño de compuestos que interfieran o modularan su actividad, junto a la optimización de las interacciones moleculares y la selectividad hacia el blanco elegido (Margineanu, 2014; Ratti et al., 2001; Umashankar et al., 2015).

A finales del siglo XX, sin embargo, varias compañías farmacéuticas retornaron hacia una aproximación por "fuerza bruta", pero que empleaba tecnologías novedosas para optimizar el proceso (Schedler, 2006). Dos ejemplos emblemáticos de esto son el cribado de alto rendimiento (*High-Throughput Screening*, HTS) y la química combinatoria. El HTS consiste en la evaluación de grandes quimiotecas en una amplia variedad de ensayos *in vitro*. Para ello, se recurre a miniaturización y ensayos automatizados, permitiéndoles a los investigadores identificar rápidamente aquellos candidatos prometedores en grandes bibliotecas de compuestos. Por su parte, la química combinatoria supone la síntesis acelerada -habitualmente en fase sólida- de una gran cantidad de compuestos químicos estructuralmente relacionados, mediante la combinación de fragmentos moleculares o *sintones*

predefinidos (Smith et al., 2006). En esa misma década comienza a reconocerse la necesidad de identificar tempranamente candidatos con propiedades farmacocinéticas favorables; un hito, en este sentido, fue la masivamente aceptada regla de 5 de Lipinski, una regla empírica para guiar la identificación y optimización de potenciales compuestos terapéuticos para administración oral (Lipinski et al., 1997).

El siglo XXI inicia con la finalización del *Proyecto de Genoma Humano* (PGH) (Craig Venter et al., 2001; Lander et al., 2001). Además de identificar posibles nuevos blancos farmacológicos, uno de los objetivos del PGH fue la obtención de derechos de propiedad para el uso de dichos blancos (Broder et al., 2000; Ward, 2001). Diversas tecnologías vinculadas a la genómica y a la proteómica, incluidas las tecnologías de microarreglos (*microarrays*) y secuenciación, han sido utilizadas en el proceso de descubrimiento de fármacos para identificar y validar dianas farmacológicas, descubrir biomarcadores de enfermedades y diseñar terapias más eficaces (Neha et al., 2013).

En comparación con el HTS y la química combinatoria, DFAC intenta emplear una búsqueda mejor orientada y, por lo tanto, permitiría identificar nuevos compuestos con una fracción del costo de las mencionadas técnicas (Rahman et al., 2012). DFAC encuentra aplicación, principalmente, en tres instancias del proceso de descubrimiento de fármacos: en el proceso de selección virtual para identificar compuestos líderes para estudios experimentales; durante la optimización de las propiedades farmacodinámicas y farmacocinéticas de los compuestos líderes, y; en el diseño de nuevos fármacos (Sliwoski et al., 2013).

La bioinformática y la quimioinformática son herramientas fundamentales en el campo del DFAC. Entre otros aportes, permiten predecir estructura y función de productos génicos, dilucidar sitios de unión a ligando, predecir poses o configuraciones de unión e interacciones relevantes del complejo ligando-blanco molecular, cuantificar la diversidad molecular de quimiotecas, y muestrear representativamente grandes colecciones de compuestos químicos (Ahamed et al., 2020; Borrel et al., 2018; Dibyajyoti et al., 2013; Zhao et al., 2020).

Por último, cabe destacar que en la actualidad, la mejor comprensión de los desórdenes complejos ha dado lugar a un tercer paradigma en el que el foco se pone

en fármacos que sean capaces de modular simultáneamente la actividad de múltiples blancos farmacológicamente relevantes. A este nuevo paradigma se lo conoce como **farmacología de redes o farmacología de sistemas** (Margineanu, 2016; Masoudi-Nejad et al., 2013).

2.2. Métodos computacionales en el descubrimiento de fármacos

Como ya se mencionó, el DFAC es capaz de aumentar la tasa de aciertos de nuevos compuestos farmacológicos porque utiliza una búsqueda mucho más racional que el cribado exhaustivo tradicional o su versión moderna, el HTS. No sólo pretende explicar las bases moleculares de la actividad terapéutica, sino también predecir posibles derivados que mejoren la actividad. El DFAC se puede clasificar en dos grandes categorías generales: **metodologías basadas en la estructura** y **metodologías basadas en el ligando**.

El DFAC basado en la **estructura** se basa en el conocimiento de la estructura de la proteína objetivo para predecir el complejo de unión al ligando, incluyendo la pose y las energías de interacción. Por su parte, el DFAC basado en el **ligando** explota el conocimiento de moléculas activas e inactivas ya conocidas a través de búsquedas de similitud química o la construcción de modelos QSAR (Kalyaanamoorthy et al., 2011).

Generalmente, se prefiere el DFAC basado en la estructura cuando se dispone de datos estructurales de alta resolución de la proteína diana, mientras que se elige el DFAC basado en el ligando cuando hay poca o ninguna información estructural del blanco disponible (Figura 2.1). La meta fundamental del DFAC es diseñar compuestos que se unan eficiente y selectivamente al blanco molecular y que, a la vez, presenten propiedades de absorción, distribución, metabolismo, excreción y toxicidad (ADMET) adecuadas (Jorgensen, 2010). Una aplicación exitosa de estos métodos resultaría en un compuesto que podría ser validado *in vitro* e *in vivo*, y cuyo sitio de unión podría ser confirmado, idealmente a través de la elucidación de la estructura del co-cristal (Sliwoski et al., 2013).

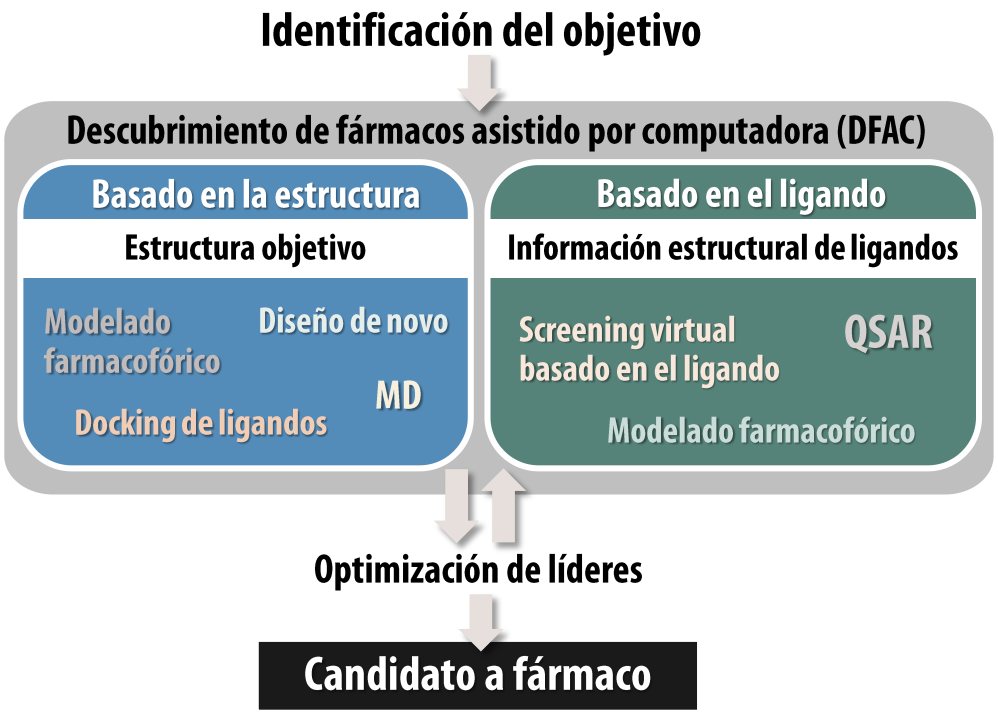


Figura 2.1. DFAC durante el desarrollo de nuevos fármacos. Usualmente se identifica un blanco molecular contra el que se debe desarrollar un fármaco. Dependiendo de la disponibilidad de información sobre la estructura, se utiliza un enfoque basado en la estructura o un enfoque basado en el ligando. Una campaña DFAC exitosa permitirá la identificación de múltiples compuestos líderes. Dicha identificación a menudo va seguida de varios ciclos de optimización. Finalmente, los compuestos lideres resultantes se prueban en estudios *in vivo* para identificar candidatos a fármacos.

A continuación, se describen las tres aplicaciones principales del DFAC:

(1) Cribado de bibliotecas virtuales de compuestos. También conocido como cribado farmacológico virtual de alto rendimiento (*virtual High-Throughput Screening*, vHTS) o screening virtual, permite a los investigadores concentrar los recursos en compuestos que tienen mayores probabilidades de presentar la actividad de interés. De esta manera, se pueden identificar compuestos activos analizando y ensayando una menor cantidad de compuestos, ya que los compuestos que con alta confianza se predice que serán inactivos pueden omitirse. Evitar el ensayo de una gran cantidad de compuestos (presuntamente) inactivos ahorra dinero y tiempo (Sliwoski et al., 2013). El vHTS puede realizarse de diversas formas, incluyendo búsquedas de similitud química por *fingerprints*, selección de compuestos por actividad biológica predicha a través de modelos QSAR, mapeo de farmacóforos y *docking* virtual de compuestos en la molécula

diana de interés (Enyedy et al., 2008). Estos métodos permiten la identificación de "hits" en la biblioteca. Es importante tener en cuenta que el vHTS no pretende identificar un compuesto farmacológico que esté listo para pruebas clínicas, sino más bien identificar andamiajes activos novedosos que no se hayan asociado previamente al blanco en cuestión (Sliwoski et al., 2013).

(2) Optimización de compuestos líderes durante el desarrollo de fármacos. El desarrollo de nuevas drogas, como ya hemos mencionado, puede tener costos del orden de miles de millones de dólares, contribuyendo a esa suma la síntesis y pruebas de los análogos de compuestos líderes (Basak, 2012). El costo comparativamente bajo del DFAC en comparación con la síntesis química y la caracterización biológica de los compuestos hace que estos métodos sean atractivos para enfocar, reducir y diversificar el espacio químico que se explora (Enyedy et al., 2008).

(3) El diseño de fármacos *de novo*. Esto supone el diseño de compuestos novedosos, en lugar de realizar un cribado de bibliotecas de compuestos ya conocidos. Para la construcción de las nuevas moléculas, se utilizan diferentes algoritmos que se definen generalmente como técnicas de enlace o de crecimiento. Los algoritmos de enlace implican el acoplamiento de pequeños fragmentos o grupos funcionales, como anillos, grupos acetilo, ésteres, etc., a diferentes partes del sitio de unión, a los que luego se los une adicionando fragmentos que permiten conectar los distintos grupos que interaccionan con los sitios adyacentes. Los algoritmos de crecimiento, por otro lado, comienzan a partir de un único fragmento colocado en el sitio de unión al que se agregan, eliminan y modifican los fragmentos para mejorar la actividad. Como en el caso del vHTS, el objetivo del diseño de fármacos *de novo* no es diseñar un único compuesto con una alta actividad y propiedades ADMET aceptables, sino más bien diseñar un compuesto líder que pueda mejorarse posteriormente (Sliwoski et al., 2013; Yuan et al., 2020).

Dado que el presente trabajo de tesis se enfoca en la predicción del parámetro farmacocinético $K_{p,uu}$, el cual es el resultado de diferentes procesos biológicos e interacciones con distintas estructuras celulares, las metodologías basadas en la

estructura no resultan de particular interés en este caso (a lo sumo, podrían servir para predecir interacciones con biomoléculas específicas relevantes para la biodisponibilidad central, como los transportadores ABC antes mencionados). Por lo tanto, se decidió aplicar metodologías basadas en el ligando.

Estas metodologías, a su vez, pueden clasificarse en tres grandes grupos: enfoques basados en el farmacóforo, metodologías basadas en la similitud molecular, y metodologías basadas en descriptores moleculares. La presente tesis se centra específicamente en las metodologías basadas en **descriptores moleculares**.

2.3. Descriptores moleculares

Los descriptores moleculares son índices numéricos que codifican información relacionada con la estructura molecular. Pueden ser propiedades fisicoquímicas experimentales de moléculas o índices teóricos calculados mediante fórmulas matemáticas o algoritmos computacionales (Todeschini et al., 2009). Todeschini y Consonni brindan la definición más aceptada hasta el momento: *"los descriptores moleculares son el resultado final de un procedimiento lógico y matemático que transforma la información química, codificada dentro de una representación simbólica de una molécula, en un número útil o en el resultado de un experimento estandarizado"* (Todeschini et al., 2000). Los descriptores moleculares se sustentan en diversas teorías o formalismos, como ser la química cuántica, la teoría de la información, la teoría de grafos, etc., y se utilizan para modelar una gran variedad de propiedades de los compuestos químicos, en campos de investigación tan variados como la toxicología, química analítica, fisicoquímica, química medicinal y química ambiental (Todeschini et al., 2009).

La información capturada por los descriptores moleculares puede variar desde simples propiedades a granel hasta variables tridimensionales complejas. En particular, se pueden usar diferentes niveles de complejidad, también conocidos como "dimensionalidad", para representar cualquier molécula dada (ver Figura 2.2). Acorde al nivel de dimensionalidad, los descriptores se clasifican en (Grisoni et al., 2018):

» **Descriptores 0-dimensionales (0D).** La representación molecular más simple es la fórmula química, es decir, la especificación de los elementos químicos que componen una molécula y el número de veces que cada uno aparece en la misma. Esta representación es independiente de cualquier conocimiento sobre conectividad atómica. Por lo tanto, los descriptores moleculares obtenidos de la fórmula química se denominan descriptores 0D. Los descriptores 0D son muy simples de calcular e interpretar, pero muestran un bajo contenido de información y un alto grado de degeneración, es decir, pueden tener valores iguales para moléculas diferentes. Algunos ejemplos de descriptores 0D son recuentos de átomos (por ejemplo, número de átomos de carbono), peso molecular y suma o promedio de propiedades atómicas (por ejemplo, volúmenes atómicos de van der Waals).

» **Descriptores 1-dimensionales (1D).** En una dimensión, las moléculas se perciben como un conjunto de subestructuras, como grupos funcionales o fragmentos centrados en átomos. Esta representación no requiere el conocimiento completo de las estructuras moleculares. La representación 1D de la molécula se refleja en los descriptores binarios que codifican la presencia/ausencia de subestructuras dadas, o en los descriptores que miden las frecuencias de ocurrencia de las mismas.

» **Descriptores 2-dimensionales (2D).** Esta representación agrega un nivel de información adicional a la representación 1D, al considerar también cómo están conectados los átomos. Por lo general, la molécula se representa como un grafo, cuyos vértices representan los átomos y los ejes o aristas, los enlaces. A partir de una representación de grafo, varios cuantificadores numéricos de topología molecular se derivan matemáticamente de manera directa o indirecta (por ejemplo, realizando operaciones algebraicas sobre matrices topológicas asociadas al grafo). Son comúnmente conocidos como **índices topológicos**, y codifican las denominadas propiedades topológicas (por ejemplo, adyacencia, conectividad), siendo generalmente sensibles a características estructurales como el tamaño, la forma, la simetría, la ramificación y la ciclicidad. A menudo, también se consideran propiedades químicas específicas de los átomos, por ejemplo, masa y electronegatividad (Consonni et al., 2012), o la presencia de

dadores/aceptores de enlaces de hidrógeno (Fechner et al., 2003; Reutlinger et al., 2013; Schneider et al., 1999). Por lo tanto, los índices topológicos pueden dividirse en dos categorías (Basak et al., 1997): (1) *índices topo-estructurales*, que codifican sólo información sobre adyacencia y distancias de enlace entre átomos; (2) *índices topo-químicos*, que cuantifican información sobre topología pero también propiedades químicas específicas de los átomos, como su identidad química y el estado de hibridación.

» **Descriptores 3-dimensionales (3D)**. Se puede agregar un nivel adicional de complejidad al percibir a la molécula no sólo en términos de tipos de átomos, conectividad y adyacencia, sino también como un objeto geométrico en el espacio, caracterizado por la configuración espacial de sus átomos. En otras palabras, la molécula se define en términos de tipos de átomos y sus coordenadas x-y-z. Los descriptores derivados de la representación 3D tienen un alto contenido de información (Kubinyi et al., 1998) y son particularmente útiles para modelar propiedades farmacéuticas y biológicas (Consonni et al., 2002; Nettles et al., 2006; Schuur et al., 1996). Cuando se trata de la representación 3D, se deben tener en cuenta varios problemas relacionados con la optimización geométrica de las moléculas, tales como la influencia del método de optimización en los valores de coordenadas (Rybinska et al., 2016); la presencia de más de un confórmero de energía mínima similar para moléculas altamente flexibles, y; la diferencia entre la geometría bioactiva y la geometría optimizada (el grado de deformación depende del número de enlaces capaces de rotar libremente en la molécula) (Nicklaus et al., 1995). Por estos motivos, el costo/beneficio del uso de descriptores 3D depende del caso, y siempre debe evaluarse cuidadosamente (Nettles et al., 2006).

» **Descriptores 4-dimensionales (4D).** Además de la geometría molecular, también se puede introducir una "cuarta dimensión", la cual generalmente intenta identificar y caracterizar cuantitativamente las interacciones entre la molécula y el sitio activo de un receptor biológico. Por ejemplo, las representaciones 4D pueden estar *basadas en conjuntos*, es decir, pueden incluir flexibilidad conformacional y libertad de alineación, a través de un conjunto de características espaciales de diferentes miembros de un conjunto de

entrenamiento (Andrade et al., 2010; Hopfinger et al., 1997), o representando a cada ligando por un conjunto de conformaciones, estados de protonación y/u orientaciones (Vedani, Briem, et al., 2000; Vedani, McMasters, et al., 2000).

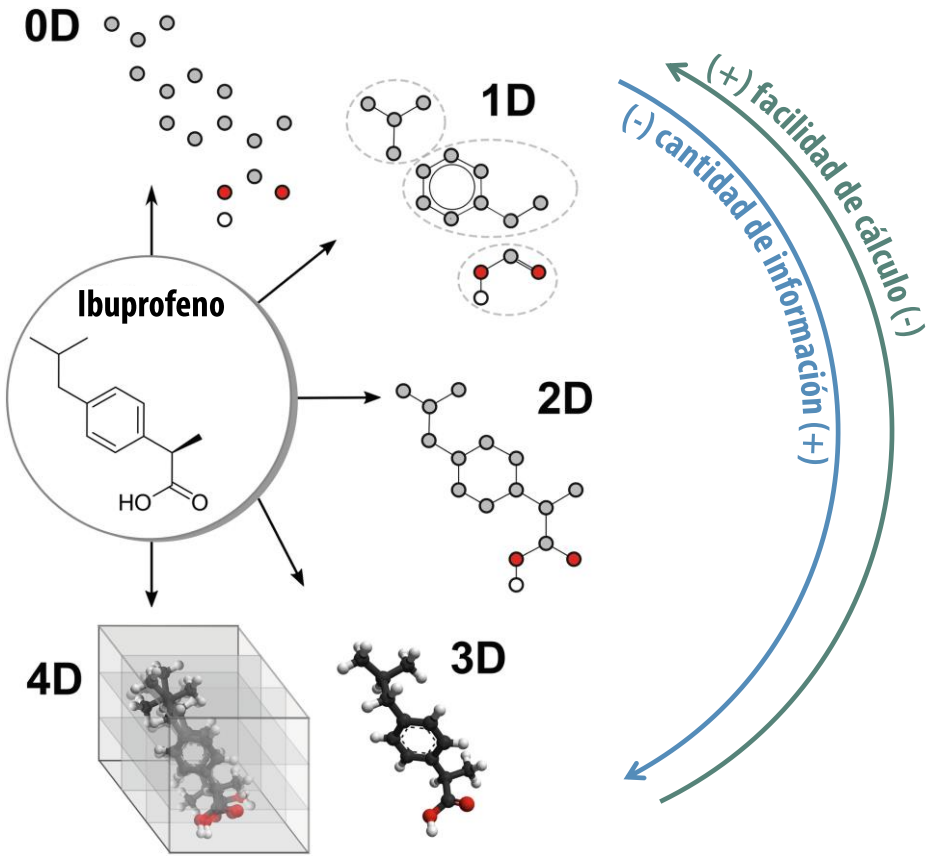


Figura 2.2. Ejemplo gráfico de diferentes representaciones moleculares de la misma estructura (ibuprofeno, aquí representado como una estructura 2D). También se representa la relación entre la dimensionalidad elegida y el contenido de la información/facilidad de cálculo.

Para un análisis más práctico, es posible englobar a los descriptores en dos grandes categorías de acuerdo con su dimensionalidad: los de baja dimensionalidad o independientes de la conformación (0, 1 y 2-dimensionales), y los de alta dimensionalidad o dependientes de la conformación (3 y 4-dimensionales).

2.4. Relación estructura-actividad cuantitativa (*Quantitative Structure–Activity Relationship*, QSAR)

Debido a su naturaleza numérica, los descriptores moleculares permiten capturar la información teórica que surge de la estructura molecular y vincularla con alguna

propiedad de la molécula (Todeschini et al., 2009). Por lo tanto, los descriptores moleculares se han convertido en el soporte de muchas aplicaciones de DFAC. Bajo esta perspectiva, el paradigma estructura-actividad se puede formular de la siguiente manera:

$$P = f(x_1, x_2, \dots, x_p) \tag{2.1}$$

donde P es la propiedad biológica/fisicoquímica que se desea explicar y/o predecir (la respuesta o salida del modelo), la cual se expresa como una función matemática de algunas características estructurales, codificadas por los p descriptores moleculares $(x_1, x_2, \dots, x_p)$. Una vez que la relación f se ha inferido a partir de un conjunto de entrenamiento, la propiedad P de un compuesto químico nuevo o no testeado se puede predecir en base a su estructura molecular, calculando los descriptores moleculares seleccionados. Este enfoque es lo que se conoce generalmente como **QSAR,** un método DFAC basado en el ligando que, como se comentó previamente, fue introducido hace más de 50 años por Hansch y Fujita (Hansch et al., 1964). Desde entonces y hasta ahora, QSAR sigue siendo un método eficiente para construir modelos matemáticos que intentan encontrar una relación estadísticamente significativa entre la estructura química y una propiedad biológica/fisicoquímica cuantitativa (pIC50, pEC50, Ki, etc.) o cualitativa (activa/inactiva, tóxico/no tóxico, etc.) utilizando generalmente técnicas de regresión o clasificación (Cherkasov et al., 2014).

Inicialmente, el modelado QSAR se limitó a pequeñas series de compuestos congenéricos, y métodos de regresión lineal. Hoy en día, el modelado QSAR ha crecido, se ha diversificado y evolucionado hacia el análisis de conjuntos de datos muy grandes que comprenden miles de estructuras químicas diversas, y que utilizan una amplia variedad de técnicas estadísticas y de aprendizaje automático (Cherkasov et al., 2014; Ekins et al., 2015; Goh et al., 2017; Mitchell, 2014).

La figura 2.3 muestra esquemáticamente las etapas e información necesaria para construir un modelo QSAR. Dichas etapas son (Grisoni et al., 2018):

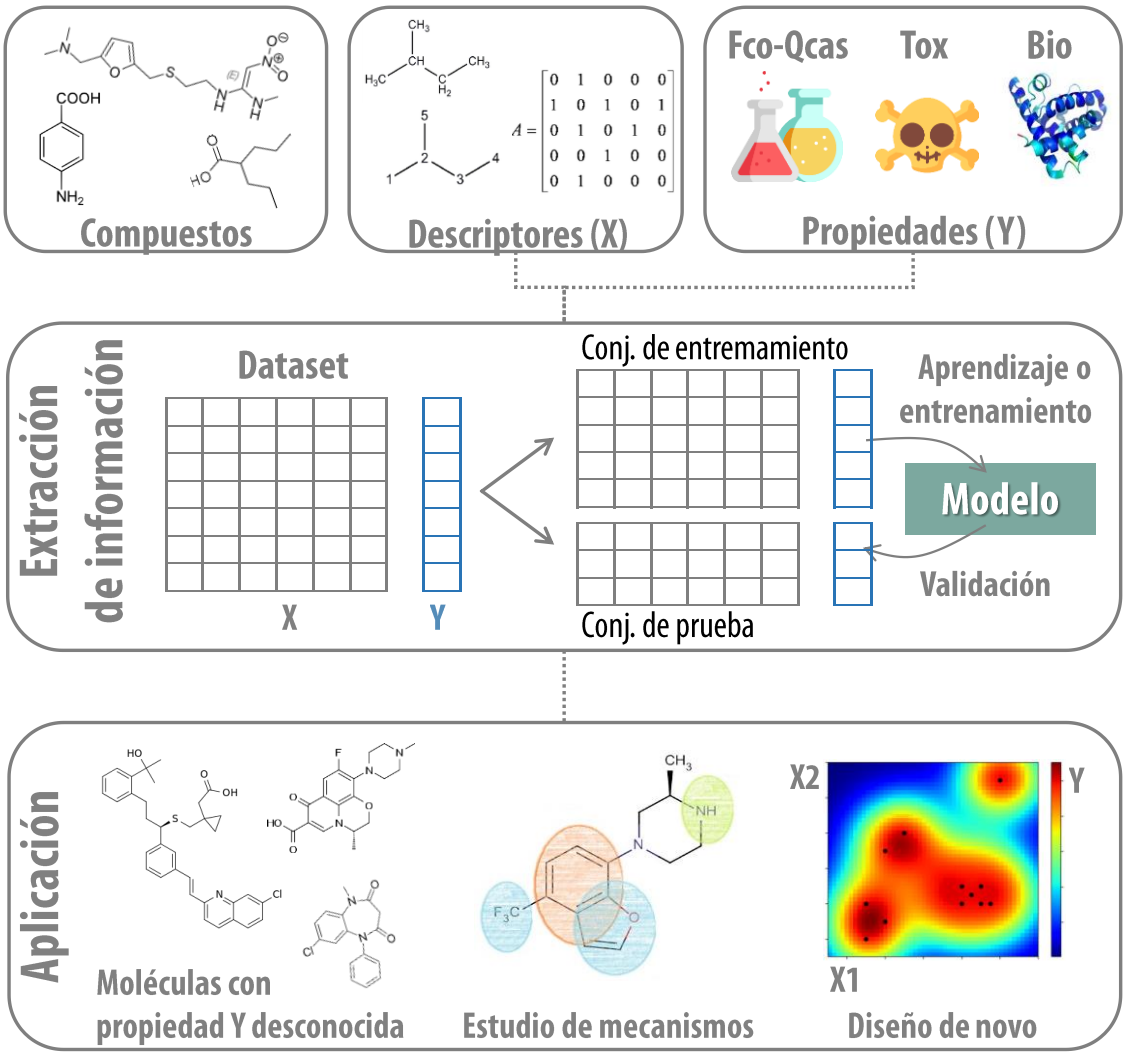


Figura 2.3. Pasos principales del desarrollo y uso del modelado QSAR: a partir de un conjunto de moléculas con propiedades experimentales conocidas (p. ej., fisicoquímicas, toxicológicas y biológicas), se pueden calcular varios tipos de descriptores moleculares. El conjunto de datos obtenido, descriptores moleculares más propiedades experimentales, se utiliza luego en la fase de extracción de la información para obtener un modelo QSAR confiable y validado. El modelo se puede aplicar más tarde para predecir las propiedades de moléculas no probadas, para obtener conocimientos del mecanismo a través de la interpretación de los descriptores moleculares y/o para diseñar moléculas novedosas.

1. **Selección y curado de datos.** La elección del conjunto inicial de moléculas, con sus correspondientes valores observados de la propiedad modelada, influirá en la cantidad y tipo de información capturada y transferida a los modelos. Cualquier error en los datos se propagará al modelo desarrollado y limitará su confiabilidad y aplicabilidad. Las estructuras moleculares, y sus representaciones, son el punto de partida del cálculo de los descriptores y, por lo tanto, un curado adecuado de dichas estructuras tiene un impacto directo en

el resultado del modelado (Fourches et al., 2010). Los errores en las estructuras pueden reflejarse en valores de descriptores erróneos y, en consecuencia, en resultados de modelos poco confiables. Al mismo tiempo, también es crucial asegurar que los datos experimentales y/o propiedades biológicas/fisicoquímicas a modelar sean confiables y no contengan errores ni valores anómalos para garantizar la utilidad del modelo (Furusjö et al., 2006). En esta etapa, el curado de datos (eliminar duplicados, corregir errores en las representaciones, eliminar ambigüedades en las estructuras y los identificadores químicos de los compuestos seleccionados (Mansouri et al., 2016) y analizar valores atípicos) y la evaluación de los protocolos experimentales (Grisoni et al., 2015), son dos aspectos claves para garantizar que no se introduzcan errores en el flujo de trabajo de modelado.

2. **Cálculo de los descriptores moleculares.** Una vez que todas las estructuras moleculares se han verificado y curado, los descriptores moleculares se calculan a partir de la representación molecular elegida, utilizando uno o más softwares disponibles. Los descriptores moleculares calculados se convierten entonces en las variables independientes que se utilizarán para desarrollar los modelos de interés.

3. **Extracción de información.** En esta fase, la información sobre la relación entre los descriptores moleculares y la propiedad a predecir se extrae y formaliza en un modelo matemático. Generalmente se requieren varios pasos, como (1) división de los datos en un conjunto de entrenamiento (*training set*) y en un conjunto de prueba (*test set*), el primero utilizado para la calibración del modelo, el segundo para la evaluación del modelo; (2) elección de la técnica de modelado apropiada, de acuerdo con los alcances del proyecto y el desempeño; (3) selección de variables (Cassotti et al., 2014; Derksen et al., 1992; Goldberg et al., 1988; Grisoni et al., 2014; Shen et al., 2004) para identificar los mejores descriptores para modelar la propiedad de interés y aumentar la estabilidad, el rendimiento y la capacidad de interpretación del modelo y; (4) evaluación del modelo a través de diferentes medidas, como el error cuadrático medio (RMSE) y la capacidad predictiva del modelo (Q^2) (Cramer, Bunce, et al., 1988;

Todeschini et al., 2016) para respuestas cuantitativas, y sensibilidad y tasa de buenas clasificaciones (Sokolova et al., 2009) para respuestas cualitativas.

Un factor adicional a considerar al desarrollar modelos de estructura-actividad es el llamado **dominio de aplicabilidad** (o dominio de aplicación) (Dragos et al., 2009; Sahigara et al., 2012), el cual se define como la región del espacio químico o, más precisamente, del espacio de los descriptores, donde las predicciones pueden considerarse confiables y se cumplen los supuestos del modelo. Las moléculas que caen fuera del dominio de aplicabilidad pueden ser muy diferentes de las moléculas utilizadas para calibrar el modelo y, por lo tanto, pueden caracterizarse por propiedades estructurales no representadas en los datos de entrenamiento o por valores de descriptores por fuera del rango de las moléculas del conjunto de entrenamiento (Grisoni et al., 2018).

4. **Utilización del modelo.** El modelo validado se puede usar para diversas aplicaciones, como predecir las propiedades de moléculas no probadas (Grisoni et al., 2015; Novič et al., 2010; Sabljic, 2001), diseñar nuevas moléculas con propiedades deseables (Miyao et al., 2010; Munteanu et al., 2010) y/o formalizar matemáticamente la relación entre las características estructurales y la propiedad de interés (Cramer, Patterson, et al., 1988; Grisoni et al., 2016; Marrero Ponce, 2004; Nembri et al., 2016).

Los expertos en I&D de la industria farmacéutica parecen coincidir en que la detección temprana de potenciales fallas en los ensayos clínicos es una estrategia clave para mejorar el éxito global y la productividad (FDA, 2004; Paul et al., 2010; Woosley et al., 2007) al reducir el tiempo y los gastos del desarrollo (Lombardo et al., 2003; Oprea, 2002; van de Waterbeemd et al., 2003), lo que a menudo se refiere como el paradigma *fallar temprano, fallar barato*.

En este contexto, los organismos reguladores más importantes del mundo han reconocido el papel clave que las tecnologías de quimio y bioinformática podrían desempeñar en el desarrollo de fármacos (Talevi et al., 2012). En particular, la metodología QSAR no sólo permite predecir la actividad de nuevos compuestos químicos diseñados sino que también permite eliminar aquellos candidatos con pobres propiedades farmacocinéticas o aquellos con alta probabilidad de mostrar

toxicidad en las etapas clínicas finales (Lipinski et al., 1997; O'Brien et al., 2005; Salum et al., 2009).

En la presente tesis se aplicará la teoría QSAR para el desarrollo de modelos computacionales que sean capaces de predecir el parámetro farmacocinético $K_{p,uu,}$ y, de esa manera, reducir las chances de que compuestos con altas probabilidades de presentar problemas de biodisponibilidad a nivel del SNC avancen a futuras instancias de investigación. Para ello, se utilizarán diversas técnicas de modelado y descriptores de baja dimensionalidad, que no requieren un análisis conformacional de las estructuras de entrenamiento ni de las de la base de datos sometida posteriormente a cribado, lo que reduce considerablemente el costo computacional y los tiempos necesarios para la etapa de modelado.

2.5. Breve revisión de los modelos *in silico* de $K_{p,uu}$ desarrollados hasta el momento

La Tabla 2.1 resume los modelos QSAR de regresión y clasificación *in silico* que han sido desarrollados hasta el momento para predecir el parámetro farmacocinético $K_{p,uu}$ (Chen et al., 2011; Fridén et al., 2009; Loryan et al., 2015; Varadharajan et al., 2015; Zhang et al., 2016a).

El primer modelo QSAR para predecir $K_{p,uu}$ fue reportado en 2009 por Fridén *et al.* (Fridén et al., 2009), el cual utilizaba un enfoque de regresión basado en un conjunto de entrenamiento de 41 fármacos con valores de $K_{p,uu}$ obtenidos por el método de *slice.* En términos de poder predictivo, el modelo tuvo un desempeño modesto (Q^2 = 0.452 y RMSE = 3.49 en el conjunto de prueba), tal vez debido al pequeño número de descriptores moleculares (16) considerados en el conjunto de posibles predictores. Sin embargo, el trabajo de Fridén inició una nueva forma de entender los factores que afectan la distribución de fármacos a través de la BHE, y su conjunto de datos disponible públicamente se ha utilizado posteriormente con fines de evaluación comparativa.

En un artículo posterior, Chen *et al.*, reportaron nuevos modelos de regresión/clasificación ampliando el conjunto de datos de Fridén con nuevos valores de $K_{p,uu}$ obtenidos de proyectos internos de descubrimiento de fármacos (es

decir, no disponibles públicamente) (Chen et al., 2011). Los autores construyeron dos tipos de modelos: los modelos directos, que predicen directamente $K_{p,uu}$, y los modelos indirectos, que predicen $V_{u,cerebro}$, K_p y $f_{u,plasma}$ para luego, mediante la Ec. 2.2, estimar $K_{p,uu}$. Su mejor modelo, una combinación o ensamblado de tres modelos (que incluía directos e indirectos), superó al modelo anterior de Fridén, logrando un porcentaje de buenas clasificaciones (%BC) del 85% en el conjunto de prueba. Luego, en 2015, el mismo grupo actualizó sus modelos en un artículo de Varadharajan y colaboradores (Varadharajan et al., 2015). Trabajando con un conjunto de datos más grande (expandido con 99 nuevos datos internos) y nuevos descriptores, construyeron un nuevo ensamblado de modelos. Según lo declarado por los autores, se logró cierta mejora en el poder predictivo, pero a expensas de una mayor complejidad.

$$K_{p,uu} = \frac{K_p}{V_{u,cerebro} \cdot f_{u,plasma}} \tag{2.2}$$

Cabe destacar que ninguno de estos trabajos explicitaba ningún paso de curado de la información, a la vez que se basaban en datos internos no divulgados, lo cual conspira contra la reproducibilidad.

También en 2015, Loryan *et al.* (Loryan et al., 2015) reportaron nuevos modelos QSAR para $K_{p,uu}$. Desarrollaron modelos de regresión PLS inferidos y validados a partir de un conjunto de 40 compuestos no divulgados (29 para fines de entrenamiento, 11 para validación externa). Los valores de $V_{u,cerebro}$ se estimaron utilizando el método de *slice* y los valores de K_p se determinaron *in vivo* en ratas o ratones, en condiciones de estado no estacionario. Los modelos arrojaron buenos resultados en el conjunto de prueba de 11 compuestos, pero mostraron un desempeño más discreto en el conjunto de datos de Fridén.

Los dos últimos modelos desarrollados hasta la fecha son los reportados por Zhang *et al.* (Zhang et al., 2016b) y Dolgikh *et al.* (Dolgikh et al., 2016). En el primer caso, los autores compilaron un conjunto de datos de 846 compuestos parcialmente revelados, con valores de $K_{p,uu}$ estimados a partir de valores *in vivo* de K_p de rata (tanto en condiciones de estado estacionario como no estacionario), mientras que las $f_{u,plasma}$ y $f_{u,cerebro}$ se extrajeron de experimentos *in vitro* de diálisis de equilibrio.

El mejor desempeño se obtuvo con un modelo clasificatorio, desarrollado utilizando la técnica de bosques aleatorios o *random forest* (RF), el cual logró un %BC del 73% contra el conjunto de prueba.

Por su parte, Dolgikh *et al.* (Dolgikh et al., 2016) aplicaron una estrategia similar a la de Chen *et al.* y Varadharajan *et al.*, mediante la construcción de modelos de regresión directa e indirecta que luego se usan como clasificadores. Curiosamente, los autores incluyeron información sobre el flujo de salida o eflujo debido a la glicoproteína P (P-gp) como una variable adicional. Los modelos se basaron en un conjunto de datos de casi 1000 valores internos de $K_{p,uu}$ calculados por combinación de valores de K_p obtenidos en ratones con $f_{u,plasma}$ y $f_{u,cerebro}$ (ambas estimadas por diálisis de equilibrio en diferentes especies) obtenidas a los 5 y 60 minutos después de una sola dosis intravenosa (es decir, en estado no estacionario). El mejor modelo reportado por Dolgikh *et al.* no superó a los anteriores cuando se desafió contra el conjunto de prueba externo (Tabla 2.1). Sin embargo, una posible ventaja es el dominio de aplicabilidad más amplio de este modelo, como lo sugiere el buen desempeño frente al conjunto de datos de Fridén.

De la discusión anterior, se puede observar que modelar $K_{p,uu}$ es una tarea compleja, principalmente debido a la escasez de datos experimentales disponibles públicamente y al ruido experimental asociado a la determinación experimental del parámetro. En la medida que los conjuntos de datos utilizados para fines de modelado y validación compilan datos experimentales obtenidos de diferentes laboratorios y mediante distintas técnicas experimentales, el paso de curado se convierte en un aspecto crucial del proceso de construcción de modelos.

Los modelos *in silico* que se informarán aquí para predecir el parámetro $K_{p,uu}$ se basan en un conjunto de datos disponible públicamente, con un conjunto de entrenamiento equilibrado, utilizando un procedimiento de curado explícito y siguiendo las buenas prácticas de desarrollo de modelos QSAR (Cherkasov et al., 2014; Tropsha, 2010).

Tabla 2.1. Resumen de los modelos QSAR desarrollados hasta el momento, utilizando $K_{p,uu}$ como parámetro de modelado. Cuando los autores obtuvieron más de un modelo del mismo conjunto de datos, se presentan los valores del mejor modelo.

Referencia	***Algoritmos***	***Descriptores***	***Conjunto de Datos***	***N***	***Conjunto de Prueba***				
					R^2	***RMSE***	***% BC***	***AUROC***	***MCC***
Friden *et al.* (Fridén et al., 2009)	PLS	16 descriptores 2D	*entrenamiento*	41	0.452 **(Q^2)**	3.99 **(x-fold)**	NA	NA	NA
			prueba	145					
Chen *et al.* (Chen et al., 2011)	SVM, RF	196 descriptores 2D y 3D	*entrenamiento*	173	0.58	0.46	85%	NA	0.72
			prueba	73					
Loryan *et al.* (Loryan et al., 2015)	PLS	188 descriptores 1D, 2D y 3D	*entrenamiento*	29	0.82	0.31	NA	NA	NA
			prueba	11					
Varadharajan *et al.* (Varadharajan et al., 2015)	SVM, RF	196 descriptores 2D, 3D & *signature*	*entrenamiento*	242	0.65	0.45	84.3%	NA	0.675
			prueba	104					
Zhang *et al*. (Zhang et al., 2016b)	Clasif bayesiano, RF, Árbol de decisión, NN, SVM, LDA	Varios conjuntos de descriptores fisicoquímicos y 2D.	*entrenamiento*	677	NA	NA	73%	0.77	0.52
			prueba	169					
Dolgikh *et al*. (Dolgikh et al., 2016)	SVM, RF	1800 descriptores 2D	*entrenamiento*	1030	0.53	0.57	80%	NA	0.6
			prueba	91					

PLS: mínimos cuadrados parciales; SVM: máquinas de soporte vectorial; RF: bosques aleatorios; NN: redes neuronales; LDA: análisis lineal discriminante; R^2: coeficiente de determinación; Q^2: R^2 de la validación cruzada interna; RMSE: error cuadrático medio; %BC: % de buenas clasificaciones; AUROC: área bajo la curva de la curva Característica Operativa del Receptor (ROC); MCC: coeficiente de correlación de Matthew; NA: no disponible.

2.6. Objetivos

Por todo lo dicho hasta aquí, en el presente trabajo de tesis nos planteamos los siguientes objetivos:

Objetivo General

Desarrollar modelos computacionales destinados a predecir el parámetro farmacocinético $K_{p,uu,}$ para ser aplicados como filtros *in silico* durante el desarrollo de nuevos fármacos para el tratamiento de desórdenes del SNC.

Objetivos Específicos

» Compilar, a partir de literatura, una base de datos de compuestos tipo-fármaco (*drug-like*) cuyo valor de $K_{p,uu}$ (coeficiente de reparto entre las concentraciones de fármaco libre en fluido intersticial cerebral y plasma en el estado estacionario) haya sido determinado experimentalmente.

» Utilizar la base de datos compilada en el punto anterior para el desarrollo y validación de modelos computacionales basados en descriptores topológicos de baja dimensionalidad, capaces de predecir la biodisponibilidad de un compuesto tipo-fármaco en el SNC.

» Evaluar la performance de los modelos desarrollados en el punto anterior.

» Validar experimentalmente las predicciones de los modelos generados.

Referencias

Adam, M. (2005). Integrating research and development: The emergence of rational drug design in the pharmaceutical industry. *Studies in History and Philosophy of Science Part C :Studies in History and Philosophy of Biological and Biomedical Sciences*, *36*(3), 513–537. https://doi.org/10.1016/j.shpsc.2005.07.003

Ahamed, T. K. S., & Muraleedharan, K. (2020). A cheminformatic study on chemical space characterization and diversity analysis of 5-LOX inhibitors. *Journal of Molecular Graphics and Modelling*, *100*, 107699. https://doi.org/10.1016/j.jmgm.2020.107699

Aminov, R. (2017). History of antimicrobial drug discovery: Major classes and health impact. *Biochemical Pharmacology*, *133*, 4–19. https://doi.org/10.1016/j.bcp.2016.10.001

Andrade, C. H., Pasqualoto, K. F. M., Ferreira, E. I., & Hopfinger, A. J. (2010). 4D-QSAR: Perspectives in Drug Design. *Molecules*, *15*(5), 3281–3294. https://doi.org/10.3390/molecules15053281

Basak, S. C. (2012). Chemobioinformatics: the advancing frontier of computer-aided drug design in the post-genomic era. *Current Computer-Aided Drug Design*, *8*(1), 1–2. https://doi.org/10.2174/157340912799218507

Basak, S. C., Gute, B. D., & Grunwald, G. D. (1997). Use of Topostructural, Topochemical, and Geometric Parameters in the Prediction of Vapor Pressure: A Hierarchical QSAR Approach. *Journal of Chemical Information and Computer Sciences*, *37*(4), 651–655. https://doi.org/10.1021/ci960176d

Borrel, A., Kleinstreuer, N. C., & Fourches, D. (2018). Exploring drug space with ChemMaps.com. *Bioinformatics*, *34*(21), 3773–3775. https://doi.org/10.1093/bioinformatics/bty412

Broder, S., & Venter, J. C. (2000). Sequencing the Entire Genomes of Free-Living Organisms: The Foundation of Pharmacology in the New Millennium. *Annual Review of Pharmacology and Toxicology*, *40*(1), 97–132. https://doi.org/10.1146/annurev.pharmtox.40.1.97

Cassotti, M., Grisoni, F., & Todeschini, R. (2014). Reshaped Sequential Replacement algorithm: An efficient approach to variable selection. *Chemometrics and Intelligent Laboratory Systems*, *133*, 136–148. https://doi.org/10.1016/j.chemolab.2014.01.011

Chen, H., Winiwarter, S., Fridén, M., Antonsson, M., & Engkvist, O. (2011). In silico prediction of unbound brain-to-plasma concentration ratio using machine learning algorithms. *Journal of Molecular Graphics & Modelling*, *29*(8), 985–995. https://doi.org/10.1016/j.jmgm.2011.04.004

Cherkasov, A., Muratov, E. N., Fourches, D., Varnek, A., Baskin, I. I., Cronin, M. et al. (2014). QSAR modeling: where have you been? Where are you going to? *Journal of Medicinal Chemistry*, *57*(12), 4977–5010. https://doi.org/10.1021/jm4004285

Consonni, V., & Todeschini, R. (2012). Multivariate Analysis of Molecular Descriptors. In *Statistical Modelling of Molecular Descriptors in QSAR/QSPR, Vol. 2*. Dehmer, M., Varmuza, K. & Bonchev, D (Eds), Wiley, Hoboken, NJ. pp. 111–147. https://doi.org/10.1002/9783527645121.ch4

Consonni, V., Todeschini, R., & Pavan, M. (2002). Structure/Response Correlations and Similarity/Diversity Analysis by GETAWAY Descriptors. 1. Theory of the Novel 3D Molecular Descriptors. *Journal of Chemical Information and Computer Sciences*, *42*(3), 682–692. https://doi.org/10.1021/ci015504a

Craig Venter, J., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G. et al. (2001). The sequence of the human genome. *Science*, *291*(5507), 1304–1351. https://doi.org/10.1126/science.1058040

Cramer, R. D., Bunce, J. D., Patterson, D. E., & Frank, I. E. (1988). Crossvalidation, Bootstrapping, and Partial Least Squares Compared with Multiple Regression in Conventional QSAR Studies. *Quantitative Structure-Activity Relationships*, *7*(1), 18–25. https://doi.org/10.1002/qsar.19880070105

Cramer, R. D., Patterson, D. E., & Bunce, J. D. (1988). Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. *Journal of the American Chemical Society*, *110*(18), 5959–5967. https://doi.org/10.1021/ja00226a005

Derksen, S., & Keselman, H. J. (1992). Backward, forward and stepwise automated subset selection algorithms: Frequency of obtaining authentic and noise variables. *British Journal of Mathematical and Statistical Psychology*, *45*(2), 265–282. https://doi.org/10.1111/j.2044-8317.1992.tb00992.x

Dibyajyoti, S., Talha Bin, E., & Swati, P. (2013). Bioinformatics:The effects on the cost of drug discovery. *Galle Medical Journal*, *18*(1), 44-50. https://doi.org/10.4038/gmj.v18i1.5511

Dolgikh, E., Watson, I. A., Desai, P. V, Sawada, G. A., Morton, S., Jones, T. M., & Raub, T. J. (2016). QSAR Model of Unbound Brain-to-Plasma Partition Coefficient, Kp,uu,brain: Incorporating P-glycoprotein Efflux as a Variable. *Journal of Chemical Information and Modeling*, *56*(11), 2225–2233. https://doi.org/10.1021/acs.jcim.6b00229

Dragos, H., Gilles, M., & Alexandre, V. (2009). Predicting the Predictability: A Unified Approach to the Applicability Domain Problem of QSAR Models. *Journal of Chemical Information and Modeling*, *49*(7), 1762–1776. https://doi.org/10.1021/ci9000579

Ekins, S., de Siqueira-Neto, J. L., McCall, L.-I., Sarker, M., Yadav, M., Ponder, E. L. et al. (2015). Machine Learning Models and Pathway Genome Data Base for Trypanosoma cruzi Drug Discovery. *PLoS Neglected Tropical Diseases*, *9*(6), e0003878. https://doi.org/10.1371/journal.pntd.0003878

Enyedy, I. J., & Egan, W. J. (2008). Can we use docking and scoring for hit-to-lead optimization? *Journal of Computer-Aided Molecular Design*, *22*(3–4), 161–168. https://doi.org/10.1007/s10822-007-9165-4

Fechner, U., Franke, L., Renner, S., Schneider, P., & Schneider, G. (2003). Comparison of correlation vector methods for ligand-based similarity searching. *Journal of Computer-Aided Molecular Design*, *17*(10), 687–698. https://doi.org/10.1023/B:JCAM.0000017375.61558.ad

FDA, Food and Drug Administration. (2004). *Innovation/Stagnation: Challenge and Opportunity on the Critical Path to New Medical Products.*Disponible en: *https://c-path.org/wp-content/uploads/2013/08/FDACPIReport.pdf.* Fecha de acceso: Diciembre 2021

Fourches, D., Muratov, E., & Tropsha, A. (2010). Trust, but verify: on the importance of chemical structure curation in cheminformatics and QSAR modeling research. *Journal of Chemical Information and Modeling*, *50*(7), 1189–1204. https://doi.org/10.1021/ci100176x

Fridén, M., Winiwarter, S., Jerndal, G., Bengtsson, O., Wan, H., Bredberg, U. et al. (2009). Structure-brain exposure relationships in rat and human using a novel data set of unbound drug concentrations in brain interstitial and cerebrospinal fluids. *Journal of Medicinal Chemistry*, *52*(20), 6233–6243. https://doi.org/10.1021/jm901036q

Furusjö, E., Svenson, A., Rahmberg, M., & Andersson, M. (2006). The importance of outlier detection and training set selection for reliable environmental QSAR predictions. *Chemosphere*, *63*(1), 99–108. https://doi.org/10.1016/J.CHEMOSPHERE.2005.07.002

Goh, G. B., Hodas, N. O., & Vishnu, A. (2017). Deep learning for computational chemistry. *Journal of Computational Chemistry*, *38*(16), 1291–1307. https://doi.org/10.1002/jcc.24764

Goldberg, D. E., & Holland, J. H. (1988). Genetic Algorithms and Machine Learning. *Machine Learning*, *3*(2-3), 95–99. https://doi.org/10.1023/A:1022602019183

Grisoni, F., Cassotti, M., & Todeschini, R. (2014). Reshaped Sequential Replacement for variable selection in QSPR: comparison with other reference methods. *Journal of Chemometrics*, *28*(4), 249–259. https://doi.org/10.1002/cem.2603

Grisoni, F., Ballabio, D., Todeschini, R., & Consonni, V. (2018). Molecular descriptors for structure–activity applications: A hands-on approach. In *Methods in Molecular Biology*. Nicolotti O. (Ed.), vol. 1800, Humana Press, NY. pp. 3–53. https://doi.org/10.1007/978-1-4939-7899-1_1

Grisoni, F., Consonni, V., Vighi, M., Villa, S., & Todeschini, R. (2016). Investigating the mechanisms of bioconcentration through QSAR classification trees. *Environment International*, *88*, 198–205. https://doi.org/10.1016/J.ENVINT.2015.12.024

Grisoni, F., Consonni, V., Villa, S., Vighi, M., & Todeschini, R. (2015). QSAR models for bioconcentration: is the increase in the complexity justified by more

accurate predictions? *Chemosphere*, *127*, 171–179. https://doi.org/10.1016/j.chemosphere.2015.01.047

Hansch, C., & Fujita, T. (1964). *p* -σ-π Analysis. A Method for the Correlation of Biological Activity and Chemical Structure. *Journal of the American Chemical Society*, *86*(8), 1616–1626. https://doi.org/10.1021/ja01062a035

Hopfinger, A. J., Wang, S., Tokarski, J. S., Jin, B., Albuquerque, M., Madhav, P. J., & Duraiswami, C. (1997). Construction of 3D-QSAR Models Using the 4D-QSAR Analysis Formalism. *Journal of the American Chemical Society*, *119*(43), 10509–10524. https://doi.org/10.1021/ja9718937

Jorgensen, W. L. (2010). Pulled from a protein's embrace. *Nature*, *466*(7302), 42–43. https://doi.org/10.1038/466042a

Kalyaanamoorthy, S., & Chen, Y. P. P. (2011). Structure-based drug design to augment hit discovery. *Drug Discovery Today*, *16*(17–18), 831–839. https://doi.org/10.1016/J.DRUDIS.2011.07.006

Kubinyi, H., Folkers, G., & Martin, Y. C. (1998). *3D QSAR in drug design*. Dordrecht: Kluwer/ESCOM, Leiden, ND.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J. et al. (2001). Initial sequencing and analysis of the human genome. *Nature*, *409*(6822), 860–921. https://doi.org/10.1038/35057062

Lipinski, C. A., Lombardo, F., Dominy, B. W., & Feeney, P. J. (1997). Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced Drug Delivery Reviews*, *23*(1–3), 3–25. https://doi.org/10.1016/S0169-409X(96)00423-1

Lombardo, F., Gifford, E., & Shalaeva, M. (2003). In Silico ADME Prediction: Data, Models, Facts and Myths. *Mini-Reviews in Medicinal Chemistry*, *3*(8), 861–875. https://doi.org/10.2174/1389557033487629

Lopez-Munoz, F., & Alamo, C. (2009). Monoaminergic Neurotransmission: The History of the Discovery of Antidepressants from 1950s Until Today. *Current Pharmaceutical Design*, *15*(14), 1563–1586. https://doi.org/10.2174/138161209788168001

Loryan, I., Sinha, V., Mackie, C., Van Peer, A., Drinkenburg, W. H., Vermeulen, A. et al. (2015). Molecular properties determining unbound intracellular and extracellular brain exposure of CNS drug candidates. *Molecular Pharmaceutics*, *12*(2), 520–532. https://doi.org/10.1021/mp5005965

Mansouri, K., Grulke, C. M., Richard, A. M., Judson, R. S., & Williams, A. J. (2016). An automated curation procedure for addressing chemical errors and inconsistencies in public datasets used in QSAR modelling . *SAR and QSAR in Environmental Research*, *27*(11), 911–937. https://doi.org/10.1080/1062936X.2016.1253611

Margineanu, D. G. (2014). Systems biology, complexity, and the impact on antiepileptic drug discovery. *Epilepsy & Behavior*, *38*, 131–142. https://doi.org/10.1016/j.yebeh.2013.08.029

Margineanu, D. G. (2016). Neuropharmacology beyond reductionism – A likely prospect. *Biosystems*, *141*, 1–9. https://doi.org/10.1016/j.biosystems.2015.11.010

Marrero Ponce, Y. (2004). Total and local (atom and atom type) molecular quadratic indices: significance interpretation, comparison to other molecular descriptors, and QSPR/QSAR applications. *Bioorganic & Medicinal Chemistry*, *12*(24), 6351–6369. https://doi.org/10.1016/j.bmc.2004.09.034

Masoudi-Nejad, A., Mousavian, Z., & Bozorgmehr, J. H. (2013). Drug-target and disease networks: polypharmacology in the post-genomic era. *In Silico Pharmacology*, *1*(1), 17. https://doi.org/10.1186/2193-9616-1-17

Mitchell, J. B. O. (2014). Machine learning methods in chemoinformatics. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, *4*(5), 468–481. https://doi.org/10.1002/wcms.1183

Miyao, T., Arakawa, M., & Funatsu, K. (2010). Exhaustive Structure Generation for Inverse-QSPR/QSAR. *Molecular Informatics*, *29*(1–2), 111–125. https://doi.org/10.1002/minf.200900038

Montinari, M. R., Minelli, S., & De Caterina, R. (2019). The first 3500 years of aspirin history from its roots – A concise summary. *Vascular Pharmacology*, *113*, 1–8. https://doi.org/10.1016/J.VPH.2018.10.008

Munteanu, C.R, Fernandez-Blanco, E., A. Seoane, J., Izquierdo-Novo, P., Angel Rodriguez-Fernandez, J., Maria Prieto-Gonzalez, J., … Pazos, A. (2010). Drug Discovery and Design for Complex Diseases through QSAR Computational Methods. *Current Pharmaceutical Design*, *16*(24), 2640–2655. https://doi.org/10.2174/138161210792389252

Neha, S., & Harikumar, S. L. (2013). Use of genomics and proteomics in pharmaceutical drug discovery and development: A review. *International Journal of Pharmacy and Pharmaceutical Sciences*, *5*, 24–28.

Nembri, S., Grisoni, F., Consonni, V., Todeschini, R., Nembri, S., Grisoni, F., … Todeschini, R. (2016). In Silico Prediction of Cytochrome P450-Drug Interaction: QSARs for CYP3A4 and CYP2C9. *International Journal of Molecular Sciences*, *17*(6), 914. https://doi.org/10.3390/ijms17060914

Nettles, J. H., Jenkins, J. L., Bender, A., Deng, Z., Davies, J. W., & Glick, M. (2006). Bridging Chemical and Biological Space: "Target Fishing" Using 2D and 3D Molecular Descriptors. *Journal of Medicinal Chemistry*, *49*(23), 6802–6810. https://doi.org/10.1021/jm060902w

Nicklaus, M. C., Wang, S., Driscoll, J. S., & Milne, G. W. (1995). Conformational changes of small molecules binding to proteins. *Bioorganic & Medicinal Chemistry*, *3*(4), 411–428. https://doi.org/10.1016/0968-0896(95)00031-b

Novič, M., Vračko, M., Novič, M., & Vračko, M. (2010). QSAR Models for Reproductive Toxicity and Endocrine Disruption Activity. *Molecules 15*(3), 1987–1999. https://doi.org/10.3390/MOLECULES15031987

O'Brien, S. E., & de Groot, M. J. (2005). Greater Than the Sum of Its Parts: Combining Models for Useful ADMET Prediction. *Journal of Medicinal Chemistry*, *48*(4), 1287–1291. https://doi.org/10.1021/jm049254b

Oprea, T. (2002). Virtual Screening in Lead Discovery: A Viewpoint. *Molecules*, *7*(1), 51–62. https://doi.org/10.3390/70100051

Paul, S. M., Mytelka, D. S., Dunwiddie, C. T., Persinger, C. C., Munos, B. H., Lindborg, S. R., & Schacht, A. L. (2010). How to improve R&D productivity: the pharmaceutical industry's grand challenge. *Nature Reviews Drug Discovery*, *9*(3), 203–214. https://doi.org/10.1038/nrd3078

Rahman, M. M., Karim, M. R., Ahsan, M. Q., Khalipha, A. B. R., Chowdhury, M. R., & Saifuzzaman, M. (2012). Use of computer in drug design and drug discovery: A review. *International Journal of Pharmaceutical and Life Sciences*, *1*(2). https://doi.org/10.3329/ijpls.v1i2.12955

Ratti, E., & Trist, D. (2001). The continuing evolution of the drug discovery process in the pharmaceutical industry. *Il Farmaco*, *56*(1–2), 13–19. https://doi.org/10.1016/S0014-827X(01)01019-9

Reddy, M. R., & Parrill, A. L. (1999). Overview of Rational Drug Design. In Parrill, A. B. & Reddy, M. R. (Eds) *Rational Drug Design, ACS Symposium Series,* Vol. 719, pp. 1–11. https://doi.org/10.1021/bk-1999-0719.ch001

Reutlinger, M., Koch, C. P., Reker, D., Todoroff, N., Schneider, P., Rodrigues, T., & Schneider, G. (2013). Chemically Advanced Template Search (CATS) for Scaffold-Hopping and Prospective Target Prediction for 'Orphan' Molecules. *Molecular Informatics*, *32*(2), 133–138. https://doi.org/10.1002/minf.201200141

Rybinska, A., Sosnowska, A., Barycki, M., & Puzyn, T. (2016). Geometry optimization method versus predictive ability in QSPR modeling for ionic liquids. *Journal of Computer-Aided Molecular Design*, *30*(2), 165–176. https://doi.org/10.1007/s10822-016-9894-3

Sabljic, A. (2001). QSAR models for estimating properties of persistent organic pollutants required in evaluation of their environmental fate and risk. *Chemosphere*, *43*(3), 363–375. https://doi.org/10.1016/S0045-6535(00)00084-9

Sahigara, F., Mansouri, K., Ballabio, D., Mauri, A., Consonni, V., & Todeschini, R. (2012). Comparison of Different Approaches to Define the Applicability Domain of QSAR Models. *Molecules*, *17*(5), 4791–4810. https://doi.org/10.3390/molecules17054791

Salum, L. B., & Andricopulo, A. D. (2009). Fragment-based QSAR: perspectives in drug design. *Molecular Diversity*, *13*(3), 277–285. https://doi.org/10.1007/s11030-009-9112-5

Schedler, D. J. A. (2006). Book review: Drug Discovery: A History, by Walter

Sneader. *Journal of Chemical Education*, *83*(2), 215. https://doi.org/10.1021/ed083p215.1

Schneider, G., Neidhart, W., Giller, T., & Schmid, G. (1999). "Scaffold-Hopping" by Topological Pharmacophore Search: A Contribution to Virtual Screening. *Angewandte Chemie International Edition*, *38*(19), 2894–2896. https://doi.org/10.1002/(SICI)1521-3773(19991004)38:19<2894::AID-ANIE2894>3.0.CO;2-F

Schuur, J. H., Selzer, P., & Gasteiger, J. (1996). The Coding of the Three-Dimensional Structure of Molecules by Molecular Transforms and Its Application to Structure-Spectra Correlations and Studies of Biological Activity. *Journal of Chemical Information and Computer Sciences*, *36*(2), 334–344. https://doi.org/10.1021/ci950164c

Shen, Q., Jiang, J.-H., Jiao, C.-X., Shen, G., & Yu, R.-Q. (2004). Modified particle swarm optimization algorithm for variable selection in MLR and PLS modeling: QSAR studies of antagonism of angiotensin II antagonists. *European Journal of Pharmaceutical Sciences*, *22*(2–3), 145–152. https://doi.org/10.1016/J.EJPS.2004.03.002

Sliwoski, G., Kothiwale, S., Meiler, J., & Lowe, E. W. (2013). Computational Methods in Drug Discovery. *Pharmacological Reviews*, *66*(1), 334–395. https://doi.org/10.1124/pr.112.007336

Smith, C. G., & O'Donnell, J. T. (2006). The process of new drug discovery and development, 2nd. Ed. Smith C. G. & O'Donnell J. T. (Eds), Informa Healthcare, Inc., NY.

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, *45*(4), 427–437. https://doi.org/10.1016/J.IPM.2009.03.002

Talevi, A. (2016). Network Pharmacology and Epilepsy. In: Talevi A., Rocha L. (eds) Antiepileptic Drug Discovery. Methods in Pharmacology and Toxicology. Humana Press, NY. https://doi.org/10.1007/978-1-4939-6355-3_18

Talevi, A., L. Bellera, C., Di Ianni, M., R. Duchowicz, P., E. Bruno-Blanch, L., & A. Castro, E. (2012). An Integrated Drug Development Approach Applying

Topological Descriptors. *Current Computer Aided-Drug Design*, *8*(3), 172–181. https://doi.org/10.2174/157340912801619076

Todeschini, R., Consonni, V., & Gramatica, P. (2009). Chemometrics in QSAR. In: Brown S. D. & Walczak, B. (Eds.) *Comprehensive Chemometrics*, Elsevier, The Netherlands. https://doi.org/10.1016/B978-044452701-1.00007-7

Todeschini, R., Ballabio, D., & Grisoni, F. (2016). Beware of Unreliable Q^2 ! A Comparative Study of Regression Metrics for Predictivity Assessment of QSAR Models. *Journal of Chemical Information and Modeling*, *56*(10), 1905–1913. https://doi.org/10.1021/acs.jcim.6b00277

Todeschini, R., & Consonni, V. (2000). Handbook of Molecular Descriptors. Wiley-VCH Verlag GmbH, Germany. https://doi.org/10.1002/9783527613106

Tropsha, A. (2010). Best Practices for QSAR Model Development, Validation, and Exploitation. *Molecular Informatics*, *29*(6–7), 476–488. https://doi.org/10.1002/minf.201000061

Umashankar, V., & Gurunathan, S. (2015). Drug discovery: an appraisal. *International Journal of Pharmacy and Pharmaceutical Sciences*, *7*(4), 59–66.

van de Waterbeemd, H., & Gifford, E. (2003). ADMET in silico modelling: towards prediction paradise? *Nature Reviews Drug Discovery*, *2*(3), 192–204. https://doi.org/10.1038/nrd1032

van Drie, J. H. (2012). Monty Kier and the origin of the pharmacophore concept. *Internet Electronic Journal of Molecular Design*, *11*(1), 271–279.

Varadharajan, S., Winiwarter, S., Carlsson, L., Engkvist, O., Anantha, A., Kogej, T. et al. (2015). Exploring In Silico Prediction of the Unbound Brain-to-Plasma Drug Concentration Ratio: Model Validation, Renewal, and Interpretation. *Journal of Pharmaceutical Sciences*, *104*(3), 1197–1206. https://doi.org/10.1002/jps.24301

Vecchio, I., Tornali, C., Bragazzi, N. L., & Martini, M. (2018). The discovery of insulin: An important milestone in the history of medicine. *Frontiers in Endocrinology*, *9,* 613. https://doi.org/10.3389/fendo.2018.00613

Vedani, A., Briem, H., Dobler, M., Dollinger, H., & McMasters, D. R. (2000). Multiple-

Conformation and Protonation-State Representation in 4D-QSAR: The Neurokinin-1 Receptor System. *Journal of Medicinal Chemistry*, *43*(23), 4416–4427. https://doi.org/10.1021/jm000986n

Vedani, A., McMasters, D. R., & Dobler, M. (2000). Multi-conformational Ligand Representation in 4D-QSAR: Reducing the Bias Associated with Ligand Alignment. *Quantitative Structure-Activity Relationships*, *19*(2), 149–161. https://doi.org/10.1002/1521-3838(200004)19:2<149::AID-QSAR149>3.0.CO;2-9

Walsh, G. (2013). Pharmaceutical biotechnology: concepts and applications. Wiley & Sons (Verlag). ISBN 978-1-118-68575-4 (ISBN)

Ward, S. J. (2001). Impact of Genomics in Drug Discovery. *BioTechniques*, *31*(3), 626–634. https://doi.org/10.2144/01313dd01

Woosley, R. L., & Cossman, J. (2007). Drug Development and the FDA's Critical Path Initiative. *Clinical Pharmacology & Therapeutics*, *81*(1), 129–133. https://doi.org/10.1038/sj.clpt.6100014

Yuan, Y., Pei, J., & Lai, L. (2020). LigBuilder V3: A Multi-Target de novo Drug Design Approach. *Frontiers in Chemistry*, *8*, 142. https://doi.org/10.3389/fchem.2020.00142

Zhang, Y. Y., Liu, H., Summerfield, S. G., Luscombe, C. N., & Sahi, J. (2016a). Integrating *in Silico* and *in Vitro* Approaches To Predict Drug Accessibility to the Central Nervous System. *Molecular Pharmaceutics*, *13*(5), 1540–1550. https://doi.org/10.1021/acs.molpharmaceut.6b00031

Zhang, Y. Y., Liu, H., Summerfield, S. G., Luscombe, C. N., & Sahi, J. (2016b). Integrating in Silico and in Vitro Approaches To Predict Drug Accessibility to the Central Nervous System. *Molecular Pharmaceutics*, *13*(5), 1540–1550. https://doi.org/10.1021/acs.molpharmaceut.6b00031

Zhao, J., Cao, Y., & Zhang, L. (2020). Exploring the computational methods for protein-ligand binding site prediction. *Computational and Structural Biotechnology Journal*, *18*, 417–426. https://doi.org/10.1016/j.csbj.2020.02.008

Capítulo 3

Metodologías de modelado

De manera análoga al análisis QSAR anteriormente descripto, el método de Relaciones Cuantitativas Estructura-Propiedad (en inglés, *Quantitative structure-property relationships*, QSPR) puede describirse de manera general como una aplicación de métodos estadísticos y análisis de datos para desarrollar modelos químico-matemáticos que puedan predecir con exactitud una propiedad biológica de compuestos químicos en función de sus estructuras (Tropsha, 2010).

Desarrollar modelos QSPR implica los siguientes pasos (Gramatica, 2013; Roy et al., 2011), que corresponden a los implementados en la etapa computacional del trabajo de tesis:

3.1. Conjuntos o bases de datos

- Recopilación y curado de un conjunto de datos de compuestos químicos a los que se les ha medido o determinado experimentalmente la propiedad que se desea modelar (variable respuesta).
- Partición del conjunto de datos en conjuntos de entrenamiento y de prueba.

› Cálculo de un conjunto de descriptores moleculares (variables explicativas).

3.2. Modelado

› Selección de variables independientes y generación de los modelos (métodos de modelado).

3.3. Evaluación del poder explicativo de los modelos

3.4. Evaluación de la robustez y poder predictivo de los modelos (validaciones interna y externa)

3.5. Ponderación de la importancia relativa de las variables independientes incorporadas a los modelos

3.6. Cálculo del dominio de aplicación de los modelos desarrollados

A continuación, se describen detalladamente las actividades desarrolladas correspondientes a cada uno de los puntos enumerados previamente.

3.1. Bases de datos

3.1.1. Recopilación y curado de las bases de datos de la propiedad de interés

En primer lugar, se realizó la compilación de un set de datos correspondientes a compuestos a los que se les hubiera medido previamente, de manera experimental, la propiedad de interés. En principio, por lo tanto, el set de datos quedó conformado por una representación molecular de los mencionados compuestos y su valor observado de la variable a modelar. En nuestro caso, la propiedad de interés a modelar es el parámetro $K_{p,uu}$, el cual, como se comentó en los capítulos anteriores, es el parámetro farmacocinético de mayor biorrelevancia para determinar la distribución del fármaco en el SNC (Morales et al., 2017). Por lo tanto, se realizó una búsqueda bibliográfica de compuestos con sus valores observados de $K_{p,uu}$ considerando publicaciones hasta febrero de 2017. A partir de tal búsqueda se obtuvieron datos de $K_{p,uu}$ de 711 compuestos.

Luego se realizó el curado de la base de datos utilizando diferentes criterios de inclusión/exclusión. Como primer criterio de inclusión, se consideraron sólo los valores de $K_{p,uu}$ obtenidos en condiciones de estado estacionario, o representativos del estado estacionario por transformación. En segundo lugar, únicamente los datos obtenidos por las técnicas de homogenato, microdiálisis o *slice* fueron considerados para los fines de modelado, mientras que los datos experimentales obtenidos por otras técnicas se descartaron. Cuando el valor de $K_{p,uu}$ de un compuesto estaba reportado por más de una de las técnicas mencionadas, los datos se priorizaron según la siguiente jerarquía: 1) microdiálisis; 2) *slice* y; 3) homogenato. Por último, se verificó que no hubiera compuestos duplicados. Después de estos criterios de selección, la base de datos se redujo a 157 compuestos.

Las estructuras moleculares de los compuestos seleccionados fueron estandarizadas como representación molecular 2D con el software Standardizer 16.7.4.0 (ChemAxon, 2016) acorde a la siguiente secuencia de comandos: *Strip Salt, Remove Solvents, Clear Stereo, Remove Absolut Stereo, Aromatize, Neutralize, Add explicit Hydrogen* y *Clean 2D.* **La base de datos así generada se denominó MSH (157 compuestos),** por las siglas de las técnicas experimentales consideradas (microdialisis, slice, homogenato).

A continuación, y a partir de la base de datos MSH, se generó un nuevo conjunto de datos que solamente contuviera compuestos con valores de $K_{p,uu}$ obtenidos por las técnicas de microdiálisis y *slice*, es decir, excluyendo la técnica de homogenato. **A esta nueva base de datos formada por 109 compuestos se la denominó MS.**

A su vez, al conjunto de datos anterior se lo refinó aún más para examinar la influencia de los sustratos de los transportadores ABC en los resultados de modelado. Para ello, se excluyeron los compuestos que, según la base de datos de DrugBank (Law et al., 2014), eran sustratos de la P-glicoproteína (P-gp) y/o de la Proteína de Resistencia al Cáncer de Mama (*Breast Cancer Resistance Protein*, BCRP), los dos transportadores de eflujo con mayores niveles de expresión en la BHE de individuos sanos (Al-Majdoub et al., 2019; Dauchy et al., 2008; Uchida et al., 2011). **El conjunto de datos resultante quedó formado por 67 compuestos, y se denominó "conjunto de datos MS refinado".** Por lo tanto, y a manera de resumen, se formaron tres bases de datos: MSH, MS y MS refinado, de 157, 109 y 67

compuestos, respectivamente. La figura 3.1 sintetiza el procedimiento para la obtención de los diferentes conjuntos de datos mencionados.

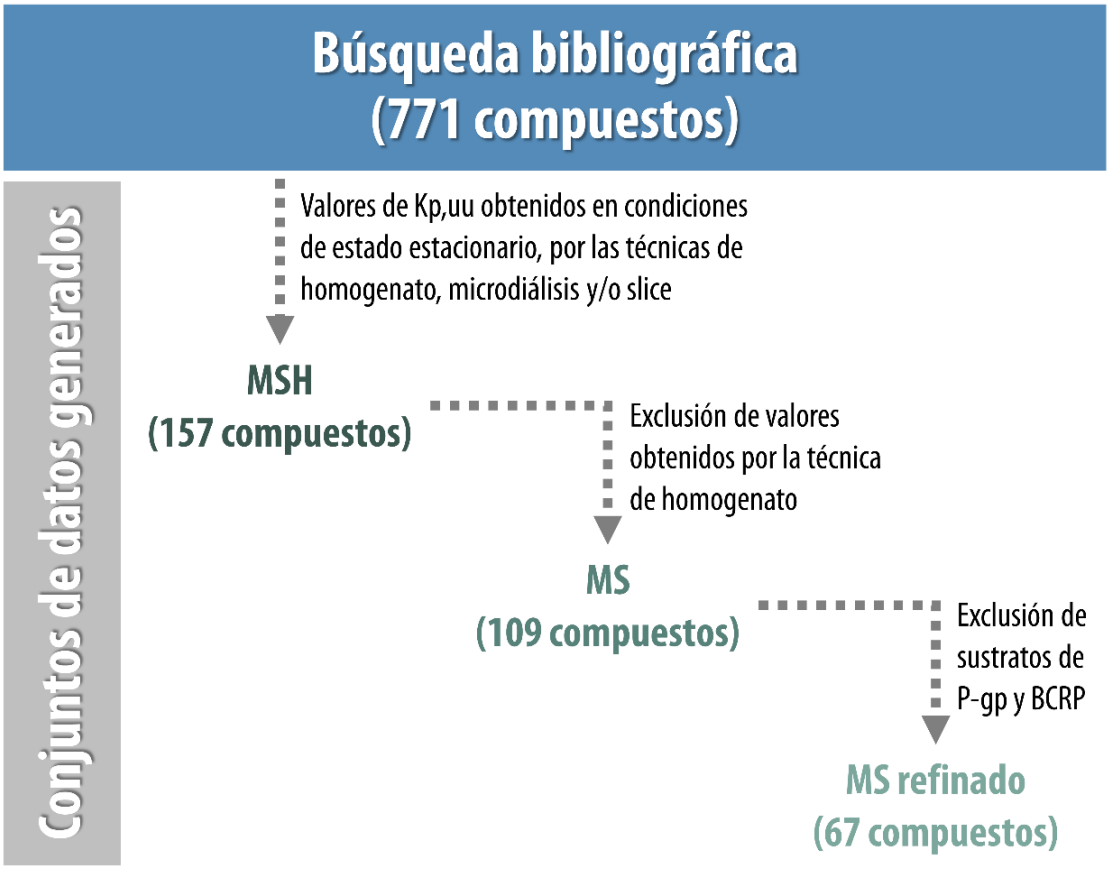


Figura 3.1. Flujo de trabajo para la generación de los tres conjuntos de datos utilizados en el presente trabajo.

Para mitigar el ruido vinculado a datos obtenidos de diferentes laboratorios y diferentes escenarios experimentales, se decidió explorar modelos clasificatorios (Alberca et al., 2016; Talevi, L. Bellera, et al., 2012). Por esta razón, definimos un esquema de clasificación binario (alta y baja biodisponibilidad -BD- del fármaco libre en SNC) usando para ello un valor de corte de $K_{p,uu}$ de 0,4. Es decir, se consideró que los compuestos con $K_{p,uu}$ mayor o igual al valor de corte ($K_{p,uu} \geq 0{,}4$) poseen alta BD libre en el SNC, mientras que se consideran de baja BD libre en el SNC a aquellos con $K_{p,uu}$ inferior al valor de corte ($K_{p,uu} < 0{,}4$). Este valor corte intenta ser más conservador que los usados previamente, ya que, como se comentó en el capítulo anterior, el valor de corte más alto reportado en estudios previos fue de 0,3,

correspondiente a los modelos de Zhang *et al.* (Y. Y. Zhang et al., 2016). Una vez clasificados los compuestos según el valor de corte propuesto, los conjuntos de datos MSH, MS y MS refinado consistieron en 74, 44 y 29 compuestos de alta y 83, 65 y 38 compuestos de baja BD libre en el SNC, respectivamente.

Con el objetivo de caracterizar la diversidad química y el espacio químico cubierto por los compuestos compilados, se realizaron gráficas de mapa de calor que ilustran la disimilaridad molecular entre las estructuras químicas, utilizando el paquete de R gplots (Warnes et al., 2016) La disimilaridad molecular se calculó con el paquete de R fingerprint (Guha, 2016), utilizando ECFP_6 como sistema de *fingerprints* o huellas digitales moleculares, y la distancia de Tanimoto como medida de disimilaridad. Adicionalmente, se efectuó un análisis de componentes principales (*Principal Component Analysis*, PCA), utilizando el paquete de R factoextra (Kassambara et al., 2017). Para este propósito, se utilizaron los siguientes ocho descriptores fisicoquímicos clásicos, calculados con Dragon 6.0 (Milano Chemometrics, 2011): peso molecular [MW], área de superficie polar topológica [TPSA (Tot)], coeficiente de partición octanol-agua de Moriguchi [MLOGP], número de átomos donantes para enlaces de hidrogeno (N y O) [nHDon], número de átomos aceptores para enlaces de hidrogeno (N, O, F) [nHAcc], número de enlaces rotables [RBN], número de anillos (número ciclomático) [nCIC] y suma de los volúmenes atómicos de van der Waals [Sv]. Estos descriptores son ampliamente reconocidos como parámetros clave en el descubrimiento de fármacos (Rankovic, 2017; Wager et al., 2010).

También se analizó la distribución de frecuencia de estos descriptores en los conjuntos de datos mediante la construcción de histogramas. Todas las gráficas fueron realizadas empleando el paquete de R ggplot2 (Wickham, 2009).

En el apartado de *Material suplementario*, al final de esta tesis, se presentan todas las estructuras químicas de los compuestos que constituyeron los diferentes conjuntos de datos.

3.1.2. Partición de los conjuntos de datos en los conjuntos de entrenamiento y de prueba

Generalmente, el set de datos recopilado y curado se divide en (al menos) dos subconjuntos de datos, llamados conjunto de entrenamiento y conjunto de prueba. El conjunto de entrenamiento se utiliza para calibrar los modelos, mientras que el conjunto de prueba se utiliza para verificar la capacidad predictiva de los modelos previamente ajustados con los datos del conjunto de entrenamiento. El objetivo de esta división es validar externamente los modelos desarrollados con compuestos que no han sido utilizados durante el procedimiento de ajuste (Martin et al., 2012).

En este paso, un interrogante relevante es cómo dividir al conjunto de datos en los conjuntos de entrenamiento y de prueba de manera tal que éstos sean representativos de toda la información química que se encuentra en el conjunto de datos. Una de las metodologías utilizadas para este propósito es la partición al azar, la cual ha demostrado ser una aproximación adecuada cuando se cuenta con un conjunto de datos de gran tamaño. Sin embargo, ese no es el caso aquí, ya que el conjunto más grande (MSH) cuenta con 157 compuestos. En estas situaciones, donde la base de datos posee una extensión acotada, se suele asignar la mayor parte de los compuestos al conjunto de entrenamiento, para poder contar con la mayor información posible a la hora de inferir los modelos, dejando por lo tanto un conjunto de prueba relativamente pequeño. Para realizar una partición con estas características, las **metodologías racionales** de muestreo han demostrado generar mejores resultados que los métodos al azar (Golbraikh et al., 2003; Leonard et al., 2006; Martin et al., 2012).

Basándonos en esta última premisa, se utilizó la herramienta de agrupamiento jerárquico LibraryMCS v16.7.4.0 (ChemAxon) en combinación con el algoritmo de optimización de agrupamiento *k-means* del lenguaje de programación R (R Core Team, 2017). El objetivo principal es distinguir dentro de una base de datos de amplia diversidad química, agrupamientos de estructuras con características similares, a fin de guiar la partición del set de datos en conjuntos de entrenamiento y de prueba representativos, asegurando así la mayor cobertura posible del espacio químico en ambos conjuntos.

LibraryMCS se basa en la máxima subestructura común (MSC) para agrupar un set de estructuras químicas sin recurrir a la comparación exhaustiva de a pares. Una subestructura común es definida como una subestructura presente en dos moléculas diferentes, conteniendo los mismos tipos de átomos y de enlaces químicos. La MSC es la subestructura común que contiene la mayor cantidad de átomos (Kawabata, 2011). Para el conjunto de estructuras consideradas, LibraryMCS construye una matriz de similitud y en base a los coeficientes de similitud selecciona el par de estructuras más similares (Talevi, A. Castro, et al., 2012), considerando que las dos estructuras con el mayor coeficiente de similitud tienen mayor probabilidad de presentar la MSC. Una vez que esta posible MSC se ha definido, se agrupan todos los compuestos del conjunto de datos que la incluyan (ver Figura 3.2).

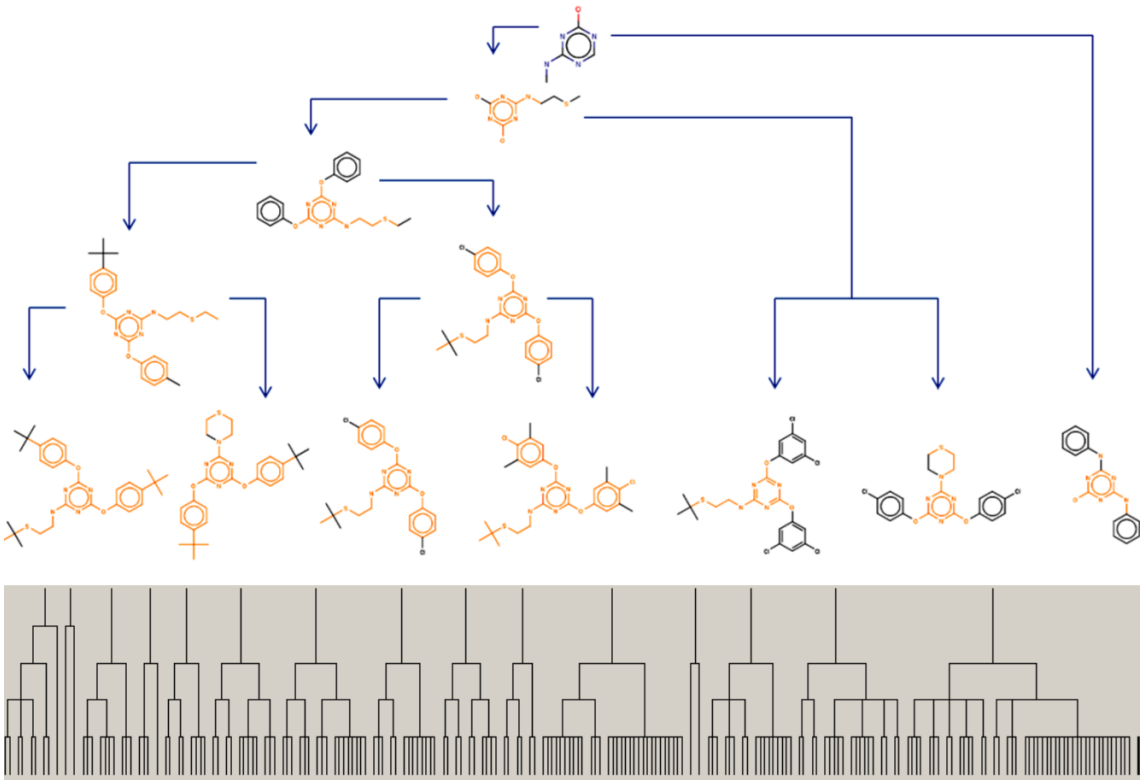


Figura 3.2. Se presenta un ejemplo de agrupamiento basados en MSCs. Las estructuras de la base de datos que no poseen una subestructura común con el resto con un mínimo número de átomos especificado por el usuario, son representadas de manera aislada en el dendrograma. Imágenes originales en *https://docs.chemaxon.com/display/docs/library-mcs-libmcs-clustering.md.*

El proceso se repite iterativamente hasta que no se encuentran más pares de estructuras con un valor del coeficiente de similitud por encima del valor de corte

utilizado por el algoritmo, o hasta que las estructuras sin agrupar no presenten una MSC de igual o mayor tamaño a la especificada por el usuario, en cuyo caso tales estructuras se presentan aisladas y constituyen compuestos atípicos, es decir, no se incluyen en ninguno de los agrupamientos (Hariharan et al., 2011). En este caso, se fijó que la MSC en torno a la cual se desarrolló el agrupamiento jerárquico debía tener al menos 9 átomos (valor por default del programa utilizado).

Los grupos de compuestos así obtenidos fueron optimizados luego utilizando el **algoritmo *K-means***, el cual constituye un algoritmo de optimización de agrupamientos (*clustering*) que procede de acuerdo a los siguientes pasos generales:

a) Encontrar o proponer una partición inicial de los N compuestos en K grupos (por ser una "partición", los K grupos no se superponen, y su suma devuelve el conjunto original)

b) Calcular la variación en un criterio de agrupamiento cuando un compuesto de un grupo se mueve a otro

c) Realizar un cambio que maximice la separación entre grupos (es decir, la mayor mejora en el valor del criterio de agrupamiento);

d) Repetir el proceso anterior hasta que ningún desplazamiento de compuestos de un grupo a otro produzca una mejoría.

En otras palabras, el método *K-means* inicia con un número (*K*) de grupos definidos por el usuario y mueve los objetos (compuestos, en este caso) entre esos grupos con el objetivo específico de: 1) minimizar la variabilidad dentro de los agrupamientos, y; 2) maximizar la variabilidad entre los grupos.

Para ello, utiliza la distancia euclídea al cuadrado como medida de las distancias entre los objetos y el centroide de los agrupamientos (Hartigan et al., 1979; Mauser et al., 1977); el criterio de agrupamiento es la minimización de la suma de las distancias euclídeas al cuadrado dentro de los grupos (Everitt et al., 2011). Es decir, lo que se busca minimizar es:

$$\sum_{k=1}^{K} \sum_{i=1}^{N_k} (x_{ki} - \bar{x}_k)^2 \qquad (3.1)$$

Donde K representa el número de grupos en los que se particionan los N compuestos, N_k es el número de compuestos en el grupo k, x_{ki} denota el vector p-dimensional de observaciones (descriptores moleculares) del compuesto i del grupo k, y $\bar{x}_k$ es el vector que representa la media de los elementos del grupo k en el espacio p-dimensional (considerando que se utilizan p variables, por ejemplo p descriptores moleculares, para caracterizar cada compuesto). A partir de la partición inicial, cada compuesto se relocaliza hacia el grupo cuya media le resulta más próxima, y el proceso se repite hasta encontrar convergencia.

En este caso, el número de grupos de la partición inicial se definió a partir del resultado del agrupamiento jerárquico de LibraryMCS. Las "semillas", es decir, aquellos elementos que se asignan inicialmente a los grupos y a partir de los cuales se construyen los agrupamientos iniciales, fueron elegidas aleatoriamente a partir de los agrupamientos obtenidos vía LibraryMCS para las categorías de alta y baja BD del fármaco libre en el SNC. Se permitieron 50 ciclos de optimización para alcanzar la convergencia.

Para caracterizar a cada objeto (compuesto químico) y definir el espacio p-dimensional en el que se aplicaría el agrupamiento *K-means*, se seleccionaron distintos descriptores moleculares representativos de diferentes aspectos de la molécula calculados con Dragon 6.0 (Milano Chemometrics, 2011): peso molecular (MW), logaritmo del coeficiente de reparto octanol-agua de Moriguchi (MLOGP), área de la superficie polar (TPSA(tot)), número de grupos aceptores de enlaces de hidrógeno (nHAcc), índice de información del contenido atómico (IAC) y suma de volúmenes atómicos de van der Waals (Sv). Estos descriptores fueron normalizados y aplicados para el cálculo de las distancias antes mencionadas.

En resumen, se aplicaron dos metodologías de agrupamiento en serie, de manera independiente para los compuestos de alta y baja BD en el SNC, para así formar un conjunto de entrenamiento con una composición de clase equilibrada, a fin de evitar sesgos hacia la categoría predominante, y representativo del espacio químico abarcado por el conjunto de datos en su totalidad. Los compuestos que no fueron asignados el conjunto de entrenamiento se utilizaron para el conjunto de prueba. Además, se tuvo en cuenta que en el conjunto de prueba haya un número representativo (al menos 10 compuestos) de cada clase. También durante la

partición se intentó mantener una relación de 70/30 entre el conjunto de entrenamiento y conjunto de prueba de la clase de compuestos menos representada en el set de datos. El procedimiento de partición de los conjuntos de datos previamente descrito en esta sección se realizó de manera separada par cada uno de los tres conjuntos de datos (Tabla 3.1).

Tabla 3.1. Composición de los conjuntos de datos MSH, MS y MS Refinado. Las filas superiores indican la cantidad de compuestos de cada clase presente en los conjuntos, y las filas inferiores indican cómo se particionaron en los conjuntos de entrenamiento y de prueba, en cada caso, indicando entre paréntesis la cantidad de compuestos con (alta BD SNC / baja BD SNC).

	Conjunto de datos		
	MSH	**MS**	**MS Refinado**
Número de compuestos			
Total	157	109	67
Con alta BD de fármaco libre en SNC	74	44	29
Con baja BD de fármaco libre en SNC	83	65	38
Particiones realizadas			
Conjunto de entrenamiento	110 (55/55)	60 (30/30)	38 (19/19)
Conjunto de prueba	47 (19/28)	49 (14/35)	29 (10/19)

3.1.3. Cálculo de descriptores

Después de curar las estructuras químicas, se calcularon los descriptores moleculares utilizando el software Dragon 6.0 (Milano Chemometrics, 2011), el cual permitió calcular 3668 descriptores independientes de la conformación para cada compuesto de la base de datos.

A continuación, se eliminaron los descriptores moleculares con valores faltantes para cualquier compuesto del conjunto de entrenamiento, así como aquellos con muy baja varianza. Para ello, se consideró que una variable independiente o descriptor era de muy baja varianza cuando se cumplían simultáneamente las condiciones siguientes: (1) la relación de frecuencia entre el valor más prevalente y

el segundo valor más frecuente del descriptor era superior a 95/5, y; (2) el número de valores únicos dividido por el número total de instancias (compuestos) estaba por debajo de 0,10. Como en el caso de la partición de las bases de datos, este procedimiento también se realizó de manera independiente en cada uno de los conjuntos de datos.

3.2. Modelado. Generación de modelos - Métodos de modelado

Como se comentó anteriormente, debido a la heterogeneidad de los valores experimentales de $K_{p,uu}$ de los compuestos de la base de datos, se decidió explorar la construcción de modelos clasificadores. En comparación con los modelos habitualmente llamados "de regresión", los modelos clasificadores son particularmente útiles para manejar datos con un alto nivel de "ruido", tales como los datos sujetos a gran variabilidad inter-laboratorio o datos obtenidos mediante métodos diversas (por ejemplo, microdiálisis y homogenato). Los **modelos de regresión** son aquellos cuya variable dependiente (usualmente denominada "y" o variable respuesta) es una variable continua (por ejemplo, IC50), que se espera sea predicha de manera exacta. Por este motivo, el rendimiento de este tipo de modelos depende fuertemente de la calidad y homogeneidad de los datos experimentales a partir de los cuales se infiere el modelo.

El rendimiento de los modelos clasificadores, en contraste, está menos influenciado por la variabilidad experimental de los datos, ya que las fuentes de ruido están particularmente restringidas a aquellos puntos que se encuentran en la frontera entre las categorías en cuestión (Talevi, L. Bellera, et al., 2012). Estos **modelos clasificadores** constituyen modelos cuantitativos basados en relaciones entre las variables independientes, en este caso los descriptores moleculares, y una variable dependiente de respuesta categórica (en este caso, binaria) que puede ser nominal (en este caso, "alta BD de fármaco libre en el SNC" o "alto $K_{p,uu}$" vs. "baja BD de fármaco libre en el SNC" o "bajo $K_{p,uu}$") o su correspondiente representación numérica (en este caso, "1" y "0", en ese orden) que representa la categoría de la clase del objeto que está siendo clasificado. El término "cuantitativo" de la sigla QSPR se refiere en este caso al valor numérico de las variables independientes (Mercader et al., 2010), de modo que los modelos clasificadores proporcionan una

respuesta cualitativa del tipo SI/NO (es decir, alta o baja BD de fármaco libre en el SNC) al tiempo que conservan el análisis cuantitativo a través de las variables independientes numéricas (Caballero et al., 2006).

Por lo tanto, definido el uso de modelos clasificatorios, se aplicaron las siguientes metodologías de modelado:

» Algoritmos clasificatorios de aprendizaje automático supervisado en lenguaje de programación R (R Core Team, 2017): k-vecinos más cercanos (*k*NN, *k-nearest neighbors*), cuadrados mínimos parciales clasificatorios (cPLS, *classificatory Partial Least Squares*), máquinas de soporte vectorial (SVM, *Support Vector Machine*), bosques aleatorios (RF, *Random Forest*), máquina de potenciación por gradiente estocástico (sGBM, *Stochastic Gradient Boosting Machine*) y de potenciación por gradiente extremo (XGBoost, *eXtreme Gradient Boosting*), provistos por los paquetes kernlab (Karatzoglou et al., 2004), gbm (Greenwell et al., 2019), caret (Kuhn, 2020), pls (Mevik et al., 2016), randomForest (Liaw et al., 2002) y xgboost (Chen et al., 2019), respectivamente.

» Redes Neuronales Profundas (DNN, *Deep Neural Networks*) a través de la librería Keras (Chollet et al., 2015) correspondiente al lenguaje de programación Python (http://www.python.org).

» Un método interno (*in-house*) de modelado -aprendizaje por ensamblado- basado en subespacios aleatorios (Alberca et al., 2016, 2018), en lenguaje R (R Core Team, 2017), el cual será descrito más adelante en esta sección.

La optimización de los hiperparámetros para cada algoritmo fue realizada mediante la búsqueda por grilla, utilizando la técnica de validación cruzada de K iteraciones, siendo K=10 y repitiendo el procedimiento 5 veces para robustecer el proceso de optimización (Douglas et al., 2003). La descripción de dichos hiperparámetros y el espacio de exploración para la optimización de cada uno de ellos se encuentra descripto en la sección correspondiente a cada algoritmo.

A continuación, se presenta una breve descripción de cada uno de los mencionados algoritmos.

3.2.1. *k*-Vecinos más cercanos (kNN)

Se trata de un método no paramétrico que clasifica las instancias o casos (aquí, compuestos químicos) en función de una medida de similitud (Cover et al., 1967; James et al., 2013). En este método, cada compuesto será clasificado por mayoría de votos de sus vecinos, y se asignará a la clase más frecuente entre sus *k* vecinos más cercanos, medida por una función de distancia (euclídea, en este caso). Si k = 5, entonces el compuesto se asigna a la clase (alto $K_{p,uu}$/ bajo $K_{p,uu}$) predominante entre sus cinco vecinos más cercanos en un espacio *p*-dimensional (considerando que se utilizan *p* variables, por ejemplo *p* descriptores moleculares). El algoritmo de base en *k*NN (representado esquemáticamente en la Figura 3.3) es el siguiente:

1. Estandarización de los datos.
2. Selección de la distancia a usar.
3. Optimización del número de vecinos (*k*).
4. Cálculo de la matriz de distancias.
5. Clasificación de compuestos según la clase más representativa de sus *k* vecinos más cercanos.

El algoritmo *k*NN proporciona buenos resultados cuando las superficies de separación de las clases (1/0) no son lineales, o son particularmente complejas (una clase contenida en otra). El único hiperparámetro a optimizar en esta metodología es el valor de *k*, el cual en nuestro caso se seleccionó por el método ya mencionado de búsqueda por grilla, en donde el espacio de exploración fueron todos los números enteros desde 1 a 50.

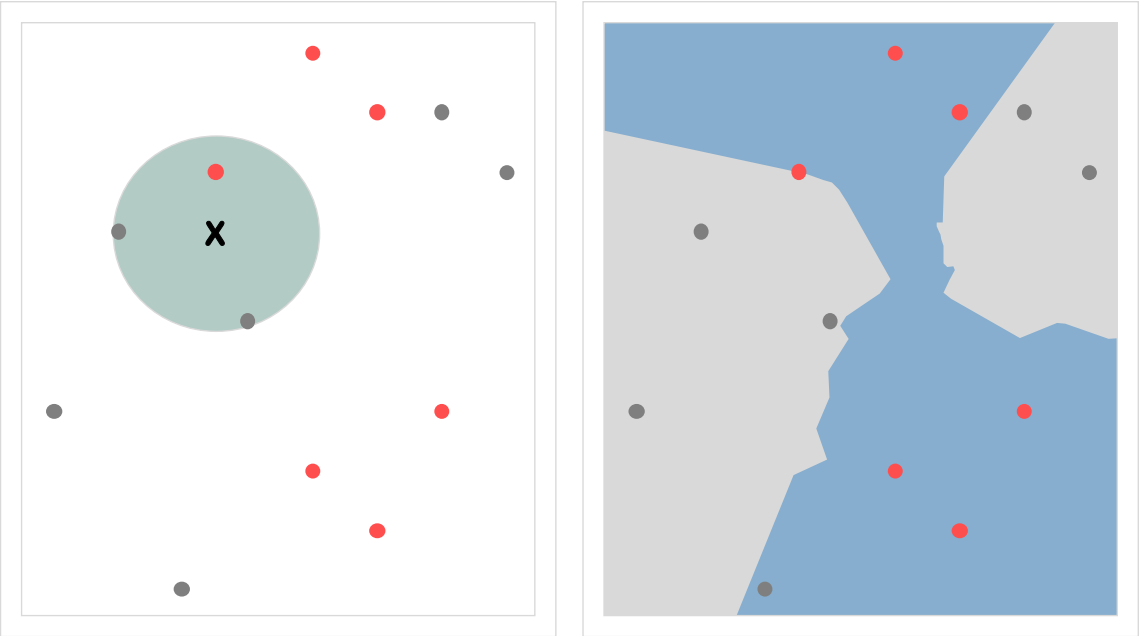


Figura 3.3. Representación de un modelo kNN, con *k*=3 y N=12, donde los compuestos de ambas clases (1/0) se representan como puntos rojos o grises, respectivamente. El gráfico de la izquierda muestra, para un compuesto X (para el cual se calculan los descriptores, pero se desconoce su clase), cómo funciona el algoritmo *kNN:* dado que *k*=3, se registra la clase de los tres vecinos más cercanos (en el círculo verde), asignándole al compuesto X la clase mayoritaria entre ellos (aquí, gris o clase 0). Procediendo de la misma manera para todos los puntos posibles, se llegan a establecer los límites de decisión o separación entre clases que se muestran en el gráfico de la derecha: cualquier observación (compuesto) que, por el valor de sus descriptores, caiga en la zona gris, será predicho como de clase 0, mientras que lo contrario sucede en la zona azul. *Imagen adaptada con permiso de James G., Witten D., Hastie T., Tibshirani R. (2013) An Introduction to Statistical Learning. Springer Texts in Statistics. Springer, NY.*

3.2.2. Cuadrados mínimos parciales clasificatorios (cPLS)

Los mínimos cuadrados parciales (PLS) surgieron como alternativa a la regresión por componentes principales (PCR, *principal component regression*). En particular, el análisis por PLS "clasificatorios" (cPLS) es una variante que se utiliza cuando la variable respuesta Y es categórica.

Tanto PLS como PCR son métodos de reducción de dimensionalidad, en los que primero se identifica un conjunto de características o variables Z_1, ..., Z_M que representan combinaciones lineales de las variables independientes originales, y luego se ajusta un modelo lineal por mínimos cuadrados utilizando estas nuevas M variables (se reduce la dimensionalidad original haciendo que $M < p$) (Hastie et al., 2009). Es decir:

$$Z_m = \sum_{j=1}^{p} \phi_{mj} X_j \tag{3.2}$$

Para algunas constantes $\phi_{m1}, \phi_{m2}, \dots, \phi_{mp}$.

$$y_i = \theta_0 + \sum_{m=1}^{M} \theta_m z_{im} + \epsilon_i \qquad i = 1,2, \dots, n \tag{3.3}$$

Mediante mínimos cuadrados ordinarios (MCO).

Reemplazando (3.2) en (3.3) se encuentra que:

$$\sum_{m=1}^{M} \theta_m z_{im} = \sum_{m=1}^{M} \theta_m \sum_{j=1}^{p} \phi_{mj} x_{ij} = \sum_{j=1}^{p} \sum_{m=1}^{M} \theta_m \phi_{mj} x_{ij} = \sum_{j=1}^{p} \beta_j x_{ij}$$

Donde:

$$\beta_j = \sum_{m=1}^{M} \theta_m \phi_{mj} \tag{3.4}$$

Así, el modelo (3.3) puede ser pensado como un caso especial del modelo original de regresión lineal. La reducción de dimensionalidad sirve para contraer los coeficientes estimados β_j, teniendo en cuenta que deben adoptar la forma (3.4).

PCR y PLS difieren en la manera de especificar las nuevas variables Z_m: mientras que en PCR se utilizan las componentes principales clásicas (es decir, nuevas variables que son combinaciones lineales de las variables originales X normalizadas, ortogonales y de máxima varianza, sin tener en cuenta a Y), en PLS se utiliza a la respuesta Y para generar las nuevas variables Z, de tal manera que éstas no sólo condensen los aspectos más informativos de las variables originales, sino que también estén correlacionadas con la respuesta. Por lo tanto, PLS busca encontrar las direcciones Z que permitan explicar tanto la variable respuesta como las predictoras.

La secuencia seguida por el algoritmo PLS es, en forma resumida, la siguiente:

1) Se estandarizan los *p* predictores

2) Se calcula la primera dirección Z_1, asignándole a cada ϕ_{1j} el coeficiente de la regresión lineal simple de Y vs. X_j. Se puede demostrar que este coeficiente es proporcional a la

correlación entre Y y X_j y, por consiguiente, que al operar de esta forma el algoritmo PLS da más peso a aquellas variables más asociadas con la respuesta.

3) Se calcula la regresión de Y vs. Z_1, para obtener $\hat{\theta}_1$

4) La dirección siguiente (Z_2) se encuentra ortogonalizando las X_j con respecto a Z_1, lo que de manera práctica se logra tomando los residuos de X_j vs. Z_1, para j=1,..., *p*. Este paso asegura que las variables Z_m son ortogonales, es decir, no se encuentran correlacionadas, ya que los residuos representan toda la información de las regresoras que no fue captada o explicada por Z_1.

5) Se repiten los pasos 2 – 4, con la matriz de residuos obtenida en 4) reemplazando a X_j, y se continúa el proceso hasta que se hayan obtenido las M direcciones Z_m.

6) Por último, se ajusta un modelo por MCO para predecir Y vs. Z_m.

Por lo tanto, un modelo de PLS trata de encontrar direcciones multidimensionales en el espacio de X para explicar la dirección de la máxima varianza multidimensional en el espacio Y. PLS es especialmente adecuada cuando la matriz de predictores tiene más variables que observaciones, y cuando hay correlación entre los valores de X (Barker et al., 2003), ambas situaciones presentes en nuestro caso.

Por lo dicho hasta aquí, se puede ver que el único hiperparámetro a optimizar en este algoritmo es el número de variables Z a incluir en el modelo final (es decir, M en las expresiones anteriores), ya que emplear el número óptimo de variables Z debería proporcionar el mejor modelo predictivo. La búsqueda de dicho parámetro se optimizó dentro del espacio de los números enteros desde 1 a 50.

3.2.3. Máquinas de Soporte Vectorial (SVM)

Este método encuentra aplicación, al menos en su forma clásica, en los escenarios de clasificación binaria, es decir, donde la variable respuesta puede tomar sólo dos valores o categorías. Es probablemente uno de los métodos basados en *kernel* más conocidos para el desarrollo de modelos (Bennett et al., 2000), y se caracteriza por su robustez (se ve menos afectado por datos duplicados) y bajo riesgo de sobreajuste (Tan et al., 2005). SVM se ha vuelto muy popular en los últimos años debido a sus aplicaciones en varios campos de reconocimiento de patrones como

bioinformática, medicina, economía y química (Basu et al., 2010; Conforti et al., 2010; Fernandez et al., 2010; Khorrami et al., 2010; Kim et al., 2010; Liew et al., 2009; Liew et al., 2010; Shen et al., 2010; Zuluaga et al., 2011).

En un espacio p-dimensional, un hiperplano es un subespacio "chato" de dimensión p-1, el cual matemáticamente se puede representar como:

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p = 0 \qquad (3.5)$$

Dado que un hiperplano divide al espacio p-dimensional en dos partes, cualquier observación en dicho espacio producirá, al calcular el lado izquierdo de la expresión (3.5), un valor igual, menor o mayor a cero según se encuentre sobre, por debajo o por encima del hiperplano, respectivamente.

Por lo tanto, para obtener un clasificador binario, lo ideal sería construir un hiperplano que separe perfectamente las observaciones del conjunto de entrenamiento de acuerdo con sus etiquetas de clase (-1/1). Si dicho hiperplano existe, cualquier observación nueva será clasificada según de qué lado del hiperplano se encuentre. Es decir, calculando el signo de (3.5). Por lo tanto, un hiperplano de separación verifica:

$$y_i\left(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}\right) > 0 \qquad \forall i = 1, \ldots, n \qquad (3.6)$$

Cuando dicho hiperplano existe (es decir, en el caso de ***observaciones linealmente separables***), una manera de elegir el mejor hiperplano entre los infinitos posibles es seleccionar el **clasificador de margen máximo** (MMC, *maximal margin classifier*), es decir, aquel que maximice la distancia (perpendicular) mínima de las observaciones al hiperplano, es decir, el *margen*.

Examinando la Figura 3.4, se ve que dos observaciones del conjunto de entrenamiento son equidistantes del MMC, y se encuentran a lo largo de las líneas que indican el ancho del margen. Estas observaciones se conocen como *vectores de soporte*, ya que son vectores en el espacio *p*-dimensional (en la Figura 3.4, *p* = 2) y "soportan" el MMC en el sentido de que, si estos puntos se movieran, entonces el MMC también se movería. Por lo tanto, el MMC depende directamente de los vectores de soporte, pero no de las otras observaciones: el movimiento de los puntos

por fuera del margen no afecta al MMC, siempre que en dicho movimiento la observación no cruce el límite establecido por el margen.

Para encontrar el MMC, se debe buscar el vector de β que maximice el margen (M) sujeto a las restricciones:

$$y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}) \geq M \qquad \forall i = 1, \ldots, n \tag{3.7}$$

$$\sum_{j=1}^{p} \beta_j^2 = 1 \tag{3.8}$$

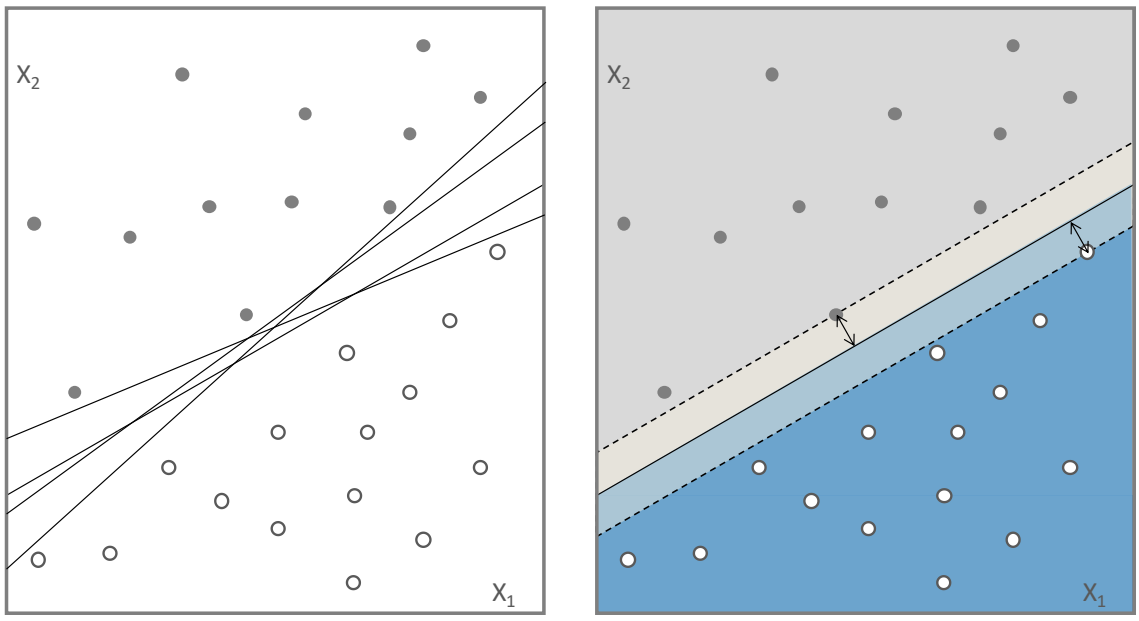


Figura 3.4. Los círculos grises llenos y vacíos representan las dos clases de observaciones. El gráfico de la izquierda muestra cuatro hiperplanos de separación posibles. En el gráfico de la derecha se observa el clasificador de margen máximo (MCC, línea continua). El margen es la distancia desde la línea continua a cualquiera de las líneas discontinuas. Los dos puntos que se encuentran en las líneas discontinuas son los vectores de soporte, y la distancia desde esos puntos al margen se indica mediante flechas. Las regiones grises y azules indican la regla de decisión creada por el hiperplano de separación. *Imagen adaptada con permiso de James G., Witten D., Hastie T., Tibshirani R. (2013) An Introduction to Statistical Learning. Springer Texts in Statistics. Springer, NY.*

Dado que M es una cantidad positiva, (3.7) asegura que cada observación se encuentra del lado correcto del hiperplano. Por su parte, se puede demostrar que la condición dada por (3.8) asegura que la distancia de la *i-ésima* observación al hiperplano está dada por $y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip})$ y, por lo tanto, M representa el margen del hiperplano encontrado (James et al., 2013).

Existen, sin embargo, casos donde los datos son ***no linealmente separables***. En estos casos, puede encontrarse una solución ampliando los conceptos anteriores para encontrar un hiperplano que logre separar *casi* todas las observaciones, generalización conocida como **clasificador de soporte vectorial** (SVC, *support vector classifier*). so, la separación sigue siendo lineal, aunque no perfecta. Matemáticamente, equivale a encontrar el vector β que maximice M sujeto a las restricciones:

$$\sum_{j=1}^{p} \beta_j^2 = 1 \tag{3.9}$$

$$y_i\left(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}\right) \geq M(1 - \epsilon_i) \tag{3.10}$$

$$\epsilon_i \geq 0, \quad \sum_{i=1}^{n} \epsilon_i \leq C \tag{3.11}$$

Donde C es un parámetro de optimización no negativo, y ϵ_i son variables de flexibilización, para permitir que observaciones individuales se encuentren en el lado incorrecto del margen o del hiperplano (ver Figura 3.5).

Si ϵ_i = 0, la observación se encuentra del lado correcto del margen, si es > a 0 se encuentra del lado incorrecto del margen, y si es > 1, la observación se encuentra del lado incorrecto del hiperplano. El parámetro C, por su parte, determina el número y la severidad de las violaciones permitidas a los márgenes y al hiperplano que serán toleradas. C=0 corresponde al caso del MMC (ecuaciones 3.7 – 3.8), y a medida que C crece más violaciones son permitidas, y más amplio será el margen. En otras palabras, valores grandes de C generan clasificadores con mayor sesgo y menor varianza, esto último debido a que incluyen más *vectores de soporte* (puntos en y dentro de los márgenes).

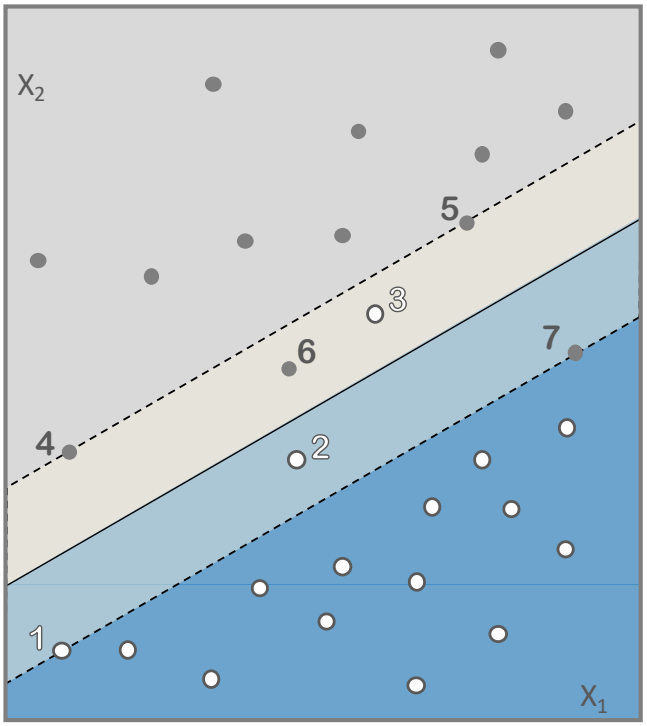


Figura 3.5. Ejemplo de un clasificador de soporte vectorial (SVC) lineal pero no perfecto. Las observaciones sin numerar se encuentran del lado correcto del margen. Observaciones grises: 4 y 5 se encuentran sobre el margen, 6 se encuentra del lado incorrecto del margen y 7 del lado incorrecto del hiperplano. Observaciones blancas: la observación 1 se encuentra sobre el margen, la 2 del lado incorrecto del margen y la 3 del lado incorrecto del hiperplano. *Imagen adaptada con permiso de James G., Witten D., Hastie T., Tibshirani R. (2013) An Introduction to Statistical Learning. Springer Texts in Statistics. Springer, NY.*

Por último, resta considerar los casos de límites no lineales entre clases, en los cuales será deseable usar funciones de los predictores, tales como cuadráticas, cúbicas o polinomios de mayor orden (ver, por ejemplo, la distribución de las observaciones en la Figura 3.6). Es decir, reemplazar la función lineal en los parámetros (expresión dentro del paréntesis a la izquierda de la ecuación (3.10)) por una función polinómica. Surgen así el concepto de SVM, como una extensión del concepto de SVC mediante el uso específico de *kernels* para generar límites no lineales entre las clases con bajo costo computacional. Un *kernel* es una función que cuantifica la similitud entre dos observaciones, $x_i \; y \; x_{i'}$, $\mathrm{K}(x_i, x_{i'})$. En SVM, los *kernel* más usados son:

Kernel polinomial de grado *d*: $$K(x_i, x_{i'}) = \left(1 + \sum_{j=1}^{p} x_{ij} x_{i'j}\right)^d \qquad (3.12)$$

Kernel de base radial (RBF): $$K(x_i, x_{i'}) = exp\left(-\frac{1}{2\sigma^2}\sum_{j=1}^{p}\left(x_{ij}-x_{i'j}\right)^2\right) \quad (3.13)$$

En el caso del kernel RBF, si una observación de prueba dada x^* está lejos de una observación de entrenamiento x_i en términos de distancia euclídea, entonces la sumatoria a la derecha de la expresión (3.13) será grande y, por lo tanto, $K(x_i, x_{i'})$ será pequeño, por lo que x_i prácticamente no jugará ningún papel en la predicción de la clase de x^*. Esto significa que el kernel radial tiene un comportamiento local, en el sentido de que sólo los puntos del conjunto de entrenamiento cercanos a una observación del conjunto de prueba tendrán efecto en su clase predicha.

Se ha utilizado SVM en diversos modelos QSPR y se ha demostrado que es robusto incluso cuando hay datos redundantes y superpuestos (Liew et al., 2009; Liew et al., 2010; Xue et al., 2006). Otra de sus ventajas es que es relativamente fácil de usar ya que hay sólo unos pocos parámetros definidos por el usuario. Por ejemplo, si se selecciona el kernel RBF, como fue en nuestro caso, sólo es necesario ajustar los parámetros C y σ. El espacio de exploración utilizado para ajustar el hiperparámetro C fue desde 2^{-5} hasta 2^{15}, y para el hiperparámetro σ fue desde 2^{-15} hasta 2^{3}.

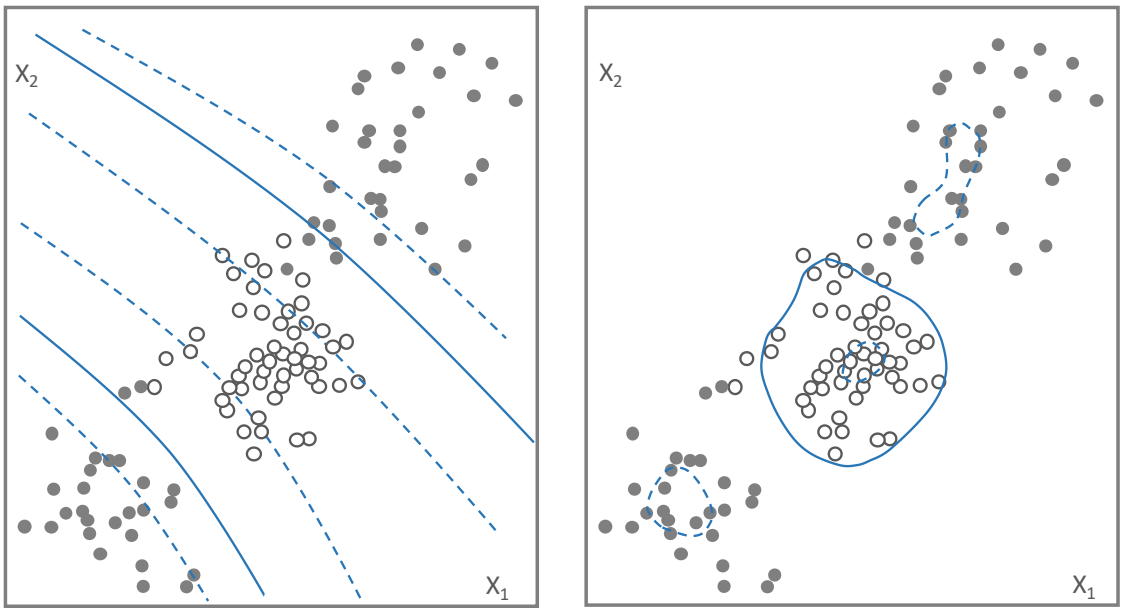


Figura 3.6. Izquierda: SVM con un kernel polinomial de grado 3 aplicada a datos donde un hiperplano lineal no sería adecuado. Derecha: para el mismo conjunto de entrenamiento, SVM de kernel radial. En este ejemplo, cualquiera de los kernel es capaz de capturar el límite de decisión. *Imagen adaptada con permiso de James G., Witten D., Hastie T., Tibshirani R. (2013) An Introduction to Statistical Learning. Springer Texts in Statistics. Springer, NY.*

3.2.4. Bosques aleatorios (RF)

Un árbol de decisión es una estructura con disposición jerárquica de nodos y ramas (Yee et al., 2012). Un árbol de decisión tiene tres tipos de nodos: un nodo raíz, nodos internos y nodos hoja. Un nodo raíz no tiene ramas entrantes, mientras que un nodo interno tiene una rama entrante y dos o más ramas salientes. Por último, los nodos hoja, también conocidos como nodos terminales, tienen una rama entrante y ninguna rama saliente. A cada nodo hoja se le asigna una propiedad objetivo, mientras que a un nodo que no es terminal (raíz o nodo interno) se le asigna un descriptor molecular que se convierte en una condición de prueba, que se ramifica en grupos de diferentes características.

La clasificación de un compuesto desconocido estará dada por el nodo hoja que alcanza después de pasar por una serie de preguntas (nodos) y respuestas (decidir qué ramas tomar), comenzando con la primera pregunta del nodo raíz. En el ejemplo de la Figura 3.7, un compuesto químico se clasificará de acuerdo con la propiedad objetivo *"y"*, si cumple una determinada condición para el descriptor molecular A. De lo contrario, se recurrirá, en el siguiente paso, al descriptor molecular B. En el ejemplo, si el valor de B es menor a 1, el compuesto analizado se etiquetará con la propiedad objetivo "*y*". De lo contrario, recibirá la etiqueta de propiedad objetivo *"ÿ"*.

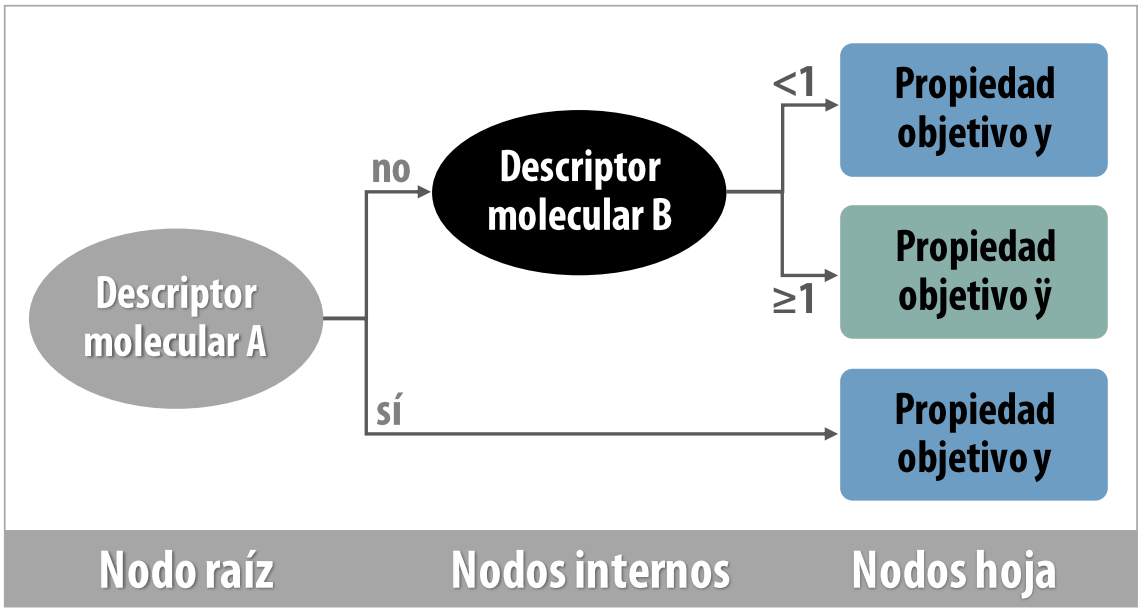


Figura 3.7. Ejemplo de árbol de decisión y sus 3 tipos de nodos.

Un árbol de decisión se construye subdividiendo sistemáticamente a los compuestos (casos o instancias) del conjunto de entrenamiento en base a reglas y relaciones. Con un conjunto dado de descriptores, se pueden construir muchas variaciones posibles de árboles, los cuales pueden tener diferentes desempeños. Existen algoritmos, como el CART (por sus siglas en inglés, *Classification and Regression Trees*) que puede utilizarse para la construcción de árboles de decisión (Breiman et al., 2017). Como primer paso, se examina si todas las instancias sin clasificar en el conjunto de entrenamiento pertenecen a una misma clase. Si es así, se creará un nodo hoja y todos estos casos se asociarán con este nodo. De lo contrario, se elige un descriptor molecular con un cierto umbral para dividir los casos en subconjuntos más pequeños, donde cada uno de estos subconjuntos forman un nuevo nodo hijo, y el proceso se repite desde el primer paso hasta que todos los casos de entrenamiento se asocian con un nodo hoja. El umbral de los descriptores moleculares que generan la mejor división se puede determinar comparando las "impurezas" (compuestos mal clasificados) en el nodo primario y los nodos secundarios; los nodos secundarios deben tener menos impurezas que el nodo primario y, por lo tanto, cuanto mayor es la diferencia de impurezas, mejor es el umbral seleccionado para dividir las muestras.

Los árboles de decisión tienen la ventaja de ser fácilmente interpretables, especialmente si son pequeños, y el desempeño del árbol de decisión no se ve tan fácilmente afectado por descriptores innecesarios. Sin embargo, un posible inconveniente es su susceptibilidad al sobreajuste. Para superar el problema del sobreajuste, se puede utilizar un bosque aleatorio (RF, del inglés, *Random Forest*).

RF utiliza la clasificación de consenso para reducir el problema del sobreajuste al tiempo que mejora el desempeño. El algoritmo RF (Breiman, 2001; James et al., 2013) es un método de "empaquetado" (*bagging*) de árboles que crea un conjunto (bosque) de árboles de decisión no correlacionados, y la predicción final se define por la votación mayoritaria de dicho conjunto. En cada árbol, un tercio del conjunto de entrenamiento se extrae aleatoriamente, y los dos tercios restantes se utilizan para modelar. El muestreo aleatorio es una forma de des-correlacionar los árboles mostrándoles diferentes conjuntos de entrenamiento. Cuando se construyen estos

árboles de decisión, y cada vez que se considera una división en un árbol, se elige una muestra aleatoria de predictores (como candidatos para la división) del conjunto completo de predictores. Esto también contribuye a la no-correlación entre los árboles, dado que evita que variables que son predictoras muy fuertes para la variable de respuesta, sean seleccionadas en muchos de los árboles. Cada árbol se hace crecer en la mayor medida posible. Por último, el árbol de decisión entrenado se usa para predecir el tercio restante de las observaciones que no se usaron para ajustar el modelo, y calcular el error-fuera-de-bolsa (*out-of-bag error*, OOB). De esta forma, RF maneja naturalmente la correlación entre los descriptores y no necesita un procedimiento de selección de descriptores por separado para obtener un buen rendimiento. Después del entrenamiento, se pueden hacer predicciones para muestras que previamente no han sido vistas por el modelo tomando el voto mayoritario de los árboles de decisión que forman el RF.

Los hiperparámetros a optimizar dentro de este algoritmo son el número de árboles que van a formar parte del RF y el número de variables muestreadas al azar como candidatos en cada división. En nuestro caso, el espacio de exploración para el número de árboles fue desde 50 hasta 3500, haciendo incrementos de a 50, y en el caso del número de variables a muestrear se decidió sondear los valores comprendidos entre la raíz cuadrada del número total de descriptores moleculares y el valor de la división del número total de variables dividido 3, redondeado hacia abajo (Hastie et al., 2009).

3.2.5. Máquina de Potenciación por Gradiente Estocástico (sGBM)

Las estrategias de potenciación o "*boosting"* son procedimientos de ensamblado (*ensemble*), que combinan varios modelos clasificadores "débiles" (aquellos que sólo representan una pequeña mejora respecto a la clasificación al azar) para producir un potente clasificador de consenso. En problemas de clasificación, la variable respuesta es binaria, es decir, por ejemplo, Y ∈ {-1; 1}. Dado un vector de variables predictoras X, un clasificador T(X) genera una clasificación predicha que puede tomar uno de dos valores, -1 o 1. El propósito del boosting es aplicar *secuencialmente* un algoritmo de clasificación débil al conjunto de datos (modificado en cada paso, ya que en cada oportunidad se toman los *pseduo* residuales r_n del paso anterior en

lugar de las y_i), para producir una secuencia de clasificadores débiles T_M. Las predicciones de todos ellos se combinan luego para producir la predicción final (Friedman, 2002; Svetnik et al., 2005).

Al tomar los residuales, el segundo modelo intenta "corregir" los errores del primero. Los modelos se agregan o ensamblan hasta que el conjunto de entrenamiento se predice perfectamente o hasta un número máximo (preespecificado) de modelos. Se llaman métodos de **boosting** porque la manera de operar del algoritmo *potencia* la performance de los árboles de decisión (u otro clasificador débil) utilizados.

En particular, describiremos cómo se usa este procedimiento para crear conjuntos de árboles de clasificación (clasificadores débiles), $\{T_1, \dots, T_M\}$, y una función aditiva, $F_M(T_1, \dots, T_M)$, que combina las salidas de estos árboles. Sea $X = \{x_1, \dots, x_p\}$ un vector p-dimensional de descriptores moleculares asociados a una molécula, e $Y(X)$ la variable binaria que representa la clase a la que pertenece la molécula (1 o -1).

Para un conjunto de entrenamiento de N compuestos, $\{(X_1, Y_1), \dots, (X_N, Y_N)\}$, el algoritmo de entrenamiento sGBM procede de la siguiente manera para construir una secuencia de M árboles:

1) $F_0(x) = 0$

2) Para $m = 1$ hasta M, hacer:

{Para $n = 1$ hasta N, hacer:

$$r_n = 2Y_n/(1 + \exp(2Y_n F_{m-1}(X_n)))$$

Finalizar n}

$$\{n_1, \dots, n_s\} = \textit{ submuestra aleatoria de } \{1, \dots, N\}$$

Ajustar un árbol (T_m) a los datos, $\{(r_{n_1}, X_{n_1}), \dots, (r_{n_s}, X_{n_s})\}$.

Luego:

$$F_m(T_1(X), \dots, T_m(X)) = F_{m-1}(T_1(X), \dots, T_{m-1}(X)) + \nu T_m(X)$$

Finalizar m

3) $\hat{Y}(X) = signo(F_M(X))$

Este algoritmo tiene dos características distintivas que explican el nombre boosting de "Gradiente Estocástico".

El término **gradiente** se refiere a que los valores r_n del algoritmo son gradientes de la función de pérdida $L(Y, F(X))$, la cual mide la diferencia entre la predicción y el valor observado:

$$r_n = \frac{\partial}{\partial F(X_n)} L(y_n, F(X_n)|_{F=F_{m-1}(X_n)}) \tag{3.14}$$

El algoritmo de clasificación utiliza la función de pérdida $-L = \log(1 + \exp(-2YF(X)))$, que se denomina función binomial de log-verosimilitud. Friedman demostró que la construcción de modelos aditivos se puede hacer de manera escalonada de tal manera que en cada etapa se construya un nuevo árbol ajustando los gradientes (Friedman, 2001). Este es un enfoque general que permite construir modelos aditivos de cualquier tipo, por ejemplo, una suma ponderada de redes neuronales (Schwenk et al., 2000), y/o que utilizan varias funciones de pérdida (Friedman, 2002).

Respecto a la componente **estocástica** que aparece en la denominación del método, se debe al hecho de que en cada iteración m, el árbol se ajusta a un subconjunto seleccionado aleatoriamente de tamaño s de los pseudo residuales $r_{n_i},\ i = 1, \dots, s$. Este muestreo aleatorio desempeña el papel de perturbación de datos similar a lo que sucedía en RF. Como en ese caso, la idea es aumentar la diversidad entre los árboles del conjunto (Friedman, 2002). A diferencia de RF, donde los árboles crecen hasta la longitud máxima, en sGBM los árboles son generalmente más pequeños, y el tamaño del árbol es un parámetro ajustable (ver a continuación).

Por lo dicho hasta aquí, se concluye que los hiperparámetros a optimizar en el algoritmo sGBM son varios. El primero es M, el número de árboles. El segundo es s, el tamaño del subconjunto aleatorio de valores de $\{r_n\}$ utilizados para construir cada árbol ($1 \ll s \leq N$). Y en tercer lugar se encuentra ν, llamado parámetro de contracción (*shrinkage*) o tasa de aprendizaje, introducido para mitigar el sobreajuste ($0 < \nu \leq 1$) (Friedman, 2001). Valores bajos de ν implican mayor contracción, o aprendizaje más lento.

El cuarto hiperparámetro del algoritmo es el tamaño de los árboles clasificadores entrenados en cada paso. El algoritmo utilizado para el modelado por sGBM en nuestro caso (el implementado por el paquete de R gbm (Greenwell et al., 2019)) tiene, a su vez, dos parámetros que permiten limitar el tamaño de los árboles clasificadores: (1) el número de divisiones del árbol, parámetro llamado "profundidad de interacción" (ID, *interaction.depth*), y; (2) el número mínimo de casos en un nodo hoja (*n.min.node*, cualquier nodo con menos de este número de muestras no se divide más). Por lo tanto, cuanto menor sean la ID y/o mayor sea el tamaño mínimo del nodo, más pequeño será el árbol.

En principio, es posible optimizar todos los parámetros de sGBM usando los datos de entrenamiento. Sin embargo, la optimización a través de, por ejemplo, una búsqueda por grilla mediante validación cruzada, generalmente requiere tiempos excesivamente largos (por su gran costo computacional). La literatura sugiere que el número de árboles (M) es uno de los parámetros más importantes (Svetnik et al., 2005). Cuando M aumenta más allá de un valor específico, el error en el conjunto de prueba comienza a aumentar (el error de entrenamiento siempre disminuye con M), lo que evidencia sobreajuste. Para evitarlo, se debe determinar el M óptimo para cada conjunto de datos. Para esto se realizó la búsqueda del número óptimo de árboles entre 100 y 3000, teniendo el resto de los hiperparámetros fijos en los siguientes valores (por default): ν = 0,01; ID = 1; *n.min.node* = 10; s/N = 0,5.

Un bajo valor de ν evita el sobreajuste, mientras que la optimización de M reduce el error en el conjunto de prueba. ID y *n.min.node* tienen valores bajos ya que, como se dijo anteriormente, en los métodos de *boosting* los clasificadores generados en cada iteración tienen en cuenta la información de los anteriores, por lo que el ensamblado de árboles pequeños continúa siendo eficiente, a la vez que el modelo resulta más interpretables debido a la menor complejidad de los clasificadores. La fracción de datos utilizada en cada etapa (s/N) suele fijarse en 0,5 (si bien para bases de datos grandes puede ser menor) ya que así se optimiza el costo computacional, a la vez que se mejora la exactitud de la predicción por aleatorización (Hastie et al., 2009).

3.2.6. Potenciación por Gradiente Extremo (XGBOOST)

Al igual que sGBM, XGBOOST es un meta-algoritmo para construir un clasificador robusto a partir de un conjunto de clasificadores débiles (Chen et al., 2016, 2019; Sheridan et al., 2016; Wang et al., 2018). Ambos algoritmos comparten la "estrategia aditiva" de optimización: dada una molécula i con un vector de descriptores X_i, un modelo de consenso usa funciones aditivas K para predecir el resultado:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(X_i) \ , \ f_k \in F \tag{3.15}$$

Aquí F es el conjunto de todos los posibles árboles de decisión. La función f_k en cada uno de los k pasos asigna a los valores de los descriptores en X_i un determinado valor de salida. XGBOOST intenta minimizar la siguiente función objetivo *regularizada*:

$$OBJ = \sum l(\hat{y}_i, y_i) + \sum \Omega(f_k) \tag{3.16}$$

$$donde\ \Omega(f) = \gamma T + \frac{1}{2}\lambda\|\omega\|^2 \tag{3.17}$$

En la ecuación 3.16, el primer término involucra una función de pérdida diferenciable, l, que mide la diferencia entre la predicción $\hat{y}_i$ y el objetivo y_i (análoga a la función L en sGBM). El segundo término, por su parte, tiene el objetivo de penalizar la complejidad del modelo, ya que como se ve en la ecuación 3.17, Ω es función de T y de ω: el número de hojas en el árbol y la puntuación en cada hoja, respectivamente. Los factores γ y λ, por su parte, son constantes que permiten controlar el grado de regularización o penalización. El término de regularización Ω ayuda a suavizar las ponderaciones finales para evitar el sobreajuste. Se espera que la función objetivo regularizada tenderá a seleccionar un modelo que emplee funciones simples y predictivas.

En XGBOOST, la función de pérdida se expande según la expansión de Taylor de segundo orden para optimizar rápidamente la función objetivo. Además de la función objetivo regularizada, la contracción y el submuestreo de columnas (descriptores moleculares) son dos estrategias adicionales que se utilizan para reducir aún más el sobreajuste (Chen et al., 2016; Sheridan et al., 2016). Después de cada paso de ajuste del modelo, la contracción escala las ponderaciones recién agregadas en un factor η (eta). Esto reduce la influencia de cada árbol y hace que el

modelo aprenda lentamente y, se espera, mejor. El submuestreo de columnas es una característica común con RF (Breiman, 2001), y consiste en considerar sólo un subconjunto aleatorio de descriptores al construir un árbol dado, lo que también acelera el proceso de ajuste al reducir el número de descriptores a considerar.

Por lo tanto, además de una mejor performance en ciertos escenarios, XGBOOST se caracteriza por implicar menor costo computacional que sGBM, lo que permite, además de procesar grandes bases de datos en tiempos moderados, optimizar todos los parámetros mediante validación cruzada y búsqueda en grilla.

Además de los hiperparámetros comunes con sGBM (número de árboles, profundidad de interacción, parámetro de contracción -aquí llamado eta- y fracción de muestras utilizadas), XGBOOST posee dos parámetros adicionales propios: la fracción de columnas muestreadas en cada iteración y la mínima reducción de la función de pérdida requerida para realizar una partición adicional en un nodo hoja del árbol (gamma). El espacio de exploración para la optimización de los diferentes hiperparámetros fue el siguiente:

- Numero de árboles: 50 a 1000, tomando cada 50.
- Profundidad de interacción: 2, 4 y 6.
- Eta: 0,003; 0,01 y 0,3.
- Gamma: 0 y 1.
- Fracción (submuestreo) de columnas utilizadas para construir cada árbol: 0,6, 0,8 y 1.
- Fracción de muestras utilizadas para construir cada árbol: 0,5, 0,75 y 1 (submuestreo del conjunto de entrenamiento).

3.2.7. Redes Neuronales Profundas (DNN)

Una red neuronal es una red compuesta de unidades operativas que, por analogía estructural y funcional, se denominan "neuronas". La Figura 3.8(A) muestra una neurona en su forma detallada y simplificada. Una neurona tiene múltiples entradas (flechas de entrada) y una salida (flecha de salida), y cada entrada está asociada con un peso w_i. Cada neurona también está asociada con una función, $f(z)$, llamada función de activación (que genera el output o salida de esa neurona), y un término de sesgo predeterminado b. Por lo tanto, cuando un vector de descriptores de

entrada $X = [x_1 \dots x_N]$ de un compuesto atraviesa una neurona, su salida puede representarse matemáticamente por la siguiente ecuación:

$$O = f(\sum_{i=1}^{N} w_i x_i + b) \tag{3.18}$$

Una DNN, a su vez, se construye a partir de varias capas de neuronas, como se ilustra en la Figura 3.8(B). Normalmente, hay tres tipos de capas en una DNN:

1. la capa de entrada (capa inferior en la Fig. 3.8 (B)), donde se ingresan los valores de descriptores moleculares de una molécula.
2. la capa de salida (que en la figura se ha representado como la capa superior) donde se generan las predicciones.
3. las capas ocultas o intermedias, a las cuales hace referencia la palabra "profundas" en el nombre de la metodología.

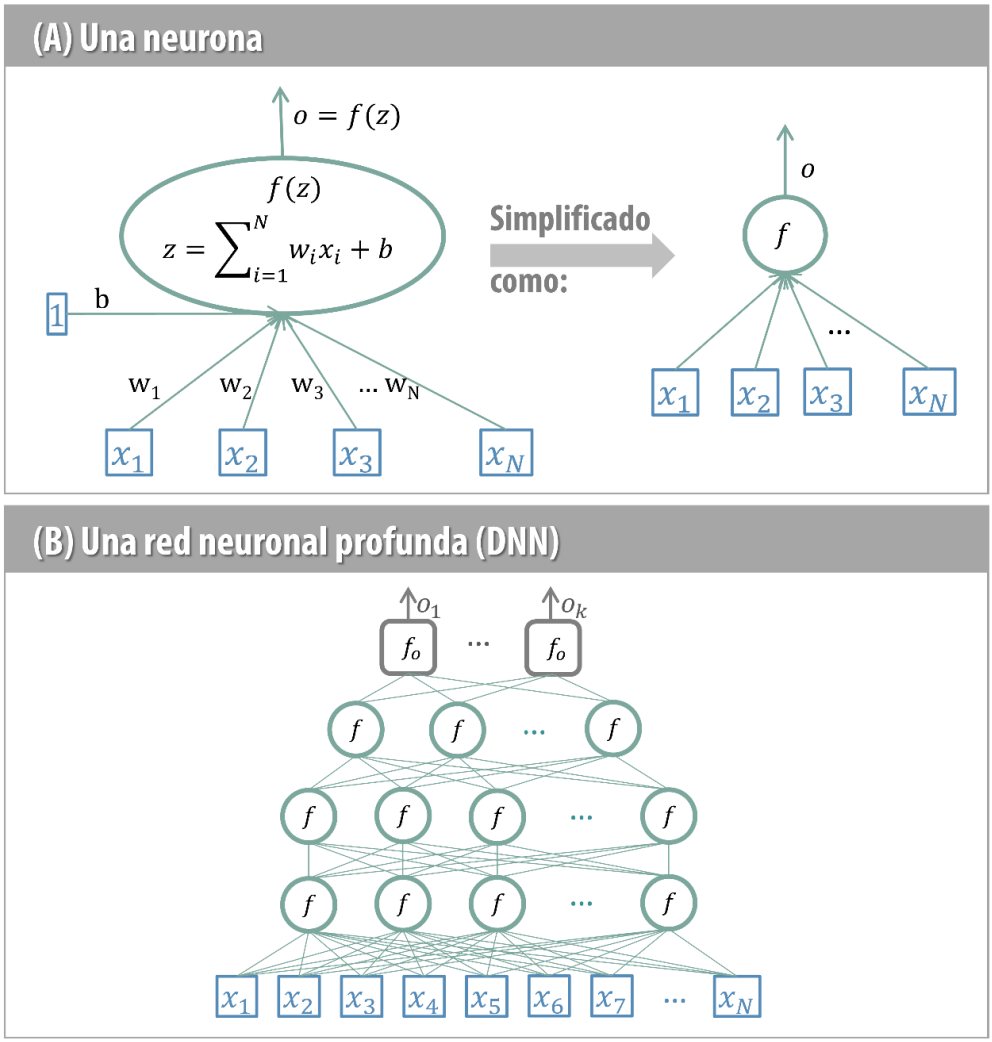


Figura 3.8. Arquitectura de DNN.

Las funciones de activación más comúnmente usadas en las capas ocultas son: (1) la función sigmoidea; o (2) la función de unidad lineal rectificada (ReLU). Ambas funciones y sus derivadas se muestran en la Figura 3.9.

La capa de salida, por su parte, puede tener una o más neuronas. De haber más de una, cada neurona de salida genera predicciones para un punto final separado (por ejemplo, el resultado del ensayo). Es decir, una DNN puede modelar múltiples puntos finales al mismo tiempo. La función de activación de las neuronas en la capa de salida suele ser una función lineal.

El diseño de una DNN, incluido el número de capas y el número de neuronas en cada capa, debe especificarse previamente, junto con la elección de la función de activación en cada neurona. El entrenamiento de una DNN implica maximizar una función objetivo (ϕ) mediante la optimización de los pesos y el sesgo de cada neurona:

$$\emptyset = (\{w_{i,j}\}; \{b_j\}; \quad i = 1, \dots, N_j; \quad j = 1, \dots, L + 1) \tag{3.19}$$

donde N_j es el número de neuronas en la capa j, y L es el número de capas ocultas. El nivel adicional de j ($L+1$) corresponde a la capa de salida.

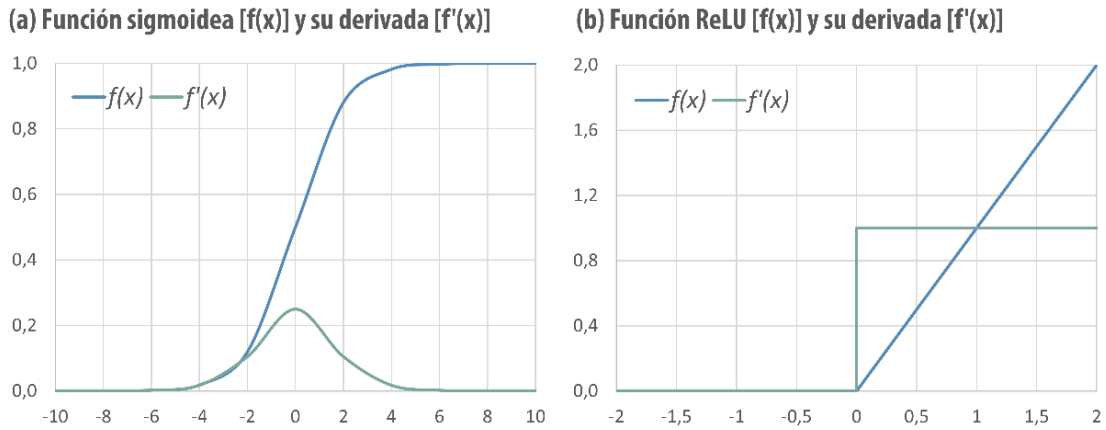


Figura 3.9. Funciones de activación (y sus derivadas) más comúnmente utilizadas en las capas ocultas. (a) Función sigmoidea; (b) Función de unidad lineal rectificada (ReLU).

El procedimiento de entrenamiento suele ser el algoritmo de retropropagación (*Back Propagation*, BP) implementado mediante un descenso de gradiente

estocástico (*Stochastic Gradient Descent*, SGD) con mini-lote (Rumelhart et al., 1986). Los valores individuales para Ø se asignan primero de manera aleatoria. Las moléculas en el conjunto de entrenamiento se mezclan aleatoriamente y luego se dividen de manera uniforme en pequeños grupos llamados "mini-lotes". Cada mini-lote se usa para actualizar los valores de Ø una vez que se utiliza el algoritmo BP. Cuando se utilizan todos los mini-lotes del conjunto de entrenamiento, se dice que el procedimiento de entrenamiento finaliza un período o "época" (*epoch*). En general, el entrenamiento de una DNN requiere muchas épocas, y en cada una se reutiliza el conjunto de entrenamiento. El número de épocas es un parámetro ajustable del método.

En nuestro caso, implementamos el algoritmo DNN utilizando Keras (Chollet et al., 2015), una biblioteca de Python, y siguiendo las recomendaciones de Ma *et al*. (Ma et al., 2015). La biblioteca Scikit-learn (Pedregosa et al., 2011) se utilizó para implementar la validación cruzada de 10 iteraciones, 3 veces, para optimizar el número de épocas durante el entrenamiento del modelo y evitar el sobreajuste. El espacio de exploración para el número de épocas fue desde 1 hasta 201, con incrementos de a 20.

3.2.8. Método interno (*in-house*) de modelado -aprendizaje por ensamblado- basado en subespacios aleatorios

Este método comenzó generando 1000 subconjuntos (subespacios aleatorios), cada uno de ellos compuesto por 200 variables independientes (descriptores) seleccionadas de forma aleatoria a partir de todos los disponibles. Esta estrategia reduce la probabilidad de encontrar correlaciones al azar (que aumenta con el número de posibles variables independientes en el grupo de descriptores) y permite la exploración estocástica del espacio de descriptores moleculares (Alberca et al., 2016; Ho, 1998).

Luego, con cada subconjunto se entrenó un modelo lineal (R Core Team, 2017), por lo que se obtuvieron 1000 modelos individuales. Para el desarrollo de los mismos, se usó una etiqueta de valor 1 para compuestos con $K_{p,uu} > 0.4$ y una etiqueta de valor 0 para compuestos con $K_{p,uu} < 0.4$. Para excluir pares de descriptores altamente correlacionados, se estableció un valor máximo del factor de inflación de la varianza

(VIF) de 2, mientras que se utilizó una relación de 10:1 entre el número de instancias de entrenamiento y el número de variables independientes permitidas en el modelo, para evitar el sobreajuste de los modelos.

Posteriormente, los mejores modelos fueron combinados (hasta 5) por los operadores mínimo y promedio (métodos de ensamblado, *ensemble learning*) para mejorar aún más su desempeño (Polikar, 2012). Para determinar qué modelos eran los mejores, se utilizó el valor del área bajo la curva ROC (por el inglés, *Receiver Operating Characteristic*) en el conjunto de entrenamiento (cuanto mayor, mejor).

3.3. Evaluación del poder explicativo de los modelos

Se utilizaron diferentes medidas para evaluar el rendimiento de los modelos clasificatorios QSPR desarrollados. Estas medidas fueron: la tasa de buenas clasificaciones o exactitud (Acc, por sus siglas en inglés); la sensibilidad (Se, la tasa de verdaderos positivos); la especificidad (Sp, la tasa de verdaderos negativos) y; el coeficiente de correlación de Matthews (MCC, por sus siglas en inglés). Estas medidas fueron calculadas empleando el paquete de R ROCR (Sing et al., 2005).

Las ecuaciones 3.21 – 3.24 muestran las expresiones de cálculo utilizadas, donde VP, FP, VN y FN significan verdaderos positivos, falsos positivos, verdaderos negativos y falsos negativos, en ese orden.

$$Se = \frac{VP}{VP+FN} \tag{3.21}$$

$$Sp = \frac{VN}{FP+VN} \tag{3.22}$$

$$Acc = \frac{VP+VN}{VP+FP+VN+FN} \tag{3.23}$$

$$MCC = \frac{VP \cdot VN - FN \cdot FP}{\sqrt{(VP+FN)(VP+FP)(VN+FN)(VP+FP)}} \tag{3.24}$$

El MCC se utiliza como una medida de la calidad de las clasificaciones binarias (Matthews, 1975). Generalmente se considera como una medida equilibrada que se puede utilizar incluso si las clases son de tamaños muy diferentes, como en nuestro caso. El MCC varía entre -1 y +1; valores más altos indican mayor acuerdo entre la clase observada y la predicha.

Para la evaluación y comparación del desempeño de los modelos generados se utilizaron también las curvas ROC (Triballeau et al., 2005), las cuales son representaciones gráficas de la Se del modelo frente a (1 – Sp). El área bajo la curva ROC (ABC_ROC) constituye una medida valiosa para evaluar si un modelo se comporta significativamente mejor que una clasificación al azar, o que otro modelo (Triballeau et al., 2005). Un modelo ideal presentará una ABC_ROC igual a 1, equivalente a una clasificación perfecta, mientras que la clasificación al azar se representa por una línea de pendiente 1 y corresponde a una ABC_ROC igual a 0,5.

Luego de la aplicación de un modelo, cada molécula del conjunto de datos que se está evaluando obtendrá un resultado numérico único (*score*). Se debe establecer un valor de corte del *score* a partir del cual se considerará alta BD del fármaco libre en el SNC. Dado que Se y Sp evolucionan de forma opuesta, no es posible optimizar ambos parámetros simultáneamente, y se debe recurrir a una solución de compromiso (Triballeau et al., 2005). Equilibrar la tasa de FP y de FN depende de consideraciones pragmáticas que deben ser juzgadas por el investigador (Hubbard et al., 2003). En nuestro caso, utilizamos como criterio el valor de MCC junto con el balance de Se y Sp para establecer el valor de corte de los modelos desarrollados.

El paquete de R pROC (Robin et al., 2011) se utilizó para el cálculo de curvas ROC, ABC_ROC y comparación de ABC_ROC utilizando el método de DeLong (DeLong et al., 1988). En la Figura 3.10 se muestra un esquema de la construcción de una curva ROC.

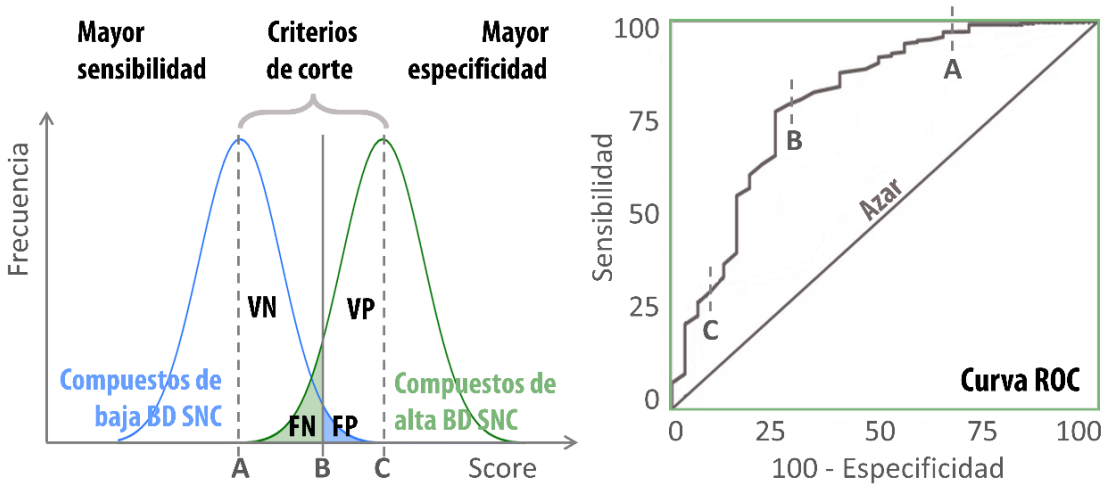


Figura 3.10. Bosquejo del armado de una curva ROC.

3.4. Validación interna o computacional de los modelos

Cada modelo generado fue validado mediante:

› **Validación cruzada de 10 iteraciones**. El conjunto de entrenamiento se dividió aleatoriamente en 10 subconjuntos de igual tamaño, de manera tal que cada compuesto haya sido removido al menos una vez; en cada ronda de validación, uno de estos subconjuntos se reservó y los nueve conjuntos restantes se usaron para inferir el modelo. El esquema anterior se repitió 500 veces para obtener un resultado robusto. En el caso de DNN, se repitió 3 veces debido al costo computacional.

› **Validación de retención**. Un conjunto de prueba independiente (obtenido al dividir a la base de datos inicialmente, cómo se comentó en la sección 3.1.2) sirvió para evaluar la capacidad predictiva de los modelos generados.

3.5. Medición de la importancia de las variables

Con el objetivo de medir la influencia de las variables independientes (descriptores) en cada uno de los modelos desarrollados, se cuantificó la "importancia de las

variables" (IV) mediante diferentes técnicas, dependiendo del algoritmo de aprendizaje automático utilizado:

› GBM. La influencia relativa de cada variable se evaluó utilizando el paquete R gbm (Greenwell et al., 2019) que implementa el método descrito por Friedman *et al*. (Friedman, 2001).

› XGBOOST. La medida de IV, calculada con el paquete R xgboost (Chen et al., 2019), representa la contribución fraccional de cada descriptor al modelo en función de la ganancia total de las divisiones de dicho descriptor.

› RF. Se utilizaron dos medidas de IV: disminución media de la precisión y disminución media de la impureza del nodo, calculada con el paquete R RandomForest (Liaw et al., 2002).

› cPLS. La medida de IV utilizada para este algoritmo se basa en las sumas ponderadas de los coeficientes de regresión absoluta. Los pesos son una función de la reducción de las sumas de cuadrados a través del número de componentes PLS y se calculan por separado para cada resultado. Por lo tanto, la contribución de los coeficientes se ponderan proporcionalmente a la reducción en las sumas de cuadrados (Kuhn, 2020).

› Para el resto de los algoritmos, se realizó un análisis de curva ROC en cada predictor utilizando el paquete R caret (Kuhn, 2020). La regla trapezoidal se utilizó para calcular las ABC_ROC, y esta área se tomó como la medida de IV.

Luego de la cuantificación de la IV, se realizó la comparación entre las variables seleccionadas como más relevantes por los modelos desarrollados por el set de datos MSH y MS, como también entre los conjuntos de datos MS y MS refinado.

3.6. Cálculo del dominio de aplicación de los modelos desarrollados

Para evitar una extrapolación excesiva, se utilizaron medidas de similitud para definir el dominio de aplicabilidad de los modelos, sobre la base de la distancia euclídea promedio entre todos los compuestos de entrenamiento y cada compuesto de prueba (S. Zhang et al., 2006). La distancia de un compuesto de prueba a su vecino

más cercano en el conjunto de entrenamiento se compara con un umbral de dominio de aplicabilidad (UDA) predefinido. Si la distancia supera este umbral, la predicción se considera poco fiable. Dicho UDA se calcula de acuerdo con la expresión:

$$UDA = \langle d \rangle + \mathrm{Z} \cdot \sigma_{\langle d \rangle} \tag{3.25}$$

El cálculo de $\langle d \rangle$ y de σ se realiza de la siguiente manera:

(1) Se obtiene el promedio de las distancias euclidianas entre todos los puntos del conjunto de entrenamiento.

(2) Luego, utilizando aquellas distancias inferiores a la media, se calcula una nueva distancia media $\langle d \rangle$ y su correspondiente desviación estándar, $\sigma_{\langle d \rangle}$.

(3) Por otra parte, Z es un valor de corte empírico, que en nuestro caso se fijó en 0,5.

Referencias

Al-Majdoub, Z. M., Al Feteisi, H., Achour, B., Warwood, S., Neuhoff, S., Rostami-Hodjegan, A., & Barber, J. (2019). Proteomic Quantification of Human Blood-Brain Barrier SLC and ABC Transporters in Healthy Individuals and Dementia Patients. *Molecular Pharmaceutics*, *16*(3), 1220–1233. https://doi.org/10.1021/acs.molpharmaceut.8b01189

Alberca, L. N., Sbaraglini, M. L., Balcazar, D., Fraccaroli, L., Carrillo, C., Medeiros, A.… Talevi, A. (2016). Discovery of novel polyamine analogs with anti-protozoal activity by computer guided drug repositioning. *Journal of Computer-Aided Molecular Design*, *30*(4), 305–321. https://doi.org/10.1007/s10822-016-9903-6

Alberca, L. N., Sbaraglini, M. L., Morales, J. F., Dietrich, R., Ruiz, M. D., Pino Martínez, A. M., … Talevi, A. (2018). Cascade Ligand- and Structure-Based Virtual Screening to Identify New Trypanocidal Compounds Inhibiting Putrescine Uptake. *Frontiers in Cellular and Infection Microbiology*, *8*, 173. https://doi.org/10.3389/fcimb.2018.00173

Barker, M., & Rayens, W. (2003). Partial least squares for discrimination. *Journal of Chemometrics*, *17*(3), 166–173. https://doi.org/10.1002/cem.785

Basu, S., Das, N., Sarkar, R., Kundu, M., Nasipuri, M., & Kumar Basu, D. (2010). A novel framework for automatic sorting of postal documents with multi-script address blocks. *Pattern Recognition*, *43*(10), 3507–3521. https://doi.org/10.1016/J.PATCOG.2010.05.018

Bennett, K. P., & Campbell, C. (2000). Support vector machines. *ACM SIGKDD Explorations Newsletter*, *2*(2), 1–13. https://doi.org/10.1145/380995.380999

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

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). Classification and regression trees, 1st Ed. Routledge, New York. https://doi.org/10.1201/9781315139470

Caballero, J., & Fernández, M. (2006). Linear and nonlinear modeling of antifungal activity of some heterocyclic ring derivatives using multiple linear regression and Bayesian-regularized neural networks. *Journal of Molecular Modeling*, *12*(2), 168–181. https://doi.org/10.1007/s00894-005-0014-x

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16*, 785–794. https://doi.org/10.1145/2939672.2939785

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H. et al. (2019). *xgboost: Extreme Gradient Boosting. R package version 0.90.0.2.*

Chollet, F., & others. (2015). *Keras*. Retrieved from https://keras.io

Conforti, D., & Guido, R. (2010). Kernel based support vector machine via semidefinite programming: Application to medical diagnosis. *Computers & Operations Research*, *37*(8), 1389–1394. https://doi.org/10.1016/J.COR.2009.02.018

Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, *13*(1), 21–27. https://doi.org/10.1109/TIT.1967.1053964

Dauchy, S., Dutheil, F., Weaver, R. J., Chassoux, F., Daumas-Duport, C., Couraud, P.-O. et al. (2008). ABC transporters, cytochromes P450 and their main transcription factors: expression at the human blood-brain barrier. *Journal of Neurochemistry*, *107*(6), 1518–1528. https://doi.org/10.1111/j.1471-4159.2008.05720.x

DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, *44*(3), 837–845. https://doi.org/10.2307/2531595

Douglas M. H., Subhash C. B., & Mills, D. (2003). Assessing Model Fit by Cross-Validation. *Journal of Chemical Information and Computer Sciences*, *43*(2), 579–586. https://doi.org/10.1021/CI025626I

Everitt, B. S., Landau, S., Leese, M., Stahl, D. (2011) Optimization Clustering Techniques. In: Shewhart, W. A. & Wilks, S. S. (Eds) *Cluster Analysis*, 5th Ed. Wiley, Hoboken, NJ. https://doi.org/10.1002/9780470977811.ch5

Fernandez, M., Ahmad, S., & Sarai, A. (2010). Proteochemometric Recognition of Stable Kinase Inhibition Complexes Using Topological Autocorrelation and Support Vector Machines. *Journal of Chemical Information and Modeling*, *50*(6), 1179–1188. https://doi.org/10.1021/ci1000532

Friedman, J. H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics*, *29*(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

Friedman, J. H. (2002). Stochastic gradient boosting. *Computational Statistics & Data Analysis*, *38*(4), 367–378. https://doi.org/10.1016/S0167-9473(01)00065-2

Golbraikh, A., Shen, M., Xiao, Z., Xiao, Y.-D., Lee, K.-H., & Tropsha, A. (2003). Rational selection of training and test sets for the development of validated QSAR models. *Journal of Computer-Aided Molecular Design*, *17*(2/4), 241–253. https://doi.org/10.1023/A:1025386326946

Gramatica, P. (2013). On the development and validation of QSAR models. In: Reisfeld, B. & Mayeno, A. (Eds.) *Computational Toxicology. Methods in molecular biology, v*ol. 930, Humana Press, Totowa, NJ. https://doi.org/10.1007/978-1-62703-059-5_21

Greenwell, B., Boehmke, B., & Cunningham, J. (2019). gbm: Generalized Boosted Regression Models. Retrieved March 7, 2017, from https://cran.r-project.org/package=gbm

Guha, R. (2016). *fingerprint: Functions for Processing Binary Fingerprint Data*. Retrieved from https://cran.r-project.org/package=fingerprint

Hariharan, R., Janakiraman, A., Nilakantan, R., Singh, B., Varghese, S., Landrum, G., & Schuffenhauer, A. (2011). MultiMCS: a fast algorithm for the maximum common substructure problem on multiple molecules. *Journal of Chemical Information and Modeling*, *51*(4), 788–806.

https://doi.org/10.1021/ci100297y

Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A K-Means Clustering Algorithm. *Applied Statistics*, *28*(1), 100-108. https://doi.org/10.2307/2346830

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning - Data Mining, Inference, and Prediction*, 2nd Ed. Springer, NY.

Ho, T. K. (1998). The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *20*(8), 832–844. https://doi.org/10.1109/34.709601

Hubbard, R., & Bayarri, M. J. (2003). Confusion Over Measures of Evidence (*p*'s) Versus Errors (α's) in Classical Statistical Testing. *The American Statistician*, *57*(3), 171–178. https://doi.org/10.1198/0003130031856

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning with applications in R,* 1st Ed. Springer, NY.

Karatzoglou, A., Smola, A., Hornik, K., & Zeileis, A. (2004). kernlab - An S4 Package for Kernel Methods in R. *Journal of Statistical Software*, *11*(9), 1–20. https://doi.org/10.18637/jss.v011.i09

Kassambara, A., & Mundt, F. (2017). factoextra: Extract and Visualize the Results of Multivariate Data Analyses. R package version 1.0.4. Retrieved from https://cran.r-project.org/web/packages/factoextra/index.html

Kawabata, T. (2011). Build-Up Algorithm for Atomic Correspondence between Chemical Structures. *Journal of Chemical Information and Modeling*, *51*(8), 1775–1787. https://doi.org/10.1021/ci2001023

Khorrami, H., & Moavenian, M. (2010). A comparative study of DWT, CWT and DCT transformations in ECG arrhythmias classification. *Expert Systems with Applications*, *37*(8), 5751–5757. https://doi.org/10.1016/J.ESWA.2010.02.033

Kim, H. S., & Sohn, S. Y. (2010). Support vector machines for default prediction of SMEs based on technology credit. *European Journal of Operational Research*,

201(3), 838–846. https://doi.org/10.1016/J.EJOR.2009.03.036

Kuhn, M. (2020). *caret: Classification and Regression Training*. Retrieved from https://cran.r-project.org/package=caret

Law, V., Knox, C., Djoumbou, Y., Jewison, T., Guo, A. C., Liu, Y. et al. (2014). DrugBank 4.0: shedding new light on drug metabolism. *Nucleic Acids Research*, *42*(Database issue), D1091-7. https://doi.org/10.1093/nar/gkt1068

Leonard, J. T., & Roy, K. (2006). On Selection of Training and Test Sets for the Development of Predictive QSAR models. *QSAR & Combinatorial Science*, *25*(3), 235–251. https://doi.org/10.1002/qsar.200510161

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R News*. Retrieved from http://ai2-s2-pdfs.s3.amazonaws.com/6e63/3b41d93051375ef9135102d54fa097dc8cf8.pdf

Liew, C. Y., Ma, X. H., Liu, X., & Yap, C. W. (2009). SVM Model for Virtual Screening of Lck Inhibitors. *Journal of Chemical Information and Modeling*, *49*(4), 877–885. https://doi.org/10.1021/ci800387z

Liew, C. Y., Ma, X. H., & Yap, C. W. (2010). Consensus model for identification of novel PI3K inhibitors in large chemical library. *Journal of Computer-Aided Molecular Design*, *24*(2), 131–141. https://doi.org/10.1007/s10822-010-9321-0

Ma, J., Sheridan, R. P., Liaw, A., Dahl, G. E., & Svetnik, V. (2015). Deep Neural Nets as a Method for Quantitative Structure–Activity Relationships. *Journal of Chemical Information and Modeling*, *55*(2), 263–274. https://doi.org/10.1021/ci500747n

Martin, T. M., Harten, P., Young, D. M., Muratov, E. N., Golbraikh, A., Zhu, H., & Tropsha, A. (2012). Does Rational Selection of Training and Test Sets Improve the Outcome of QSAR Modeling? *Journal of Chemical Information and Modeling*, *52*(10), 2570–2578. https://doi.org/10.1021/ci300338w

Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta (BBA) - Protein*

Structure, *405*(2), 442–451. https://doi.org/10.1016/0005-2795(75)90109-9

Mauser, G. A., & Hartigan, J. A. (1977). Clustering Algorithms. *Journal of Marketing Research*, *14*(1), 124-125. https://doi.org/10.2307/3151073

Mercader, A. G., Duchowicz, P. R., Fernández, F. M., & Castro, E. A. (2010). Replacement Method and Enhanced Replacement Method Versus the Genetic Algorithm Approach for the Selection of Molecular Descriptors in QSPR/QSAR Theories. *Journal of Chemical Information and Modeling*, *50*(9), 1542–1548. https://doi.org/10.1021/ci100103r

Mevik, B.-H., Wehrens, R., & Liland, K. H. (2016). pls: Partial Least Squares and Principal Component Regression. Retrieved from https://cran.r-project.org/package=pls

Morales, J. F., Montoto, S. S., Fagiolino, P., & Ruiz, M. E. (2017). Current State and Future Perspectives in QSAR Models to Predict Blood-Brain Barrier Penetration in Central Nervous System Drug R&D. *Mini Reviews in Medicinal Chemistry*, *17*(3), 247–257. https://doi.org/10.2174/1389557516666161013110813

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O. et al. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*(Oct), 2825–2830.

Polikar, R. (2012). Ensemble Learning. In: Zhang, C. & Ma, Y. (Eds.), *Ensemble Machine Learning.* Springer, Boston, MA. https://doi.org/10.1007/978-1-4419-9326-7_1

R Core Team. (2017). *R: A language and environment for statistical computing*. Retrieved from http://www.r-project.org/

Rankovic, Z. (2017). CNS Physicochemical Property Space Shaped by a Diverse Set of Molecules with Experimentally Determined Exposure in the Mouse Brain. *Journal of Medicinal Chemistry*, *60*(14), 5943–5954. https://doi.org/10.1021/acs.jmedchem.6b01469

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J. C., & Muller, M. (2011). pROC: an open-source package for R and S+ to analyze and compare

ROC curves. *BMC Bioinformatics*, *12*, 77. https://doi.org/10.1186/1471-2105-12-77

Roy, K., & Mitra, I. (2011). On Various Metrics Used for Validation of Predictive QSAR Models with Applications in Virtual Screening and Focused Library Design. *Combinatorial Chemistry & High Throughput Screening*, *14*(6), 450–474. https://doi.org/10.2174/138620711795767893

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*(6088), 533–536. https://doi.org/10.1038/323533a0

Schwenk, H., & Bengio, Y. (2000). Boosting Neural Networks. *Neural Computation*, *12*(8), 1869–1887. https://doi.org/10.1162/089976600300015178

Shen, J., Cheng, F., Xu, Y., Li, W., & Tang, Y. (2010). Estimation of ADME Properties with Substructure Pattern Recognition. *Journal of Chemical Information and Modeling*, *50*(6), 1034–1041. https://doi.org/10.1021/ci100104j

Sheridan, R. P., Wang, W. M., Liaw, A., Ma, J., & Gifford, E. M. (2016). Extreme Gradient Boosting as a Method for Quantitative Structure– Activity Relationships. *Journal of Chemical Information and Modeling, 56*(12), 2353-2360. https://doi.org/10.1021/acs.jcim.6b00591

Sing, T., Sander, O., Beerenwinkel, N., & Lengauer, T. (2005). ROCR: visualizing classifier performance in R. *Bioinformatics*, *21*(20), 3940–3941. https://doi.org/10.1093/bioinformatics/bti623

Svetnik, V., Wang, T., Tong, C., Liaw, A., Sheridan, R. P., & Song, Q. (2005). Boosting: An Ensemble Learning Tool for Compound Classification and QSAR Modeling. *Journal of Chemical Information and Modeling*, *45*(3), 786–799. https://doi.org/10.1021/ci0500379

Talevi, A., A. Castro, E., & E. Bruno-Blanch, L. (2012). Recent Studies on Similarity Measures and its Applications to Chemoinformatics and Drug Design. In: Khan, M. T. (Ed.) *Recent Trends on QSAR in the Pharmaceutical Perceptions* Bentham Science, Al Sharjah, United Arab Emirates. https://doi.org/10.2174/978160805379711201010272

Talevi, A., L. Bellera, C., Di Ianni, M., R. Duchowicz, P., E. Bruno-Blanch, L., & A. Castro, E. (2012). An Integrated Drug Development Approach Applying Topological Descriptors. *Current Computer Aided-Drug Design*, *8*(3), 172–181. https://doi.org/10.2174/157340912801619076

Tan, P.-N., Steinbach, M., & Kumar, V. (2005). *Introduction to data mining*. Retrieved from https://dl.acm.org/citation.cfm?id=1095618

Triballeau, N., Acher, F., Brabet, I., Pin, J. P., & Bertrand, H. O. (2005). Virtual screening workflow development guided by the "receiver operating characteristic" curve approach. Application to high-throughput docking on metabotropic glutamate receptor subtype 4. *J Med Chem*, *48*(7), 2534–2547. https://doi.org/10.1021/jm049092j

Tropsha, A. (2010). Best Practices for QSAR Model Development, Validation, and Exploitation. *Molecular Informatics*, *29*(6–7), 476–488. https://doi.org/10.1002/minf.201000061

Uchida, Y., Ohtsuki, S., Katsukura, Y., Ikeda, C., Suzuki, T., Kamiie, J., & Terasaki, T. (2011). Quantitative targeted absolute proteomics of human blood-brain barrier transporters and receptors. *Journal of Neurochemistry*, *117*(2), 333–345. https://doi.org/10.1111/j.1471-4159.2011.07208.x

Wager, T. T., Chandrasekaran, R. Y., Hou, X., Troutman, M. D., Verhoest, P. R., Villalobos, A., & Will, Y. (2010). Defining Desirable Central Nervous System Drug Space through the Alignment of Molecular Properties, in Vitro ADME, and Safety Attributes. *ACS Chemical Neuroscience*, *1*(6), 420–434. https://doi.org/10.1021/cn100007x

Wang, H., Liu, C., & Deng, L. (2018). Enhanced Prediction of Hot Spots at Protein-Protein Interfaces Using Extreme Gradient Boosting. *Scientific Reports*, *8*, 14285. https://doi.org/10.1038/s41598-018-32511-1

Warnes, G. R., Bolker, B., Bonebakker, L., Gentleman, R., Liaw, W. H. A., Lumley, T. et al. (2016). *gplots: Various R Programming Tools for Plotting Data*. Retrieved from https://cran.r-project.org/package=gplots

Wickham, H. (2009). *Ggplot2: elegant graphics for data analysis*. Springer, Boston,

MA.

Xue, Y., Li, H., Ung, C. Y., Yap, C. W., & Chen, Y. Z. (2006). Classification of a Diverse Set of Tetrahymena pyriformis Toxicity Chemical Compounds from Molecular Descriptors by Statistical Learning Methods. *Chemical Research in Toxicology*, *19*(8), 1030–1039. https://doi.org/10.1021/tx0600550

Yee, L. C., & Wei, Y. C. (2012). Current Modeling Methods Used in QSAR/QSPR. In: Dehmer, M., Varmuza, K. & Bonchev, D. (Eds.) *Statistical Modelling of Molecular Descriptors in QSAR/QSPR*, vol. 2. Wiley, Hoboken, NJ. https://doi.org/10.1002/9783527645121.ch1

Zhang, S., Golbraikh, A., Oloff, S., Kohn, H., & Tropsha, A. (2006). A Novel Automated Lazy Learning QSAR (ALL-QSAR) Approach: Method Development, Applications, and Virtual Screening of Chemical Databases Using Validated ALL-QSAR Models. *Journal of Chemical Information and Modeling*, *46*(5), 1984–1995. https://doi.org/10.1021/ci060132x

Zhang, Y. Y., Liu, H., Summerfield, S. G., Luscombe, C. N., & Sahi, J. (2016). Integrating in Silico and in Vitro Approaches to Predict Drug Accessibility to the Central Nervous System. *Molecular Pharmaceutics*, *13*(5), 1540–1550. https://doi.org/10.1021/acs.molpharmaceut.6b00031

Zuluaga, M. A., Magnin, I. E., Hernández Hoyos, M., Delgado Leyton, E. J. F., Lozano, F., & Orkisz, M. (2011). Automatic detection of abnormal vascular cross-sections based on density level detection and support vector machines. *International Journal of Computer Assisted Radiology and Surgery*, *6*(2), 163–174. https://doi.org/10.1007/s11548-010-0494-8

Capítulo 4

Validación de modelos

4.1. Validación externa o experimental de los modelos desarrollados con el set de datos MSH

El mejor modelo fue validado externamente de forma prospectiva utilizando la siguiente metodología: a compuestos cuyo valor de K_p en estado estacionario se encontraba disponible en bibliografía, pero no su valor de $K_{p,uu}$ (y que, por lo tanto, no formaban parte del conjunto de datos MSH), se les determinaron los valores tanto de fracción libre en plasma como de fracción libre en cerebro, para poder obtener el correspondiente valor de $K_{p,uu}$ y así compararlo con la predicción de nuestro modelo computacional desarrollado.

Con dicho propósito, se realizó una búsqueda bibliográfica para encontrar compuestos que verificaran la condición anterior, es decir, valor de K_p obtenido en estado estacionario, en ratas, y valor de $K_{p,uu}$ no reportado. Adicionalmente, se constató que ambas categorías estuvieran representadas, es decir, que quedaran incluidos compuestos tanto de alta como de baja biodisponibilidad del fármaco libre

en el SNC. Finalmente, se sumó el requisito adicional de que los mismos fueran accesibles comercialmente, seleccionando finalmente los compuestos que se listan en la Tabla 4.1 para la validación experimental de los modelos.

Tabla 4.1. Compuestos seleccionados para la validación experimental de los modelos basados en el conjunto de datos MSH. Los datos de *Kp* reportados fueron determinados en estado estacionario y en rata como modelo animal.

	Valor de K_p reportado	**Categoría (BD en SNC)**	**Referencia**
Ácido para-aminobenzoico	0,04	Baja	(Nakazono et al., 1991)
Clorfeniramina	34	Alta	(Doan et al., 2004)
Lidocaína	2,2	Alta	(Nakazono et al., 1991)
Ranitidina	0,058	Baja	(Young et al., 1988)
Teofilina	0,7	Alta	(Yasuhara et al., 1988)

4.1.1. Determinación de la fracción libre en plasma

4.1.1.1. Ultrafiltración

La técnica de ultrafiltración se basa en el uso de la fuerza centrífuga para impulsar al compuesto de prueba a través de una membrana de tamaño selectivo, que dejará pasar al compuesto libre, no unido a proteínas plasmática, el cual es determinado en el filtrado mediante un método analítico adecuado (Dow, 2006). Para obtener resultados válidos mediante esta técnica, el grado de unión inespecífica entre el compuesto a testear y el dispositivo experimental (tubo plástico + membrana) debe determinarse previamente, ya que si la misma supera el 5% se considera que los resultados no poseerán exactitud aceptable (Dow, 2006; Toma et al., 2021).

En ese caso, no se recomienda la ultrafiltración para la determinación de la unión a proteínas plasmáticas y debe investigarse una técnica alternativa como la diálisis de equilibrio, descripta en la sección 4.1.1.2, o la ultracentrifugación (Morales et al., 2017). No es aconsejable la "corrección" de los datos mediante sustracción o resta del valor de unión inespecífica ya que ésta se determina empleando una solución

blanco del fármaco, sin plasma, mientras que en el experimento real la presencia de plasma puede alterar la extensión de la unión inespecífica (y, por lo tanto, no corresponderse con la determinada en el blanco).

Protocolo experimental

Materiales:

- Solución buffer fosfato isotónico -PBS- pH = 7,4.
- Compuesto de prueba.
- Solvente adecuado (por ej. metanol, acetonitrilo, DMSO).
- Plasma de rata previamente colectado y congelado.
- Filtros para centrífuga Microcon® de 10 kDa de tamaño de poro (Merck KGaA, Darmstadt, Alemania).
- Equipo de centrifugación.

Preparación del buffer PBS: para preparar 100 mL de buffer PBS pH = 7,4 concentrado (10X) se pesaron 8 g de NaCl, 0,2 g de KCl, 1,44 g de Na_2HPO_4 y 0,24 g de KH_2PO_4, los cuales se disolvieron en agua destilada y se ajustó el pH a 7,4 (con soluciones de HCl o NaOH, según correspondiera). El mismo día del experimento se realizaba una dilución 1:10 de la solución buffer concentrada 10X en agua destilada.

Preparación de la solución madre o stock del compuesto de prueba: la misma fue preparada en un solvente adecuado a una concentración 100 veces mayor que la requerida en el experimento, de tal manera que luego de la dilución para obtener la concentración final del compuesto prueba, la concentración de solvente orgánico fuera tan sólo 1% v/v. Dado que la concentración del compuesto de prueba utilizada durante el experimento de ultrafiltración fue de 10 µM (Dow, 2006), las concentraciones de las soluciones stock fueron del orden de 1000 µM. Estas soluciones se conservaron en el freezer.

Determinación de la unión inespecífica: se realizó una dilución 1:100 de la solución madre del compuesto de prueba con buffer PBS pH = 7,4 para obtener una concentración 10 µM del compuesto prueba. Se agito la dilución y se dejó durante 20 minutos en un baño de agua a 37 °C, con el objetivo de realizar el experimento a

la temperatura fisiológica. A continuación, se transfirieron 300 µL de la muestra a la cámara dadora del tubo de filtración (Microcon®) y se centrifugó durante 20 minutos a 1000g. Al finalizar, se tomó una muestra del ultrafiltrado que se encontraba en el compartimiento aceptor del Microcon® para determinar la concentración del compuesto de prueba mediante cromatografía líquida de alta resolución (HPLC), siguiendo métodos previamente desarrollados para cada analito (ver descripción de los métodos en la sección 4.3).

El porcentaje de unión inespecífica (%UI) se determinó de acuerdo a la siguiente ecuación:

$$\%UI = 100 - \left[\frac{concentracion\ del\ analito\ en\ el\ ultrafiltrado}{concentracion\ inicial\ de\ analito\ en\ PBS} * 100\right] \quad (4.1)$$

Determinación de la unión a proteínas plasmáticas: para ello, el día del experimento se dejaba descongelar el plasma de rata a temperatura ambiente, luego de lo cual se centrifugaba durante 5 minutos a 2000g para eliminar material en suspensión. A continuación, se procedió de la misma manera que se describió previamente para la determinación de la unión inespecífica, pero realizando la dilución 1:100 de la solución madre del compuesto de prueba con el plasma de rata en lugar del buffer fosfato PBS pH = 7,4.

El porcentaje de unión a proteínas plasmáticas (%UP) fue determinado según:

$$\%\ UP = 100 - \left[\frac{concentracion\ del\ analito\ en\ el\ ultrafiltrado}{concentracion\ inicial\ de\ analito\ en\ plasma} * 100\right] \quad (4.2)$$

4.1.1.2. Diálisis de equilibrio

La fracción libre en plasma de aquellos compuestos con un %UI mayor al 5% se determinó mediante la técnica de diálisis de equilibrio, con ciertas modificaciones. Primero se describirán los fundamentos de la técnica y luego las modificaciones realizadas.

La diálisis de equilibrio se ha considerado durante mucho tiempo el método estándar de excelencia para determinar la fracción de fármaco no unido a las proteínas plasmáticas (f_u) (Chen et al., 2019; Vuignier et al., 2013). En un

experimento de diálisis de equilibrio, la celda de diálisis consta de dos cámaras, denominadas D y R (de *donor* y *receptor*, ver a la izquierda de la Fig. 4.1) separadas por una membrana de diálisis. La membrana de diálisis es semipermeable, ya que contiene poros que son lo suficientemente grandes como para permitir que pequeñas moléculas (como la droga en solución, y el solvente) se muevan libremente de un lado a otro, pero demasiado pequeños para permitir el paso de moléculas más grandes como las proteínas del plasma. El plasma enriquecido con el fármaco se agrega a un lado de la membrana de diálisis (D) y una solución buffer isotónica del otro (R). Se permite que el sistema alcance el equilibrio, y en ese momento la fracción del fármaco unida a las proteínas del plasma permanece del lado D, mientras que la concentración de fármaco libre es equivalente en ambos lados de la membrana de diálisis. La f_u es igual a la concentración de fármaco libre dividida por la concentración total de fármaco, es decir la sumatoria de lo libre más lo unido. La concentración del fármaco en el lado del plasma de la celda de diálisis (D) es equivalente a la suma del fármaco unido más no unido, mientras que la concentración del fármaco en el lado del dializado o receptor de la celda (R) es equivalente a la concentración de fármaco libre. Por lo tanto, la f_u es igual a la concentración del fármaco en el lado del dializado (concentración de fármaco libre) dividida por la concentración del fármaco en el lado del plasma (concentración total de fármaco), es decir, f_u = R / D.

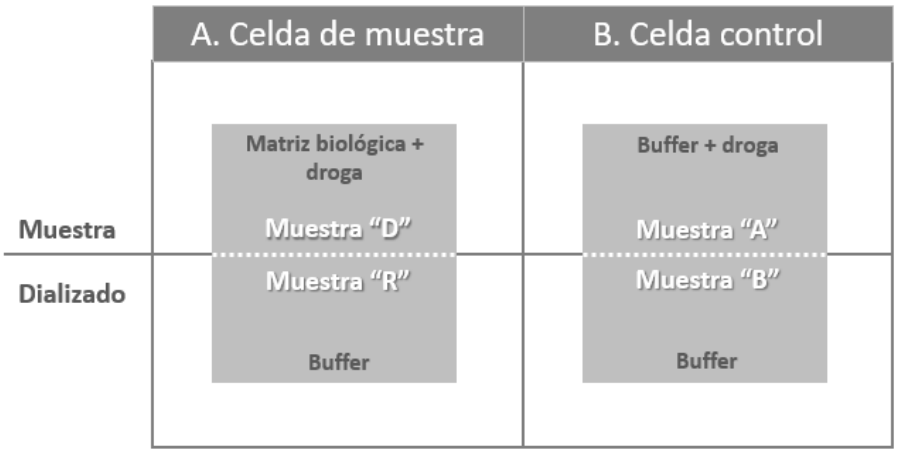


Figura 4.1. Esquema de la celda de diálisis. La sección A es para el experimento con la muestra y la sección B para el control. La línea punteada representa la membrana de diálisis.

Sobre la técnica hasta aquí descripta, se realizaron las modificaciones propuestas por Banker *et al.* (Banker et al., 2008). En dicho trabajo, se presenta una fórmula alternativa para calcular los valores de f_u. Esta nueva fórmula es compatible con la metodología de diálisis de equilibrio existente, ya que se utilizan las mismas muestras biológicas.

En la versión modificada, además de las celdas D y R, se debe incluir un control experimental para demostrar que las muestras han alcanzado el equilibrio. Esto se logra añadiendo el compuesto de prueba a la solución buffer isotónica y dializando contra la misma solución. Estas muestras de control de equilibrio se representan como muestras A y B del lado derecho de la Fig. 4.1. Si el sistema se ha incubado durante un período de tiempo suficiente, la concentración del compuesto de prueba en las muestras A y B debe ser equivalente, es decir, A/B = 1. La derivación de la nueva fórmula se describe en la Tabla 4.2.

Tabla 4.2. Ecuaciones correspondientes al método de diálisis de equilibrio con las modificaciones de Banker *et al*. R, D, A y B representan las concentraciones del analito en las cámaras correspondientes (ver Fig. 4.1), mientras que V_1 y V_2 representan el volumen de los compartimentos de muestra y dializado, respectivamente.

Ecuación original	$f_u = R/D$	(4.3)
Cuando se agrega una cantidad equivalente de fármaco a las celdas de diálisis experimental y control	$V_1D + V_2R = V_1A + V_2B$	(4.4)
Cuando la celda control alcanza el equilibrio	$A = B$	(4.5)
Sustituyendo (4.5) en (4.4)	$V_1D + V_2R = V_1B + V_2B$	(4.6)
Resolviendo la ecuación (4.6) para D	$D = \frac{(V_1 + V_2)B - V_2R}{V_1}$	(4.7)
Sustituyendo (4.7) en (4.3)	$f_u = \frac{R}{\left\{\frac{[(V_1 + V_2)B - V_2R]}{V_1}\right\}}$	(4.8)

El cálculo propuesto por Banker *et al.* permite llegar al valor de f_u utilizando solamente los valores de concentración del compuesto de prueba en las cámaras R y B, es decir, sin plasma (ambas son concentraciones del lado dializado o receptor de la membrana: R del dializado de plasma + fármaco, y B del dializado de buffer + fármaco). El reordenamiento algebraico (que se muestra en la Tabla 4.2) demuestra que la concentración del compuesto de prueba en la muestra D es equivalente a la

expresión (4.7). Resolviendo para D y sustituyendo en la ecuación original (4.3), se llega a la expresión final de cálculo, (4.8). La nueva fórmula supone que, si la recuperación del fármaco disminuye (por ejemplo, por uniones inespecíficas a los elementos del sistema, o porque aún no se ha logrado el equilibrio), A será mayor a B en el experimento control, permitiendo compensar un evento similar en la celda de diálisis donde el compuesto de prueba se añadió al plasma. La concentración del compuesto de prueba en la muestra B representa la cantidad de compuesto que se podría esperar en el lado del dializado si no hubiera unión a las proteínas del plasma, a la vez que también permite compensar por las posibles pérdidas de recuperación antes mencionadas.

Por lo tanto, cuando se usa la ecuación (4.8) sólo se debe determinar la concentración del compuesto de prueba en dos muestras por réplica, para generar valores de f_u que también tienen en cuenta cualquier pérdida de compuesto de prueba debido a la unión inespecífica al aparato de diálisis. Por otro lado, y dado que las muestras R y B son ultrafiltrados sin proteínas plasmáticas, una ventaja adicional del método es que dichas muestras pueden analizarse directamente por HPLC, eliminando la necesidad de pasos previos de preparación de muestra, tales como la precipitación de proteínas. De esta manera, esta modificación del protocolo reduce significativamente los requisitos y tiempos de manejo de muestras, a la vez que mejora la exactitud del resultado.

Protocolo experimental

Materiales:

- Solución buffer fosfato isotónico -PBS- pH = 7,4.
- Compuesto de prueba.
- Solvente adecuado (por ej. metanol, acetonitrilo, DMSO).
- Plasma de rata previamente colectado y congelado.
- Membrana de diálisis de celulosa con punto de corte de peso molecular de 12 kDa (Merck KGaA, Darmstadt, Alemania).
- Equipo de dializado diseñado en nuestro laboratorio, empleando placas de 6 well Costar Snapwell (Corning INC., Corning, NY, USA), a los cuales se les reemplazó

la membrana de policarbonato de fábrica por la membrana de diálisis (área de membrana aproximada 1,12 cm^2 según fabricante).

Preparación del buffer PBS: ídem sección 4.1.1.1

Preparación de la solución madre o stock del compuesto de prueba: ídem sección 4.1.1.1

Diálisis de equilibrio según método modificado por Banker et al.: el día anterior al experimento se prepararon las membranas de diálisis, cortándolas de forma cuidadosa en cuadrados de 3x3 cm^2 aproximadamente, los cuales se dejaron hidratar toda la noche en agua destilada.

El día del experimento, se dejó descongelar el plasma de rata a temperatura ambiente, luego de lo cual se centrifugó durante 5 minutos a 2000*g* para eliminar material en suspensión.

Como se mencionó anteriormente, el equipo empleado para realizar la diálisis fue diseñado en nuestro laboratorio, empleando placas de 6 well Costar Snapwell (Corning INC., Corning, NY, USA), pero reemplazando la membrana de policarbonato de fábrica por la membrana de diálisis (ver Figura 4.2). Este equipo permitió realizar el experimento con plasma y el experimento control en simultáneo.

Para el experimento con plasma, se colocaron en el compartimento dador de la celda de dializado 700 µL de plasma teniendo una concentración de 10 µM del compuesto prueba, y en el compartimento aceptor de la misma celda se colocaron 300 µL de buffer PBS. En el caso de la celda control, se colocaron 700 µL de buffer PBS con una concentración de 10 µM del compuesto prueba, y también se dializó contra 300 µL del mismo buffer. El sistema se dejó equilibrar por al menos 6 horas en un agitador orbital a 37 °C. Pasado este tiempo se tomaron muestras de los compartimentos aceptores tanto del experimento con plasma (R) como del control (B).

Las muestras R y B se centrifugaron durante 10 minutos a 12000*g* y se determinaron las concentraciones mediante análisis por HPLC, siguiendo métodos previamente desarrollados para cada analito o compuesto prueba. Estos métodos se describen en la sección 4.3.

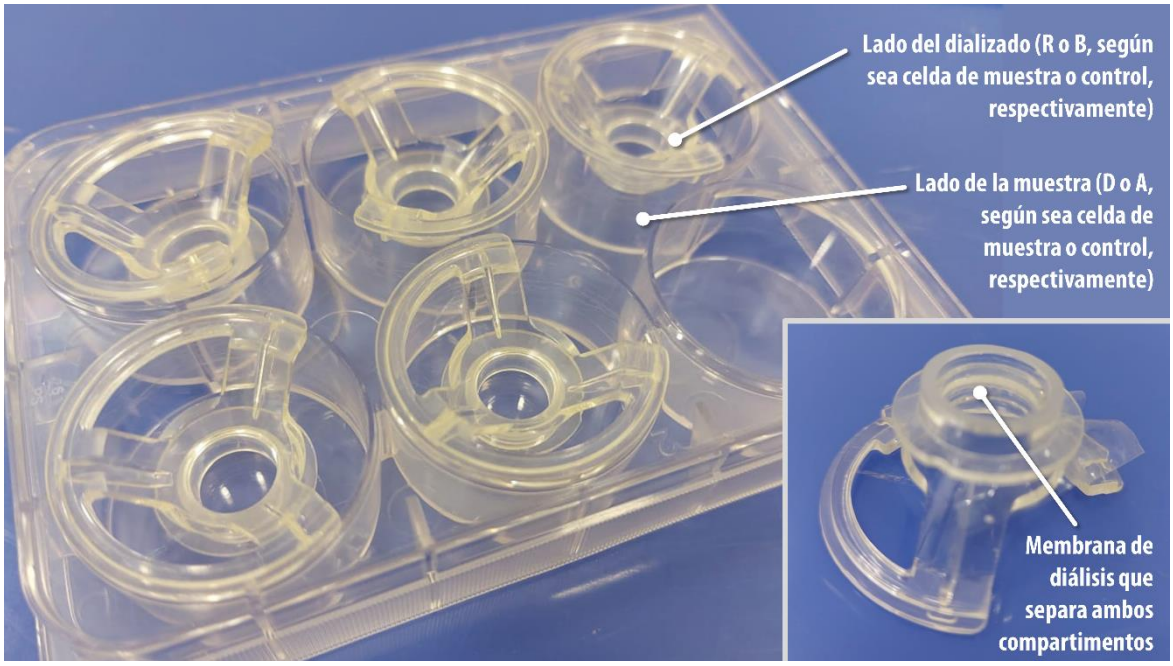


Figura 4.2. Equipo de diálisis utilizado.

Una vez determinadas las concentraciones R y B, se realizó el cálculo de la f_u utilizando la ecuación (4.8), donde V_1 es el volumen del compartimento dador y V_2 es el volumen del compartimento aceptor (volúmenes adicionados, es decir, teóricos iniciales).

4.1.2. Fracción libre en cerebro por el método del homogenato

La fracción libre en cerebro para cada compuesto se determinó mediante la técnica de diálisis de equilibrio con homogenato de cerebro descripta por Kalvass *et al.* (Kalvass et al., 2002), con las mismas modificaciones descriptas anteriormente para plasma. Esta técnica es esencialmente similar a la detallada en la sección anterior, por lo que sólo se remarcarán los aspectos diferenciales.

Materiales:

› Solución buffer fosfato de sodio 100 mM pH=7,4.
› Compuesto de prueba.
› Solvente adecuado (por ej. metanol, acetonitrilo, DMSO).

- Cerebros frescos de rata.
- Membrana de diálisis de celulosa con punto de corte de peso molecular de 12 kDa (Merck KGaA, Darmstadt, Alemania).
- Equipo de dializado diseñado en nuestro laboratorio (Fig. 4.2).

Preparación del buffer fosfato de sodio 100 mM: para preparar de 100 mL de buffer fosfato de sodio 100 mM pH=7,4 se pesaron 0,27 g de NaH_2PO_4 y 1,09 g de Na_2HPO_4, los cuales se disolvieron en agua destilada y se ajustó el pH a 7,4 (con soluciones de HCl o NaOH, según correspondiera).

Preparación de la solución madre o stock del compuesto de prueba: la misma fue preparada en un solvente adecuado a una concentración 100 veces mayor que la requerida en el experimento, de tal manera que luego de la dilución para obtener la concentración final del compuesto prueba, la concentración de solvente orgánico sea de sólo el 1% v/v. Dado que la concentración a utilizar durante el experimento debía ser de 1 µM (Loryan et al., 2013), las concentraciones de las soluciones stock fueron del orden de 100 µM. Estas soluciones se conservan en el freezer.

Determinación de la fracción libre por el método del homogenato: el día del experimento, se preparó el homogenato a partir de cerebros frescos de rata. Los cerebros fueron pesados, diluidos con 2 volúmenes de solución buffer fosfato de sodio 100 mM pH=7,4 y homogeneizados con un homogeneizador de alto corte Pro Scientific Bio-Gen Pro200 (PRO Scientific Inc., Oxford, CT, USA), durante al menos 1 minuto.

Utilizando el equipo de diálisis descripto en la sección anterior (Fig. 4.2), se realizó el experimento con homogenato de cerebro y el experimento control en simultáneo. Para el experimento con homogenato se colocaron en el compartimento dador de la celda de dializado 700 µL de homogenato de cerebro con una concentración de aproximadamente 1 µM del compuesto prueba (por dilución 1:100 de la solución stock), y en el compartimento aceptor de la misma celda se colocaron 300 µL de buffer fosfato de sodio 100 mM pH = 7,4. En el caso de la celda control, se colocaron 700 µL de buffer fosfato de sodio 100 mM pH=7,4 con una concentración de

aproximadamente 1 µM del compuesto prueba, y también se dializó contra 300 µL del mismo buffer.

El sistema se dejó equilibrar durante toda la noche en un agitador orbital a 37 °C y a continuación se tomaron muestras de los compartimentos aceptores tanto del experimento con homogenato de cerebro (R) como del control (B), las cuales se centrifugaron y analizaron mediante los métodos por HPLC descriptos en la sección 4.3.

Si bien el cálculo de la fracción libre se realiza de la misma manera que en la sección 4.1.1 (ecuación (4.8)), dado que el valor de f_u es sensible a la dilución, las fracciones libres determinadas a partir de homogenatos de tejido cerebral diluidos deben ser corregidas por un factor de dilución. Aquí emplearemos el método de corrección de dilución descripto por Kalvass *et al.* (Kalvass et al., 2002). Dicha corrección se deriva considerando el equilibrio de unión del compuesto de interés (C) a los componentes tisulares (T, componentes de unión no específicos del tejido) (Romer et al., 1979), según:

$$C + T \rightleftharpoons CT \tag{4.9}$$

El equilibrio en (4.9) puede representarse mediante una constante de disociación (K_d):

$$K_d = \frac{[C][T]}{[CT]} \tag{4.10}$$

Mientras que la fracción libre en cerebro del compuesto C estará dada por la expresión:

$$f_u = \frac{[C]}{[C]+[CT]} \tag{4.11}$$

Despejando [CT] de (4.10), reemplazado en (4.11) y reordenando, se llega a que:

$$f_u = \frac{K_d}{[T]+K_d} \tag{4.12}$$

Reordenando para despejar K_d:

$$K_d = \frac{f_u[T]}{1-f_u} \tag{4.13}$$

Aplicando la ecuación anterior a dos concentraciones diferentes de componentes de unión inespecíficos, $[T]_1$ y$[T]_2$, con sus respectivas fracciones libres asociadas, f_{u1} y f_{u2}, y haciendo el cociente, se obtiene:

$$1 = \frac{f_{u1}[T]_1(1-f_{u2})}{f_{u2}[T]_2(1-f_{u1})} \tag{4.14}$$

Definiendo la relación de $[T]_1/[T]_2$ como un factor de dilución $D \geq 1$ y simplificando, se obtiene la siguiente ecuación, donde f_{u2} es la fracción no unida medida después de una dilución de tejido cerebral conocida:

$$f_{u1}\ (no\ diluido) = \frac{{}^1/_D}{\left(\left({}^1/_{f_{u2}}\right)-1\right)+{}^1/_D} \tag{4.15}$$

Esta ecuación final fue utilizada para corregir y estimar la f_u en el tejido no diluido. Puede verse que, en ausencia de dilución, D = 1 y la ecuación (4.15) se reduce de manera tal que f_u sin diluir (f_{u1}) será igual a f_{u2}. La ecuación (4.15) muestra un comportamiento no lineal entre la fracción libre y la concentración de componentes de unión no específicos del tejido, representado por el factor de dilución en la ecuación. Este comportamiento no lineal entre ambos ya ha sido descrito previamente en bibliografía (Romer et al., 1979).

4.1.3. Cálculo del parámetro farmacocinético $K_{p,uu}$

Una vez que se completaron los cálculos de la fracciones libres tanto en plasma como en cerebro, se utilizó la siguiente ecuación para el cálculo de los valores de $K_{p,uu}$ de cada compuesto de prueba:

$$K_{p,uu} = \frac{C_{u,cerebro}}{C_{u,plasma}} = K_p \frac{f_{u,cerebro}}{f_{u,plasma}} \tag{4.16}$$

A cada compuesto se lo clasificó de alta o baja biodisponibilidad del fármaco libre en el SNC, dependiendo de si el valor de $K_{p,uu}$ obtenido era menor o mayor que 0,4, respectivamente. Luego se compararon las etiquetas predichas por el modelo computacional y las clases determinadas experimentalmente para los compuestos de prueba, para darle validez a nuestro modelo *in silico*.

Dado que el resultado final ($K_{p,uu}$) se calcula a partir de dos datos experimentales ($f_{u,cerebro}$ y $f_{u,plasma}$), cada uno de los cuales tiene un error (o desviación estándar, SD) muestral conocido, a continuación se deriva la expresión que permite propagar los errores al valor final de $K_{p,uu}$ informado[1].

En este contexto, podemos pensar a $K_{p,uu}$ como una función de dos variables, $f(x, y)$, la cual, si es diferenciable alrededor del punto (a,b), puede aproximarse en los entornos de dicho punto mediante la expansión por serie de Taylor de 1er orden:

$$f(x, y) \approx f(a, b) + \frac{\partial f}{\partial x}(x - a) + \frac{\partial f}{\partial y}(y - b) \tag{4.17}$$

Dado que X e Y representan dos variables aleatorias ($f_{u,cerebro}$ y $f_{u,plasma}$), la expresión (4.17) indica que $f(x,y) = K_{p,uu}$ se puede escribir una suma de variables aleatorias (VA) y valores constantes, por lo que se pueden aplicar las siguientes propiedades de la varianza (V) de una variable aleatoria (Walpole et al., 2012):

Propiedad 1

Sea Z una VA, y a una constante:

$$V(Z + a) = V(Z) \tag{4.18}$$

Propiedad 2:

Sea Z = aX + bY, donde X e Y son dos VA, a y b constantes. luego:

$$V(Z) = a^2V(X) + b^2V(Y) + 2ab \cdot COV(X, Y)$$

Si X e Y son independientes, su covarianza es cero, y la expresión anterior se transforma en:

$$V(Z) = a^2V(X) + b^2V(Y) \tag{4.19}$$

[1] *Si bien la expresión de cálculo de $K_{p,uu}$ incluye un tercer resultado (K_p), dado que este se tomó de fuentes bibliográficas, será considerado sin error. Por lo tanto, los errores reales finales serán algo mayores a los aquí informados.*

Aplicando la propiedad 2 (ec. 4.19) sobre la expresión (4.17), y teniendo en cuenta que, según la propiedad 1 (expresión (4.18)) la varianza de cualquier término constante es igual a cero, se tiene que:

$$V[f(x,y)] \approx \left(\frac{\partial f}{\partial x}\right)^2 V(X) + \left(\frac{\partial f}{\partial y}\right)^2 V(Y) \tag{4.20}$$

Reemplazando varianza (V) por SD al cuadrado (σ^2), y las variables X e Y por su significado experimental, la expresión (4.20) se transforma en:

$$\sigma^2_{Kp,uu} \approx \left(\frac{\partial Kp,uu}{\partial fu,cerebro}\right)^2 \sigma^2_{fu,cerebro} + \left(\frac{\partial Kp,uu}{\partial fu,plasma}\right)^2 \sigma^2_{fu,plasma} \tag{4.21}$$

Resta encontrar las expresiones para las derivadas en (4.21). Dado que:

$$K_{p,uu} = \frac{C_{u,cerebro}}{C_{u,plasma}} = K_p \frac{f_{u,cerebro}}{f_{u,plasma}} \tag{4.22}$$

$$\frac{\partial f}{\partial x} = \frac{\partial K_{p,uu}}{\partial f_{u,cerebro}} = \frac{K_p}{f_{u,plasma}} \tag{4.23}$$

$$\frac{\partial f}{\partial y} = \frac{\partial K_{p,uu}}{\partial f_{u,plasma}} = -\frac{K_p \cdot f_{u,cerebro}}{f^2_{u,plasma}} \tag{4.24}$$

Reemplazando (4.23) y (4.24) en (4.21):

$$\sigma^2_{Kp,uu} \approx \left(\frac{K_p}{f_{u,plasma}}\right)^2 \sigma^2_{fu,cerebro} + \left(-\frac{K_p \cdot f_{u,cerebro}}{f^2_{u,plasma}}\right)^2 \sigma^2_{fu,plasma} \tag{4.25}$$

Dividiendo (4.25) por la expresión de $K_{p,uu}$ (ecuación (4.16)) al cuadrado, se llega a:

$$\left(\frac{\sigma_{Kp,uu}}{K_{p,uu}}\right)^2 \approx \left(\frac{\sigma_{fu,cerebro}}{f_{u,cerebro}}\right)^2 + \left(\frac{\sigma_{fu,plasma}}{f_{u,plasma}}\right)^2 \tag{4.26}$$

Por último, se aproximan las desviaciones estándar (σ) por sus correspondientes estimas muestrales (s) y tomar la raíz cuadrada, para llegar a la expresión final empleada para el cálculo del error en $K_{p,uu}$:

$$\frac{s_{K_{p,uu}}}{K_{p,uu}} \approx \sqrt{\left(\frac{s_{fu,cerebro}}{f_{u,cerebro}}\right)^2 + \left(\frac{s_{fu,plasma}}{f_{u,plasma}}\right)^2} \tag{4.27}$$

El resultado (4.27) es la expresión conocida para la propagación de errores en productos y cocientes, el cual establece que la SD relativa del resultado se puede aproximar por la raíz de la suma de cuadrados de las SD relativas de los resultados que componen dicho producto o cociente (Skoog et al., 2013).

4.2. Validación externa de los modelos desarrollados con el set de datos MS

El mejor modelo desarrollado con el conjunto de datos MS fue validado externamente de utilizando datos obtenidos de dos maneras diferentes:

› Datos de $K_{p,uu}$ extraídos de publicaciones posteriores a la fecha del armado del conjunto de datos.

› Valores de $K_{p,uu}$ calculados a partir de datos internos de microdiálisis en estado no estacionario. Para ello, se utilizó el curso temporal de $K_{p,uu}$ para estimar el $K_{p,uu}$ en estado estacionario de acuerdo con la siguiente ecuación:

$$K_{p,uu(estado\ no\ estacionario)} = K_{p,uu}(1 - e^{-keq*t}) \quad (4.28)$$

Este manejo de datos ya se ha utilizado anteriormente (Cremers et al., 2012; Kalvass et al., 2007). Los parámetros se obtuvieron ajustando el modelo usando la técnica de mínimos cuadrados no lineales empleando la función nls en R (R Core Team, 2017). Esta función determina las estimaciones de mínimos cuadrados no lineales (ponderados) de los parámetros de un modelo no lineal.

En total, el conjunto de validación externa para los datos de $K_{p,uu}$ consistió en 10 compuestos, de los cuales 5 se extrajeron de publicaciones y 5 de datos de microdiálisis internos.

4.3. Métodos analíticos por HPLC

Para la cuantificación de los compuestos de interés en las diferentes muestras mencionadas en la sección 4.1, se utilizó un equipo UHPLC Dionex Ultimate 3000

(Thermo Scientific, Dionex, Sunnyvale, California, Estados Unidos) equipado con un detector de UV con arreglo de diodos. La fase estacionaria fue una columna Hibar C-18 (5 µm, 125 x 4,0 mm) (Merck KGaA, Darmstadt, Alemania), y en todos los casos el equipo fue operado de manera isocrática, a temperatura ambiente.

A continuación, se describen las condiciones cromatográficas particulares para cada uno de los compuestos evaluados:

Teofilina

Fase móvil: mezcla de metanol: solución buffer KH_2PO_4 20 mM pH = 2,5 (20:80).
Longitud de onda de detección: 271 nm.
Flujo: 1 mL/minuto.

Clorfeniramina

Fase móvil: mezcla de metanol: solución buffer KH_2PO_4 50 mM pH = 2,5 (40:60).
Longitud de onda de detección: 261 nm.
Flujo: 1 mL/minuto.

Ranitidina

Fase móvil: mezcla de metanol: solución buffer fosfato de sodio 20 mM pH = 7,4 (40:60).
Longitud de onda de detección: 316 nm.
Flujo: 1 mL/minuto.

Lidocaína

Fase móvil: mezcla de metanol: solución buffer KH_2PO_4 50 mM pH = 2,5 (25:75).
Longitud de onda de detección: 220 nm.
Flujo: 1 mL/minuto.

Ácido para-aminobenzoico

Fase móvil: mezcla de metanol: solución buffer KH_2PO_4 50 mM pH = 2,5 (5:95).
Longitud de onda de detección: 281 nm.
Flujo: 1 mL/minuto.

Referencias

Banker, M., & Clark, T. (2008). Plasma / Serum Protein Binding Determinations. *Current Drug Metabolism*, *9*(9), 854–859. https://doi.org/10.2174/138920008786485065

Chen, Y. C., Kenny, J. R., Wright, M., Hop, C. E. C. A., & Yan, Z. (2019). Improving Confidence in the Determination of Free Fraction for Highly Bound Drugs Using Bidirectional Equilibrium Dialysis. *Journal of Pharmaceutical Sciences*, *108*(3), 1296–1302. https://doi.org/10.1016/j.xphs.2018.10.011

Cremers, T. I. F. H., Flik, G., Hofland, C., & Stratford, R. E. (2012). Microdialysis Evaluation of Clozapine and N-Desmethylclozapine Pharmacokinetics in Rat Brain. *Drug Metabolism and Disposition*, *40*(10), 1909–1916. https://doi.org/10.1124/dmd.112.045682

Doan, K. M., Wring, S. A., Shampine, L. J., Jordan, K. H., Bishop, J. P., Kratz, J. et al. (2004). Steady-State Brain Concentrations of Antihistamines in Rats. *Pharmacology*, *72*(2), 92–98. https://doi.org/10.1159/000079137

Dow, N. (2006). Determination of Compound Binding to Plasma Proteins. *Current Protocols in Pharmacology*, *34*(1), 7.5.1-7.5.15. https://doi.org/10.1002/0471141755.ph0705s34

Kalvass, J. C., Maurer, T. S., Cory Kalvass, J., & Maurer, T. S. (2002). Influence of nonspecific brain and plasma binding on CNS exposure: implications for rational drug discovery. *Biopharmaceutics & Drug Disposition*, *23*(8), 327–338. https://doi.org/10.1002/bdd.325

Kalvass, J. C., Olson, E. R., Cassidy, M. P., Selley, D. E., & Pollack, G. M. (2007). Pharmacokinetics and pharmacodynamics of seven opioids in P-glycoprotein-competent mice: assessment of unbound brain EC50,u and correlation of in vitro, preclinical, and clinical data. *The Journal of Pharmacology and Experimental Therapeutics*, *323*(1), 346–355. https://doi.org/10.1124/jpet.107.119560

Loryan, I., Fridén, M., & Hammarlund-Udenaes, M. (2013). The brain slice method for studying drug distribution in the CNS. *Fluids and Barriers of the CNS*, *10*, 6.

https://doi.org/10.1186/2045-8118-10-6

Morales, J. F., Montoto, S. S., Fagiolino, P., & Ruiz, M. E. (2017). Current State and Future Perspectives in QSAR Models to Predict Blood- Brain Barrier Penetration in Central Nervous System Drug R&D. *Mini Reviews in Medicinal Chemistry*, *17*(3), 247–257. https://doi.org/10.2174/1389557516666161013110813

Nakazono, T., Murakami, T., Higashi, Y., & Yata, N. (1991). Study on Brain Uptake of Local Anesthetics in Rats. *Journal of Pharmacobio-Dynamics*, *14*(11), 605–613. https://doi.org/10.1248/bpb1978.14.605

R Core Team. (2017). *R: A language and environment for statistical computing*. Retrieved from http://www.r-project.org/

Romer, J., & Bickel, M. H. (1979). A method to estimate binding constants at variable protein concentrations. *Journal of Pharmacy and Pharmacology*, *31*(1), 7–11. https://doi.org/10.1111/j.2042-7158.1979.tb13411.x

Skoog, D. A., West, D. M., Holler, F. J., & Crouch, S. R. (2014). Fundamentals of analytical chemistry, 9th Ed. Cengage Learning, Boston, MA.

Toma, C. M., Imre, S., Vari, C. E., Muntean, D. L., & Tero-Vescan, A. (2021). Ultrafiltration method for plasma protein binding studies and its limitations. *Processes*, *9*(2), 382. https://doi.org/10.3390/pr9020382

Vuignier, K., Veuthey, J.-L., Carrupt, P.-A., & Schappler, J. (2013). Global analytical strategy to measure drug-plasma protein interactions: from high-throughput to in-depth analysis. *Drug Discovery Today*, *18*(21–22), 1030–1034. https://doi.org/10.1016/j.drudis.2013.04.006

Walpole, R., Myers, R., & Myers, S. (2012). *Probabilidad y estadística para ingeniería y ciencias,* 9ª Ed. Pearson Educación, México.

Yasuhara, M., & Levy, G. (1988). Kinetics of drug action in disease states XXVI: Effect of fever on the pharmacodynamics of theophylline-induced seizures in rats. *Journal of Pharmaceutical Sciences*, *77*(7), 569–570. https://doi.org/10.1002/jps.2600770704

Young, R. C., Mitchell, R. C., Brown, T. H., Ganellin, C. R., Griffiths, R., Jones, M. et al. (1988). Development of a new physicochemical model for brain penetration and its application to the design of centrally acting H2 receptor histamine antagonists. *Journal of Medicinal Chemistry*, *31*(3), 656–671. https://doi.org/10.1021/jm00398a028

Capítulo 5

Resultados y discusión

Como se describió en el capítulo anterior, se comenzó trabajando con un conjunto de datos formado en base a una búsqueda bibliográfica exhaustiva de compuestos que tuvieran reportado un valor experimental de $K_{p,uu}$; el conjunto de datos fue posteriormente sometido a un cuidadoso proceso de curado.

El conjunto de datos constaba inicialmente de 157 compuestos, 74 de alta permeabilidad en el SNC ("activos") y 83 de baja permeabilidad en el SNC ("inactivos"), y fue denominado **conjunto de datos MSH** por contener datos de $K_{p,uu}$ determinados por cualquiera de los tres métodos experimentales mencionados en los capítulos previos: microdiálisis (M), *slice* (S) u homogenato (H).

Debido a que, como se comentó en el Cap. 1, los valores de $K_{p,uu}$ obtenidos por la técnica de homogenato cerebral se consideran más variables (y, por lo tanto, menos confiables) en relación a las otras técnicas experimentales, se decidió repetir el esquema de trabajo con un segundo conjunto de datos, constituido solamente por los compuestos cuyo valor de $K_{p,uu}$ había sido determinado por microdiálisis o *slice* (**conjunto de datos MS**, 109 compuestos). A su vez, este conjunto de datos fue

refinado para explorar el efecto que pudieran tener los sustratos de transportadores ABC sobre los resultados del modelado (**conjunto de datos MS refinado**, 67 compuestos). La Figura 5.1 muestra esquemáticamente la composición de estos conjuntos de datos, incluyendo su balance de clases (alta/baja BD libre en SNC) y particiones realizadas.

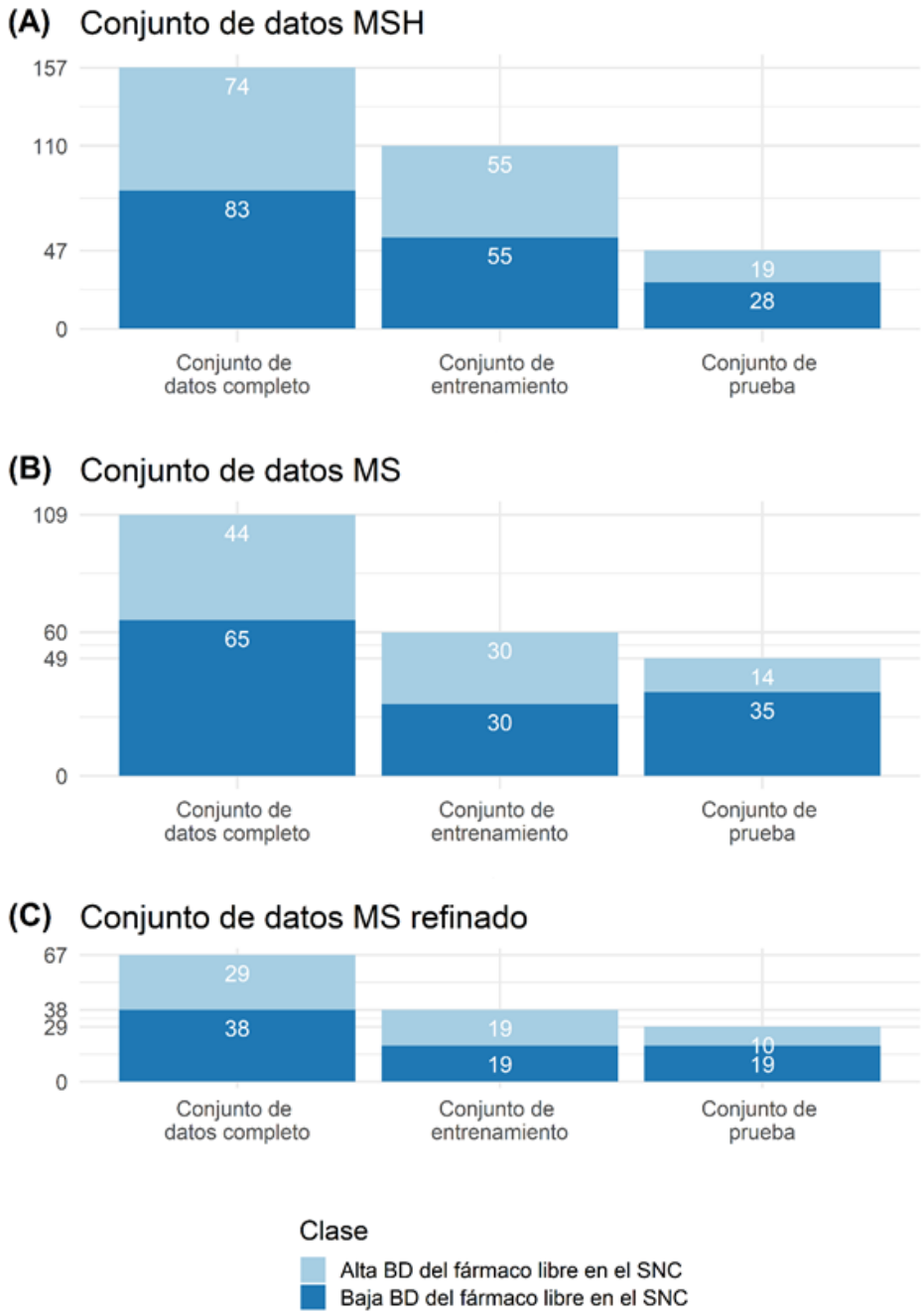


Figura 5.1. Composición de los diferentes conjuntos de datos utilizados para el desarrollo de los modelos. (A) Conjunto de datos MSH; (B) conjunto de datos MS; (C) conjunto de datos MS refinado.

El presente capítulo se encuentra estructurado en dos partes, correspondientes a los resultados obtenidos a partir del conjunto de datos MSH, por un lado, y de los conjuntos de datos MS y MS refinado, por el otro.

***

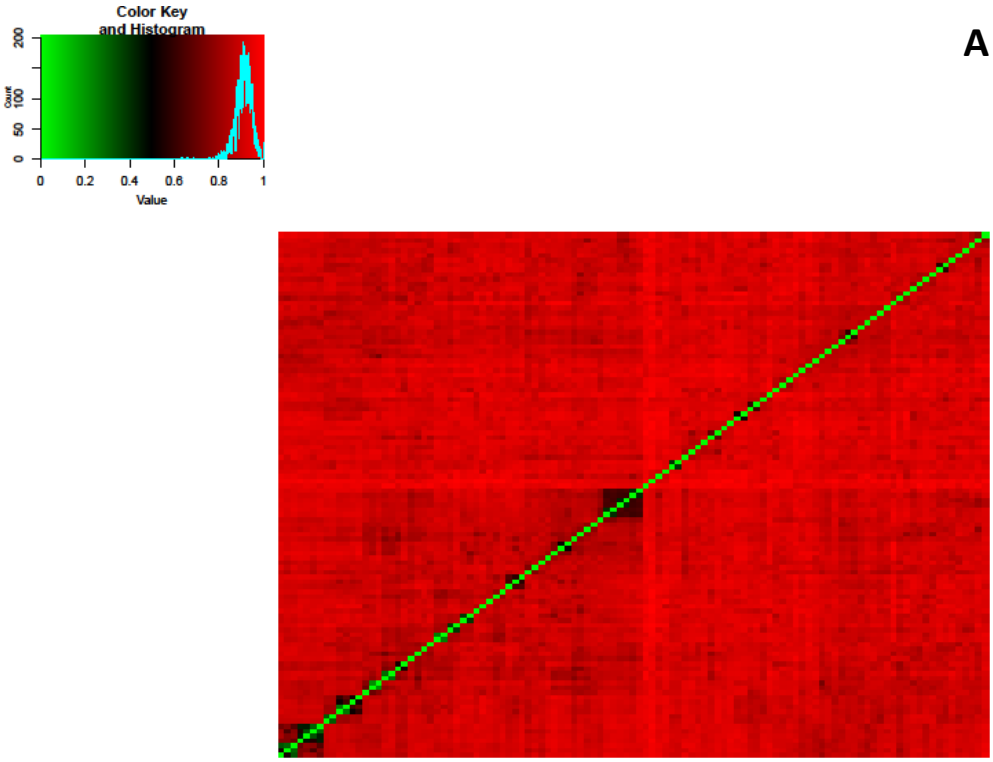
5.1. Resultados en el conjunto de datos MSH

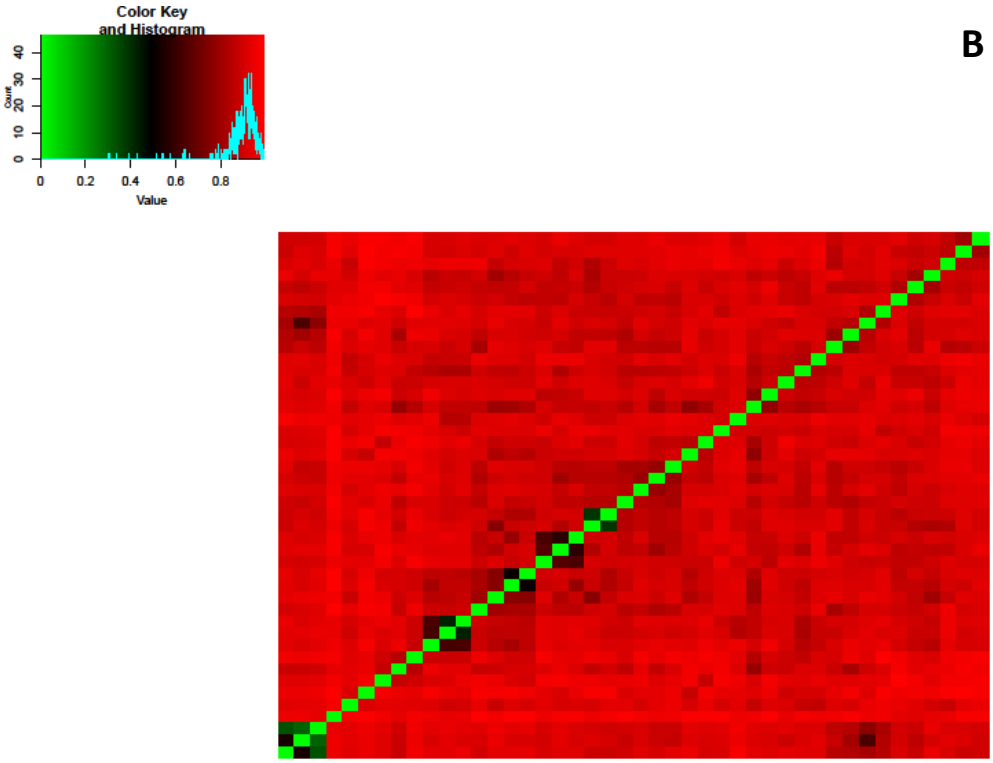
5.1.1. Conjunto de datos

Luego de la búsqueda bibliográfica inicial, se compilaron datos de $K_{p,uu}$ de 711 compuestos. A continuación, y aplicando el curado detallado en la sección 3.1, 554 compuestos fueron excluidos y tan sólo 157 compuestos pasaron a formar parte del conjunto de datos MSH (74 de alta BD libre y 83 de baja BD libre en el SNC).

En la Figura 5.2 se presentan los mapas de calor que ilustran la diversidad molecular de los compuestos del conjunto de datos MSH, en el cual se observa un elevado (y deseable) grado de diversidad molecular entre los compuestos (los bits en rojo corresponden a pares de compuestos de alta disimilitud molecular).

La distribución del conjunto de datos MSH en el espacio químico se puede visualizar en los gráficos de análisis de componentes principales (PCA, por sus siglas en inglés *Principal Component Analysis*) (Figura 5.3), así como en los histogramas de la Figura 5.4. Tanto el PCA como los histogramas se basan en ocho descriptores moleculares frecuentemente utilizados para caracterizar compuestos tipo fármaco (*druglike*): peso molecular [MW], área de superficie polar topológica [TPSA (Tot)], coeficiente de partición octanol-agua de Moriguchi [MLOGP], número de átomos donantes para enlaces de hidrogeno (N y O) [nHDon], número de átomos aceptores para enlaces de hidrogeno (N, O, F) [nHAcc], número de enlaces rotativos [RBN], número de anillos (número ciclomático) [nCIC] y suma de los volúmenes atómicos de van der Waals [Sv].


Color Key
and Histogram
Count
0
50
100
200
0
0.2
0.4
0.6
0.8
1
Value
A


Color Key
and Histogram
Count
0
10
20
30
40
0
0.2
0.4
0.6
0.8
Value
B

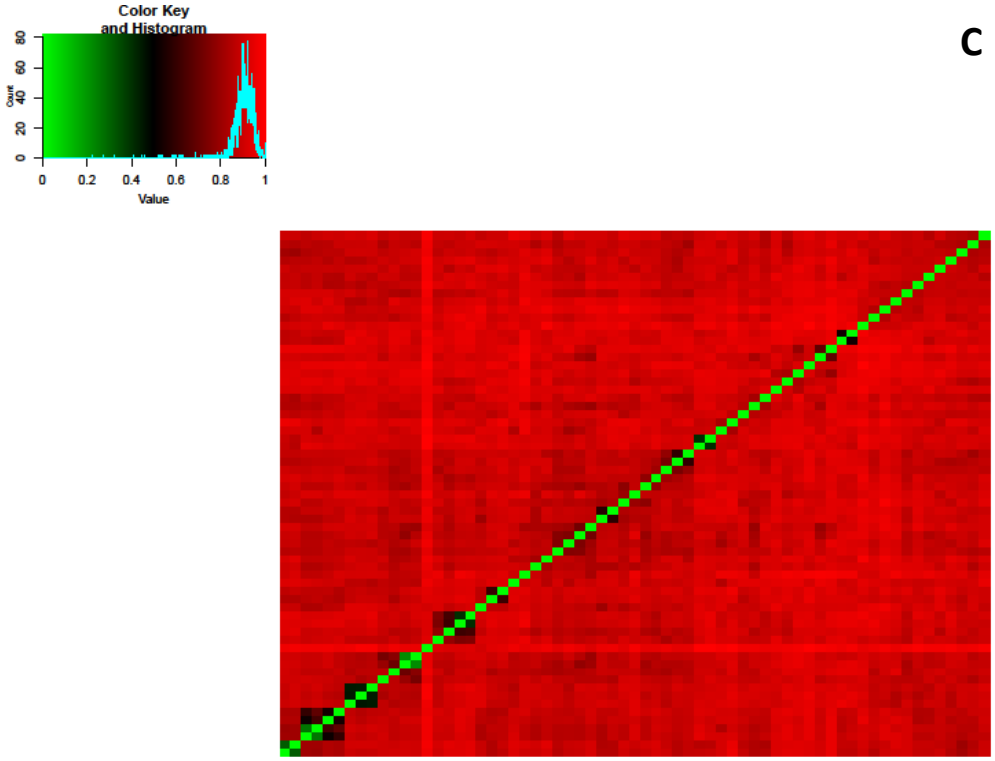


Figura 5.2. Mapas de calor que reflejan la diversidad molecular de los compuestos en el conjunto de datos MSH. Los histogramas anexos a cada gráfica muestran la referencia para los colores y la distribución de los compuestos respecto a su disimilitud medida según la distancia de Tanimoto (un bit verde en el mapa de calor representa un par de compuestos con distancia de Tanimoto igual o similar a cero, es decir, idénticos o muy similares; en el otro extremo, un bit rojo representa un par de compuestos altamente disímiles). Se utilizó ECFP_6 como sistema de *fingerprints* o huellas digitales moleculares. (A) Todos los 157 compuestos del conjunto de datos MSH. (B) 74 compuestos de alta BD como fármaco libre en el SNC. (C) 83 compuestos de baja BD como fármaco libre en el SNC.

Puede observarse que, respecto a las componentes principales de los ocho descriptores moleculares utilizados en esta instancia para caracterizar a los compuestos, el grupo de compuestos de baja BD como fármaco libre en el SNC (triángulos verdes en la figura) se encuentra ampliamente distribuido en el espacio, lo cual no ocurre para el grupo de alta BD como fármaco libre en el SNC (círculos rojos). Para esta última clase, se puede ver que la región del espacio químico que ocupan los compuestos es más restringida, y se superpone completamente con parte de la región correspondiente al grupo de baja BD en el SNC. Los histogramas en la Figura 5.4 representan la frecuencia de distribución de esos ocho descriptores moleculares en el conjunto de datos MSH. Los histogramas muestran que ambos

grupos de compuestos prácticamente no se diferencian (barras turquesas y rojas para los grupos de baja y alta BD en el SNC, respectivamente, y barras grises para el total de compuestos), observándose mayor diversidad química en los compuestos de baja BD en el SNC. Como puede esperarse, las distribuciones de frecuencia de la Figura 5.4 concuerdan con el análisis de PCA discutido previamente, mostrando una superposición considerable entre ambos grupos.

Mientras que la distribución de los compuestos en el espacio químico es muy informativa (por ejemplo, nos dice que es muy poco probable que algunas regiones del espacio químico estén ocupadas por compuestos con alta BD en el SNC), las regiones superpuestas entre ambos grupos representan un desafío para el desarrollo de modelos predictivos del parámetro $K_{p,uu}$. Por lo tanto, la utilización para el modelado de un gran número de descriptores moleculares, así como diversos algoritmos de clasificación, se justifican como una buena estrategia para poder superar este inconveniente.

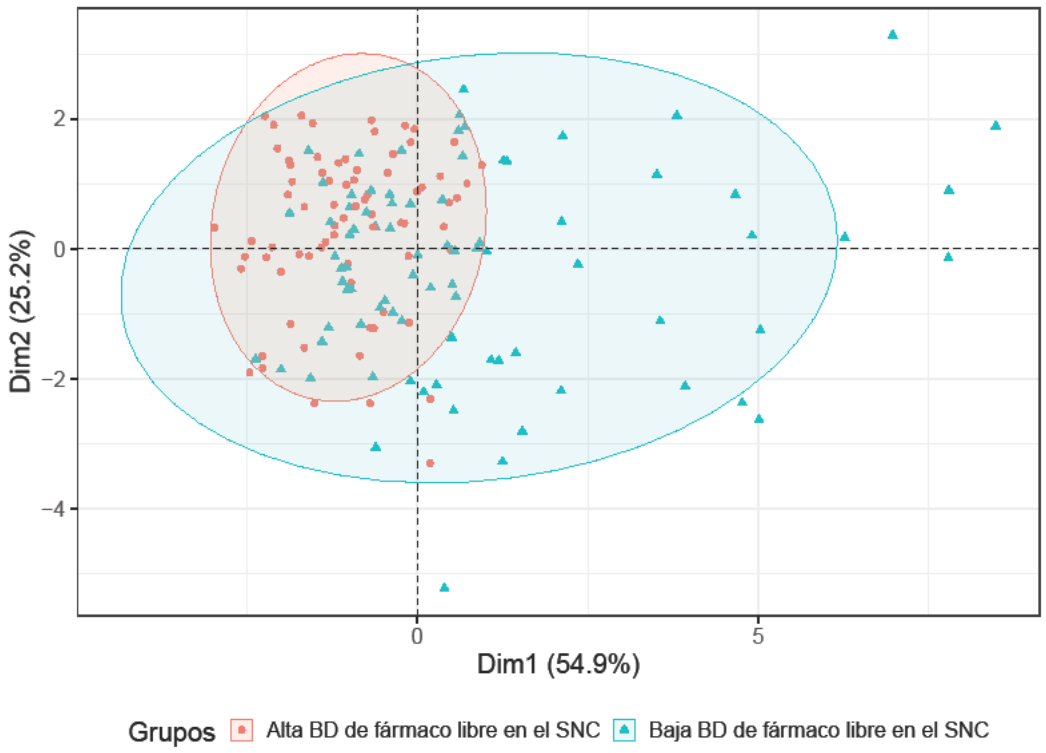


Figura 5.3. Gráfico de PCA del conjunto MSH (157 compuestos), basado en ocho descriptores fisicoquímicos (MW; TPSA (Tot); MLogP; nHDon; nHAcc; RBN; nCIC y Sv). Los triángulos verdes y círculos rojos representan a los compuestos de baja y alta BD en el SNC, respectivamente. Las elipses dibujadas corresponden a los intervalos del 90% de confianza asumiendo una distribución normal multivariada de los datos por grupo.

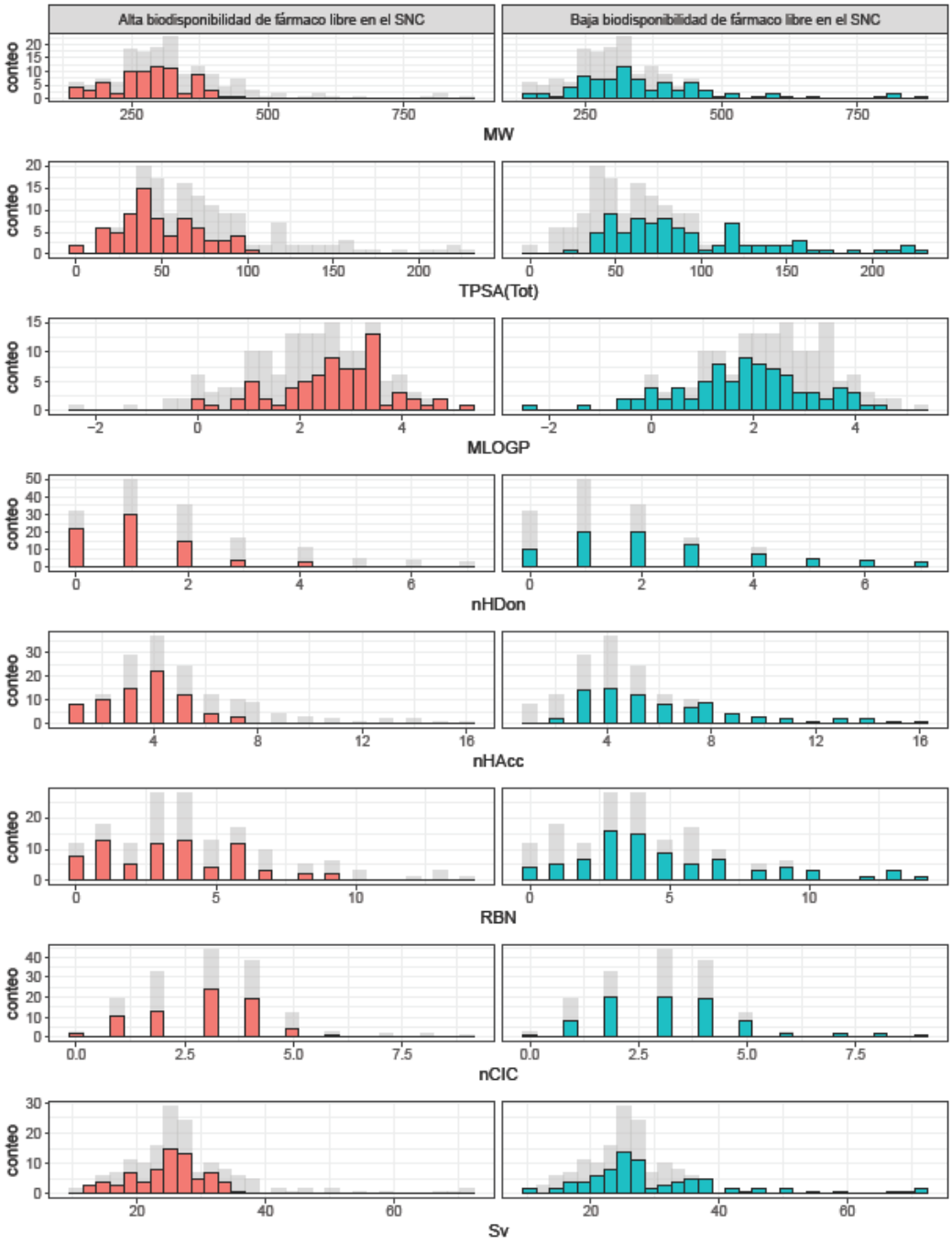


Figura 5.4. Histogramas que muestran la distribución de frecuencia de los descriptores fisicoquímicos seleccionados en todo el conjunto de datos MSH. Las barras grises representan la frecuencia en el conjunto de datos MSH total, mientras que las barras rojas y verdes corresponden a los grupos de alta y baja BD en el SNC, en ese orden.

5.1.2. Partición del conjunto de datos en los conjuntos de entrenamiento y de prueba

Recurriendo a la metodología de agrupamiento descripta en el capítulo 3, se obtuvieron separadamente los agrupamientos de las categorías de alta y baja BD del

fármaco libre en el SNC a partir de los cuales se generaron los conjuntos de entrenamiento y de prueba.

Idealmente, el conjunto de entrenamiento debe presentar una composición equilibrada entre los compuestos con alta y baja BD de fármaco libre en el SNC para evitar el sesgo hacia la predicción de una categoría en particular. Por esta razón, el 74% de cada grupo de los compuestos de alta BD del fármaco libre en el SNC y el 66% de cada grupo de compuestos de baja BD del fármaco libre en el SNC se asignaron al conjunto de entrenamiento, lo que resulta en igual número de compuestos de alta y baja BD en ese conjunto; los compuestos restantes se utilizaron como conjunto de prueba. El conjunto de entrenamiento resultante consistió en 110 compuestos (55 con alta y 55 con baja BD del fármaco libre en el SNC), mientras que el conjunto de prueba de 47 compuestos incluyó 19 compuestos con alta BD en el SNC y 28 compuestos con baja BD en el SNC (Figura 5.1.A).

5.1.3. Cálculo de descriptores

Luego de calcular los 3668 descriptores independientes de la conformación para cada compuesto utilizando el software Dragon 6.0 (Milano Chemometrics, 2011), se prosiguió con los criterios de exclusión anteriormente mencionados, finalizando con un conjunto de 1807 descriptores moleculares que fueron utilizados a los fines de modelado.

5.1.4. Modelos

La Tabla 5.1 muestra los resultados obtenidos con los modelos desarrollados utilizando el conjunto de datos MSH. Se presentan los valores de área bajo la curva ROC (ABC_ROC) y la tasa de buenas clasificaciones o exactitud (Acc) para los conjuntos de entrenamiento y prueba. Para el conjunto de prueba, se presentan también los valores de sensibilidad (Se), especificidad (Sp) y coeficiente de correlación de Matthews (MCC). Acorde a lo de descripto en la sección 3.3 del capítulo 3, el valor de MCC junto con el balance Se/Sp fueron el criterio utilizado para establecer los valores de corte del *score* para definir el criterio de alta/baja BD en el SNC de cada uno de los modelos desarrollados.

Una vez definidos los valores de corte, se calcularon los parámetros de desempeño, los cuales permiten evaluar la capacidad de clasificación de los modelos, y seleccionar el mejor modelo. La tabla también presenta los resultados de la validación cruzada. Puede apreciarse que todos los modelos tienen un buen desempeño en el conjunto de prueba, con valores de ABC_ROC, *Acc* y *MCC* entre 0,821-0,900, 76,6%-85,1% y 0,546-0,696, respectivamente. Basándonos en los valores de *MCC* en el conjunto de prueba, hay cinco algoritmos que se encuentran por debajo del valor 0,6: SVM, kNN, cPLS, DNN y el mejor modelo individual desarrollado por nuestro algoritmo interno LDA. Estos modelos también presentan valores de *Acc* en el conjunto de prueba menores a 78,7%, por lo que podríamos decir que estos algoritmos son los que exhibieron un desempeño modesto. Los ensamblados de los mejores modelos obtenidos por LDA junto con RF son los algoritmos que siguen en términos de desempeño, con valores en el conjunto de prueba de *MCC* entre 0,620 y 0,647, y valores de *Acc* desde 76,6% hasta 83,0%. Por último, los dos mejores modelos, utilizando como criterio el valor de *MCC*, fueron los que surgieron de las metodologías de *boosting* (sGBM y XGBOOST). Basándonos en su desempeño en el conjunto de prueba (el cual también se correlaciona con el resultado de la validación cruzada) podemos decir que XGBOOST condujo al mejor modelo individual, ya que sus valores de *Acc* y *MCC* en el conjunto de prueba (85,1% y 0,696, respectivamente) son los más altos de entre todos los obtenidos con los diversos algoritmos evaluados, siendo estos valores muy similares a los conseguidos por el mejor modelo de clasificación desarrollado hasta la fecha, reportado por Chen et al. (Chen et al., 2011) (*MCC* = 0,72 y *Acc* = 85%). Cabe resaltar que si se hubiera tenido solo en cuenta los valores de ABC_ROC en el conjunto de prueba (0,900 y 0,891 para sGBM y XGBOOST, respetivamente), se habría seleccionado el modelo sGBM, siendo una decisión diferente a la obtenida cuando se utilizó como criterio *Acc* y *MCC*. Como ya se comentó, el parámetro *MCC* tiene la ventaja que es una métrica equilibrada que mantiene su eficiencia incluso cuando las clases de compuestos está muy desbalanceadas, como es el caso de nuestro conjunto de prueba. Por otra parte, el balance Se/Sp para el modelo XGBOOST resulta el más conservador para optimizar el uso de recursos, en tanto se asocia a una baja tasa de falsos positivos.

Tabla 5.1. Resultados de los modelos obtenidos por los diferentes algoritmos en el conjunto de datos MSH. Se muestran los resultados tanto para el conjunto de entrenamiento como para el conjunto de prueba. El mejor modelo está resaltado en negrita. (*) *Acc* de los modelos individuales del ensamblado.

Algoritmo	**Conjunto de entrenamiento**			**Conjunto de prueba**				
	ABC_ROC	**Acc**	**Acc promedio (validación cruzada; 500 iteraciones)**	**ABC_ROC**	**Acc**	**MCC**	**Se**	**Sp**
SVM	0,904	84,5	69,4	0,840	78,7	0,557	0,579	0,929
sGBM	0,999	99,1	72,9	0,900	80,9	0,653	0,947	0,714
*k*NN	0,775	71,8	69,2	0,827	76,6	0,546	0,842	0,714
cPLS	0,967	93,6	69,4	0,828	78,3	0,547	0,684	0,852
RF	1,000	100,0	72,9	0,860	82,6	0,638	0,737	0,889
XGBOOST	**1,000**	**100,0**	**75,1**	**0,891**	**85,1**	**0,696**	**0,684**	**0,964**
DNN	1,000	100,0	70,3	0,830	78,3	0,588	0,474	1,000
LDA – mejor modelo individual	0,958	90,9	68,7	0,821	78,7	0,553	0,684	0,857
LDA – Ensamblado de 2 mejores modelos indiv. *(operador mínimo)*	0,965	90,9	(68,7; 70,5)*	0,829	76,6	0,620	1,000	0,607
LDA – Ensamblado de 5 mejores modelos indiv. (*operador promedio)*	0,987	94,5	(61,6; 62,7; 66,9; 68,7; 70,5)*	0,850	83,0	0,647	0,789	0,857

5.1.5. Dominio de aplicación

Se realizó la evaluación del dominio de aplicación acorde a la metodología descripta en la sección 3.6. Se encontró una cobertura de aproximadamente el 94% (44/47) para los compuestos del conjunto de prueba, lo que indica que muy pocas extrapolaciones fueron realizadas. Entre los compuestos incorrectamente clasificados por el modelo, ninguno corresponde a los compuestos extrapolados.

En el caso de los compuestos de la validación experimental, todos ellos se encontraron dentro del dominio de aplicación del mejor modelo.

5.1.6. Validación experimental

Luego del desarrollo y la selección del mejor modelo, se prosiguió con la validación experimental del mismo de manera prospectiva. Para esto se estimó el valor de $K_{p,uu}$ en estado estacionario de nuevos compuestos que no formaron parte del conjunto de datos MSH.

Como resultado de la búsqueda bibliográfica y posterior selección según los criterios establecidos en la metodología, se obtuvieron cinco compuestos para la validación experimental: clorfeniramina, teofilina, ranitidina, ácido para-aminobenzoico (PABA) y lidocaína. La Tabla 5.2 muestra los resultados obtenidos. La $f_{u,cerebro}$ fue determinada por el método del homogenato en todos los casos, mientras que la $f_{u,plasma}$ primeramente se intentó determinar por ultrafiltración, pero debido a que todos los compuestos presentaron un grado de unión inespecífica mayor al 5%, el valor de este parámetro se determinó mediante la técnica de diálisis de equilibrio.

El conjunto de compuestos utilizado para la validación experimental estaba formado por tres compuestos de baja BD libre en el SNC (teofilina, ranitidina y PABA) y dos de alta BD libre en el SNC (clorfeniramina y lidocaína), teniendo así datos de ambas clases. Todos los compuestos elegidos se encuentraban dentro del dominio de aplicación del modelo, por lo que las predicciones pueden considerarse confiables. Como puede observarse en la Tabla 5.2, sólo uno de los compuestos fue mal clasificado por nuestro mejor modelo. Aunque el número de compuestos evaluados experimentalmente es pequeño, la *Acc* observada parecería ser similar al 80,0%,

esto es, cercana a la obtenida en el conjunto de prueba. Si indagamos el compuesto mal clasificado, teofilina, podemos observar que su valor experimental de $K_{p,uu}$ es de 0,119 (baja BD en el SNC), siendo este el más cercano de todos los compuestos probados al punto de corte establecido para realizar la clasificación (0,4, acorde a lo establecido en la sección 3.1, capítulo 3). Dicha cercanía al punto de corte podría ser una posible explicación de la mala predicción de este compuesto. Teniendo nuestro mejor modelo un poder predictivo muy similar al del mejor modelo reportado hasta la fecha, podríamos considerarlo apto para su aplicación como herramienta para asistir el desarrollo de nuevos fármacos destinados al SNC.

Tabla 5.2. Resultados de la validación experimental prospectiva del mejor modelo obtenido con el conjunto de datos MSH.

	Clorfeni-ramina	**Teofilina**	**Ranitidina**	**PABA**	**Lidocaína**
K_p	34	0,7	0,058	0,04	2,2
$f_{u,plasma}$ **(SD)**	0,15 (0,02)	0,40 (0,06)	0,78 (0,16)	0,97 (0,11)	0,45 (0,06)
$f_{u,cerebro}$ **(SD)**	0,06 (0,01)	0,07 (0,013)	0,49 (0,08)	0,30 (0,08)	0,18 (0,04)
$K_{p,uu}$ **calculado (SD)**	13,6 (3,9)	0,119 (0,03)	0,036 (0,009)	0,012 (0,003)	0,88 (0,23)
Clase real	alta BD libre en el SNC	baja BD libre en el SNC	baja BD libre en el SNC	baja BD libre en el SNC	alta BD libre en el SNC
Clase predicha	alta BD libre en el SNC	alta BD libre en el SNC	baja BD libre en el SNC	baja BD libre en el SNC	alta BD libre en el SNC
Predicción correcta	**Sí**	No	**Sí**	**Sí**	**Sí**
Referencia	(Doan et al., 2004)	(Yasuhara et al., 1988)	(Young et al., 1988)	(Nakazono et al., 1991)	(Nakazono et al., 1991)

Algo a remarcar de los resultados obtenidos en la validación experimental es que incluso compuestos que, a partir de sus valores de $K_{p,uu,}$ podrían ser sustratos de transportadores, fueron bien clasificados por nuestro mejor modelo. Un ejemplo sería el caso de la clorfeniramina, cuyo valor de $K_{p,uu}$ estimado (13,6) sugiere que el mismo es activamente captado a nivel de la BHE. Este proceso no se encuentra

descripto en bibliografía; sin embargo, se sabe que mepiramina (un antagonista H1 clásico) es captado por las células endoteliales cerebrales a través de un mecanismo saturable, el cual puede ser inhibido por la clorfeniramina (Tamai et al., 2000).

***

5.2. Resultados en el conjunto de datos MS

5.2.1. Conjunto de datos y partición

El conjunto de datos MS se generó excluyendo del conjunto MSH aquellos compuestos cuyo valor experimental de $K_{p,uu}$ había sido obtenido por la técnica de homogenato, generando así un subconjunto de 109 compuestos (44 de alta BD libre y 65 de baja BD libre en el SNC). En cuanto a la partición del conjunto de datos, se procedió de igual manera que en la sección anterior para lograr un conjunto de entrenamiento equilibrado en términos del contenido de compuestos con alta y baja BD de fármaco libre en el SNC para evitar sesgos de predicción. Por ello, el 68% de cada grupo de los compuestos de alta BD libre en el SNC y el 46% de cada grupo de compuestos de baja BD libre en el SNC se asignaron al conjunto de entrenamiento, lo que resultó en igual número de compuestos de alta y baja BD del fármaco libre en el SNC en ese conjunto; los compuestos restantes se usaron como conjunto de prueba. El conjunto de entrenamiento así generado consistió en 60 compuestos (30 con alta y 30 con baja BD libre en el SNC), mientras que el conjunto de prueba de 49 compuestos incluyó 14 compuestos con alta BD en el SNC y 35 compuestos con baja BD en el SNC (Figura 5.1.B).

En la Figura 5.5 podemos visualizar los mapas de calor para el conjunto de datos MS. Se puede observar, también como en el caso del conjunto de datos MSH, una gran diversidad molecular. Además, en las Figuras 5.6 y 5.7 se muestra el grafico de PCA y los histogramas de los 8 descriptores fisicoquímicos relevantes para el descubrimiento de fármacos en el conjunto MS, respectivamente. Del mismo modo que en el conjunto de datos MSH, los compuestos de baja BD en el SNC presentan una mayor diversidad fisicoquímica que los compuestos de alta BD en el SNC. Incluso puede visualizarse que el grafico de PCA es bastante similar al del conjunto

MSH (Figura 5.3), por lo cual se podría pensar que ambos conjuntos tienen una complejidad similar.

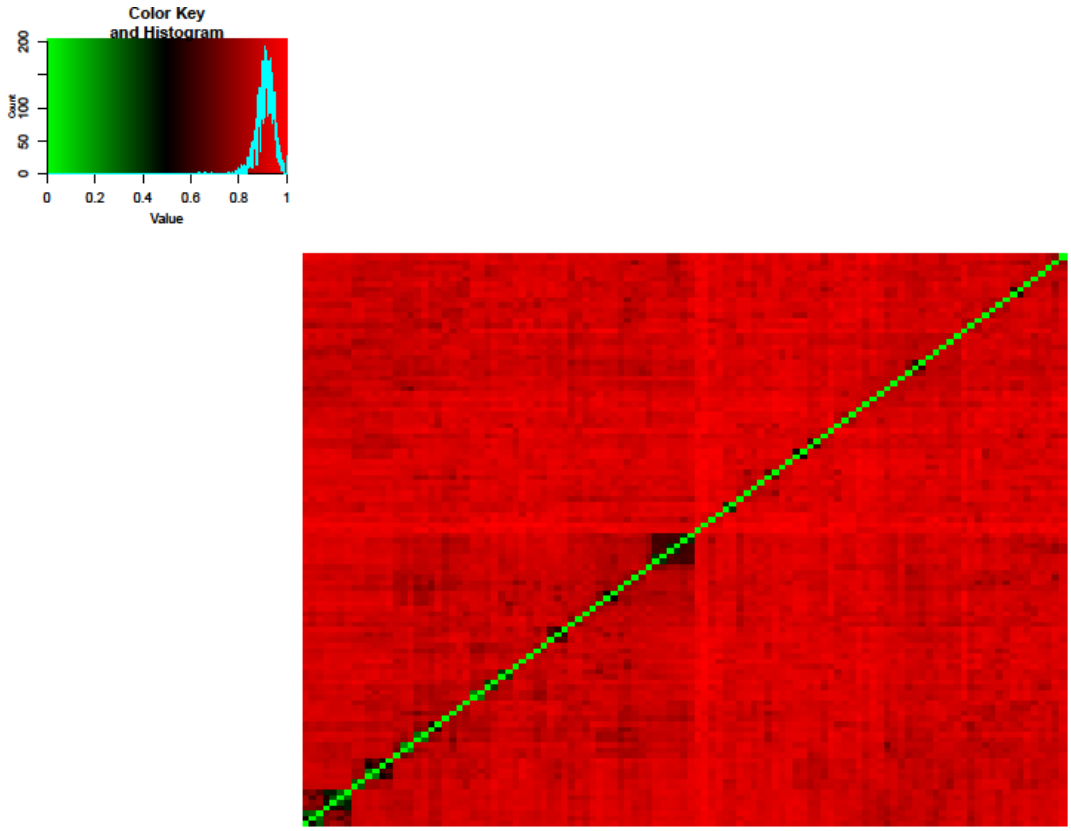


Figura 5.5. Mapa de calor que refleja la diversidad molecular de los compuestos en el conjunto de datos MS (109 compuestos). El histograma a la izquierda muestra la referencia para los colores y la distribución de los compuestos respecto a su disimilitud medida según la distancia de Tanimoto (un bit verde en el mapa de calor representa un par de compuestos con distancia de Tanimoto igual o similar a cero, es decir, idénticos o muy similares; en el otro extremo, un bit rojo representa un par de compuestos altamente disímiles). Se utilizó ECFP_6 como sistema de *fingerprints* o huellas digitales moleculares.

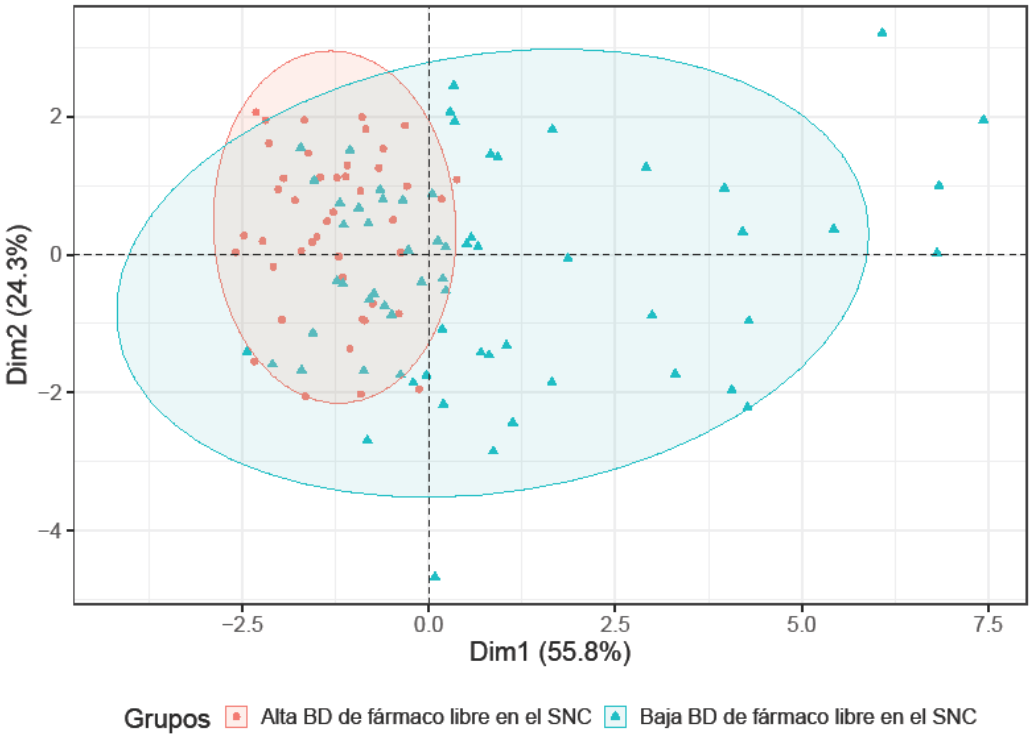


Figura 5.6. Gráfico de PCA del conjunto de datos MS (109 compuestos), basado en ocho descriptores fisicoquímicos (MW; TPSA (Tot); MLogP; nHDon; nHAcc; RBN; nCIC y Sv). Los puntos de datos están coloreados por grupo (triángulos verdes y círculos rojos para baja y alta BD en el SNC, respectivamente). Las elipses dibujadas corresponden a los intervalos del 90% de confianza asumiendo una distribución normal multivariada de los datos por grupo.

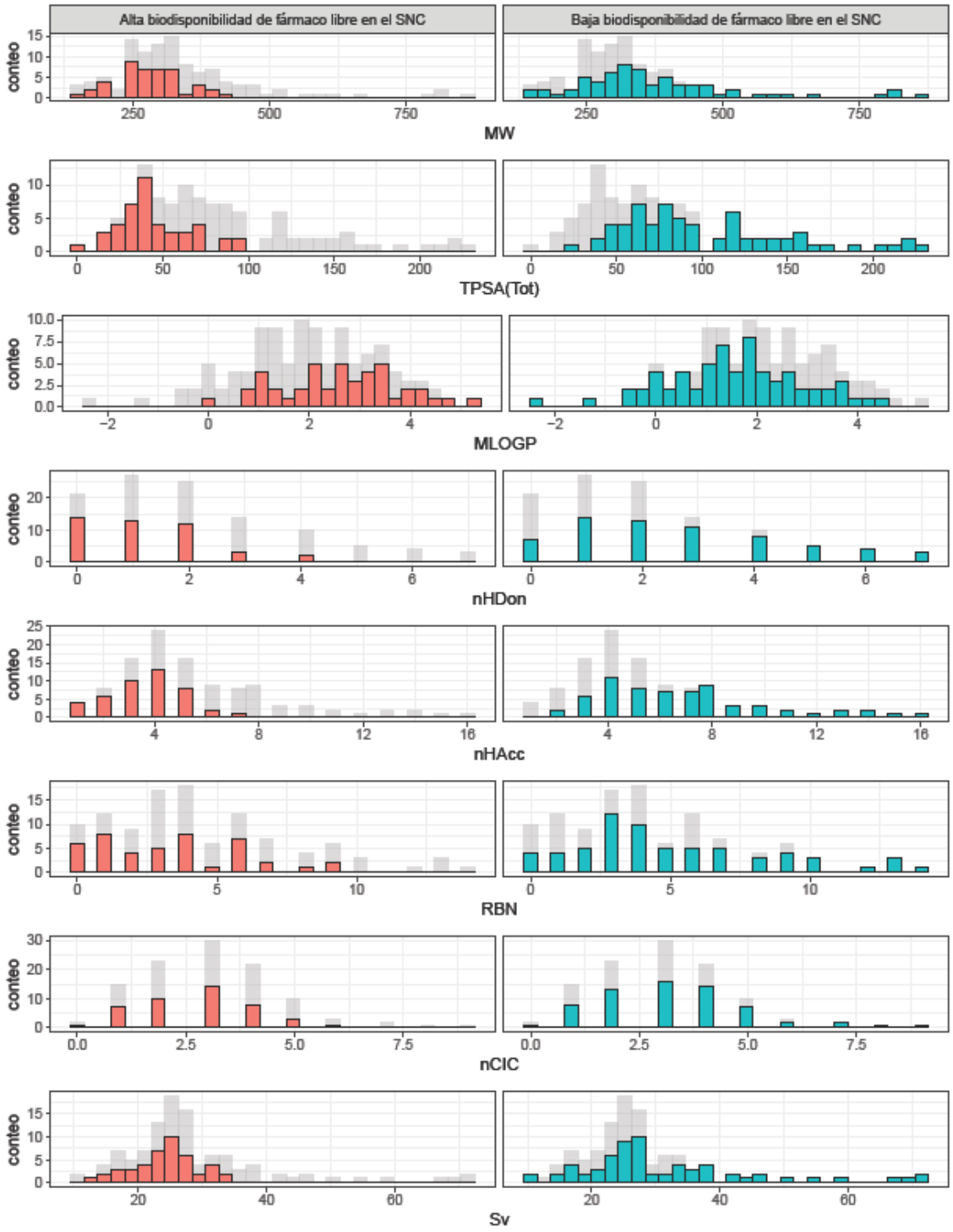


Figura 5.7. Histogramas que muestran la distribución de frecuencia de los descriptores fisicoquímicos seleccionados en todo el conjunto de datos MS. Las barras grises representan la frecuencia en el conjunto de datos MS total, mientras que las barras rojas y verdes corresponden a los grupos de alta y baja BD en el SNC, en ese orden.

5.2.2. Cálculo de descriptores

Luego de calcular los 3668 descriptores independientes de la conformación para cada compuesto utilizando el software Dragon 6.0 (Milano Chemometrics, 2011), se prosiguió con los criterios de exclusión ya mencionados, finalizando con un conjunto de 1848 descriptores moleculares que fueron utilizados a los fines de modelado.

5.2.3. Modelos

La Tabla 5.3 muestra los resultados obtenidos con los modelos desarrollados utilizando el conjunto de datos MS. Como en el caso de la Tabla 5.1, la Tabla 5.3 presenta los valores de ABC_ROC y *Acc* para los conjuntos de entrenamiento y prueba, así como los valores de Se, *Sp* y *MCC* para el conjunto de prueba. Todos estos parámetros, como en el caso del conjunto de datos MSH, fueron utilizados para evaluar la capacidad de clasificación de los modelos y para seleccionar al modelo con mejor desempeño. En la tabla se muestran, asimismo, los resultados de la validación cruzada. Puede apreciarse que los ensamblados de los mejores modelos obtenidos por LDA, así como aquellos modelos que resultaron de los algoritmos cPLS y kNN, fueron los que mostraron el peor desempeño en el conjunto de prueba (*MCC* = 0,331 y *Acc* = 72,9%, para el ensamblado con el operador mínimo, *MCC* = 0,426 y *Acc* = 64,6%, para el ensamblado con el operador promedio, *MCC* = 0,632 y *Acc* = 78,7% para cPLS y *MCC* = 0,533 y *Acc* = 81,6% para kNN). RF, SVM y el mejor modelo individual desarrollado por nuestro algoritmo interno LDA tuvieron mejores resultados. Los algoritmos de modelado restantes (sGBM, XGBOOST y DNN) funcionaron considerablemente bien, con valores de *Acc*, *MCC* y ABC_ROC en el conjunto de prueba iguales o mayores que 91,5%, 0,801 y 0,939, respectivamente. Estos valores superan a los del modelo obtenido por Chen et al. (Chen et al., 2011) (*MCC* = 0,72 y *Acc* = 85%). Sin embargo, debe tenerse en cuenta que el número de compuestos utilizados en el trabajo de Chen et al. (173 y 73 compuestos para los conjuntos de entrenamiento y de prueba, respectivamente) es superior al empleado aquí.

Teniendo en cuenta criterios adicionales como la facilidad del método de modelado y el costo computacional asociado, se decidió dejar de lado el modelo por DNN, y

continuar el análisis de los modelos que utilizan la metodología de boosting (sGBM y XGBOOST). Una comparación estadística del ABC_ROC de estos dos modelos para el conjunto de prueba (realizada utilizando la función roc.test del paquete de R pROC (Robin et al., 2011) con el método de DeLong (DeLong et al., 1988)) no mostró diferencias significativas entre ellos, pero la exploración de los valores de *Se* y *Sp* reveló que XGBOOST tiene un sesgo hacia una *Sp* alta, mientras que sGBM mostró una relación *Se*/*Sp* más equilibrada. Por lo tanto, decidimos seleccionar sGBM como el mejor modelo en este conjunto de datos.

Tabla 5.3. Resultados de los modelos obtenidos por los diferentes algoritmos en el conjunto de datos MS. Se muestran los resultados tanto para el conjunto de entrenamiento como para el conjunto de prueba. El mejor modelo está resaltado en negrita. (*) *Acc* de los modelos individuales del ensamblado.

Algoritmo	**Conjunto de entrenamiento**			**Conjunto de prueba**				
	ABC_ROC	**Acc**	**Acc promedio (validación cruzada; 500 iteraciones)**	**ABC_ROC**	**Acc**	**MCC**	**Se**	**Sp**
SVM	0,932	90,0	71,6	0,914	89,8	0,742	0,710	0,971
sGBM	**1,000**	**100,0**	**73,7**	**0,951**	**91,8**	**0,812**	**0,930**	**0,914**
*k*NN	0,820	75,0	69,1	0,835	81,6	0,533	0,360	1,000
cPLS	0,951	93,3	71,0	0,889	78,7	0,632	1,000	0,706
RF	1,000	100,0	67,4	0,937	89,4	0,732	0,620	1,000
XGBOOST	0,983	95,0	74,1	0,959	91,8	0,801	0,710	1,000
DNN	1,000	100,0	67,8	0,939	91,5	0,801	0,850	0,912
LDA – mejor modelo individual	0,996	96,7	68,3	0,910	85,4	0,728	1,000	0,794
LDA – Ensamblado de 2 mejores modelos indiv. *(operador mínimo)*	0,992	98,3	(63,3; 68,3)*	0,702	72,9	0,331	0,500	0,824
LDA – Ensamblado de 2 mejores modelos indiv. (*operador promedio)*	1,000	100,0	(63,3; 68,3)*	0,756	64,6	0,426	0,929	0,529

5.2.4. Dominio de aplicación

Se realizó la evaluación del dominio de aplicación, y se obtuvo un porcentaje de cobertura de aproximadamente el 94% (46/49) de los compuestos del conjunto de prueba, indicando nuevamente que muy pocas extrapolaciones fueron realizadas. Entre los compuestos incorrectamente clasificados por el modelo, ninguno correspondió a los compuestos extrapolados. En el caso de los compuestos de la validación externa o experimental, todos los compuestos se encontraron dentro del dominio de aplicación del mejor modelo.

5.2.5. Validación externa o experimental

La Tabla 5.4 detalla los resultados de la validación externa del mejor modelo desarrollado con la base de datos MS, realizada siguiendo la metodología descripta en la sección 4.2. El conjunto de validación externa se basó en 10 compuestos, 6 de alta y 4 de baja BD libre en el SNC. La *Acc* global fue del 90,0% (1 compuesto mal clasificado, de 10), lo cual está en concordancia con el valor de *Acc* estimado a partir del conjunto de prueba (91,8%).

Tabla 5.4. Desempeño del mejor modelo en el conjunto de validación externa.

Origen de los datos	Compuesto	Ref.	$K_{p,uu}$	Clase real	Clase predicha	Predicción correcta
Datos internos	Anfetamina	-	1,72	alta BD en el SNC	alta BD en el SNC	**Sí**
Datos internos	Aripiprazol	-	1,53	alta BD en el SNC	baja BD en el SNC	No
Datos internos	Duloxetina	-	0,86	alta BD en el SNC	alta BD en el SNC	**Sí**
Datos internos	Hidroxibupropion	-	1,90	alta BD en el SNC	alta BD en el SNC	**Sí**
Datos internos	Mecamilamina	-	1,70	alta BD en el SNC	alta BD en el SNC	**Sí**
Bibliografía	7-Hidroxitraginina	(Yusof et al., 2019)	0,08	baja BD en el SNC	baja BD en el SNC	**Sí**
Bibliografía	Bepotastina	(Kanamitsu et al., 2017)	0,12	baja BD en el SNC	baja BD en el SNC	**Sí**
Bibliografía	Ketotifeno	(Kanamitsu et al., 2017)	5,18	alta BD en el SNC	alta BD en el SNC	**Sí**
Bibliografía	Mitraginina	(Yusof et al., 2019)	0,09	baja BD en el SNC	baja BD en el SNC	**Sí**
Bibliografía	Olopatadina	(Kanamitsu et al., 2017)	0,16	baja BD en el SNC	baja BD en el SNC	**Sí**

5.2.6. Resultados en el conjunto de datos MS refinado

Por último, se trabajó con el subconjunto de datos MS refinado, obtenido al eliminar del conjunto MS aquellos compuestos calificados como sustratos de los transportadores ABC con mayores niveles de expresión en la BHE, es decir, de P-gp y/o BCRP. El resultado gráfico del análisis por PCA de este nuevo subconjunto se presenta en la Figura 5.8, en la cual se observa que, al igual que en los casos anteriores, la amplia superposición entre ambos grupos continúa representando un desafío para el desarrollo de modelos predictivos del parámetro $K_{p,uu}$.

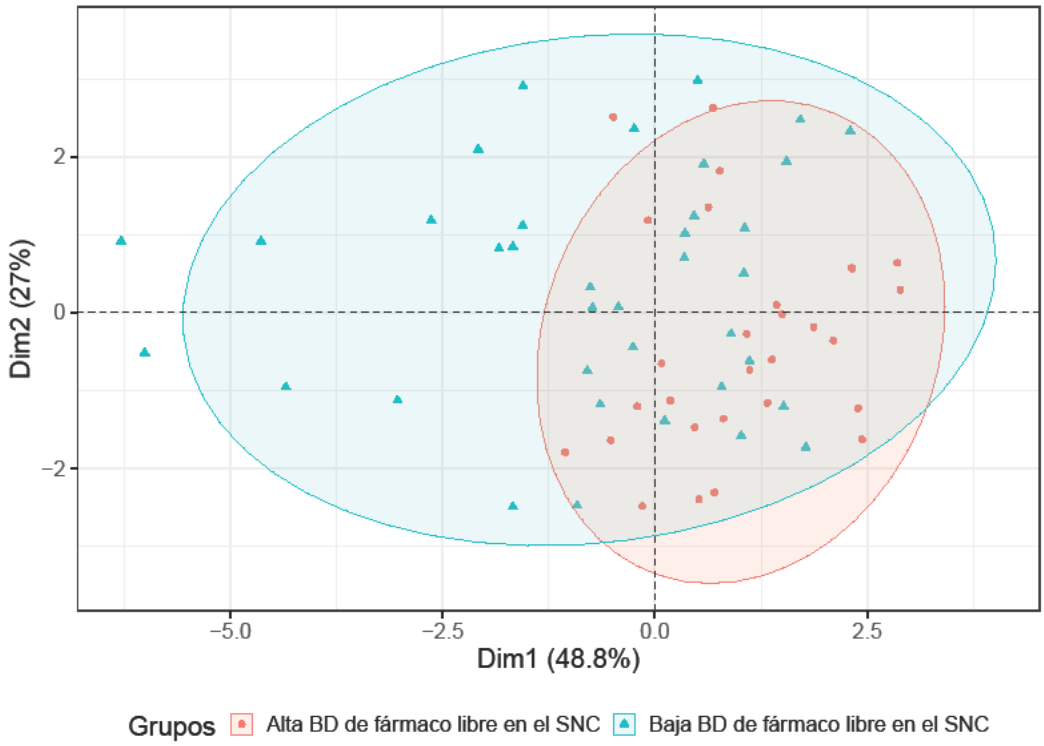


Figura 5.8. PCA del conjunto de datos MS refinado (67 compuestos) basado en ocho descriptores fisicoquímicos (MW; TPSA (Tot); MLogP; nHDon; nHAcc; RBN; nCIC y Sv). Los puntos de datos están coloreados por grupo (triángulos verdes y círculos rojos para baja y alta BD en el SNC, respectivamente). Las elipses dibujadas corresponden a los intervalos del 90% de confianza asumiendo una distribución normal multivariada de los datos por grupo.

En cuanto a la partición del conjunto MS refinado, se procedió de igual manera que en las secciones previas para lograr un conjunto de entrenamiento equilibrado en términos del contenido de compuestos con alta y baja BD de fármaco libre en el SNC. Por ello, el 65% de cada grupo de los compuestos de alta BD libre en el SNC y el 50% de cada grupo de compuestos de baja BD libre en el SNC se asignaron al conjunto de entrenamiento, lo que resultó en igual número de compuestos de alta y baja BD del fármaco libre en el SNC en ese conjunto; los compuestos remanentes se usaron como conjunto de prueba. El conjunto de entrenamiento resultante consistió en 38 compuestos (19 con alta y 19 con baja BD libre en el SNC), mientras que el conjunto de prueba de 29 compuestos incluyó 10 compuestos con alta BD en el SNC y 19 compuestos con baja BD en el SNC (Figura 5.1, C).

Los algoritmos de mejor desempeño difirieron de aquellos que se comportaron mejor en el conjunto de datos MS completo, como se puede ver en la Tabla 5.5. Esto era de

esperarse debido al hecho de que la eliminación de los sustratos de transportadores modifica en gran medida las características del conjunto de datos. Según su desempeño en el conjunto de prueba, los mejores algoritmos fueron cPLS y RF, que arrojaron resultados idénticos de *Acc* (86,2%), *MCC* (0,716), *Se* (0,900) y *Sp* (0,842), y valores de ABC_ROC de 0,863 y 0,890, respectivamente, sin diferencias estadísticamente significativas (p-valor = 0,7851). Se podría pensar que este cambio en los algoritmos con mejor desempeño pueda deberse a que el conjunto de datos MS refinado es de menor tamaño que los anteriores, y por otro lado contemplaría únicamente la difusión pasiva como mecanismo de transporte a través de la BHE, pudiendo así recurrir a técnicas de modelado más sencillas para captar la variabilidad dentro del set de datos.

Tabla 5.5. Resultados de los modelos obtenidos por los diferentes algoritmos en el conjunto de datos MS refinado. Se presentan los resultados tanto para el conjunto de entrenamiento como para el conjunto de prueba. Los mejores modelos están resaltados en negrita. (*) *Acc* de los modelos individuales del ensamblado.

Algoritmo	**Conjunto de entrenamiento**			**Conjunto de prueba**				
	ABC_ROC	**Acc**	**Acc promedio (validación cruzada de 500 iteraciones)**	**ABC_ROC**	**Acc**	**MCC**	**Se**	**Sp**
SVM	0,975	94,7	70,6	0,821	75,9	0,569	0,800	0,789
sGBM	1,000	100,0	72,8	0,911	82,8	0,701	1,000	0,737
*k*NN	0,868	78,9	67,3	0,771	72,4	0,508	0,900	0,632
cPLS	**0,983**	**97,4**	**78,1**	**0,863**	**86,2**	**0,716**	**0,900**	**0,842**
RF	**1,000**	**100,0**	**73,5**	**0,890**	**86,2**	**0,716**	**0,900**	**0,842**
XGBOOST	1,000	100,0	72,0	0,890	86,2	0,688	0,700	0,947
DNN	1,000	100,0	70,2	0,795	82,8	0,611	0,700	0,895
LDA – mejor modelo individual	1,000	100,0	97,1	0,753	79,3	0,525	0,500	0,947
LDA – Ensamblado de 2 mejores modelos indiv. *(operador mínimo)*	1,000	100,0	(88,5; 97,1)*	0,761	72,4	0,461	0,800	0,684
LDA – Ensamblado de 2 mejores modelos indiv. (*operador promedio)*	1,000	100,0	(88,5; 97,1)*	0,816	79,3	0,569	0,800	0,789

5.3. Discusión

La validación prospectiva de los modelos de clasificación desarrollados en la presente tesis pone en evidencia el poder predictivo de los mismos. Es probable que varios factores hayan contribuido a este desempeño. Entre ellos, podemos citar la diversidad de algoritmos de aprendizaje automático ensayados para encontrar los patrones de correlación entre los datos, así como la gran cantidad de descriptores moleculares utilizados para describir el conjunto de datos. Sin embargo, creemos que el cuidadoso curado de las bases de datos fue el factor clave para obtener modelos QSPR precisos. De hecho, el concepto de "basura entra – basura sale" (del inglés, *garbage in – garbage out*) es cada vez más importante en el área de modelado QSAR/QSPR, refiriéndose a que la calidad del modelo resultante depende de la calidad de los datos seleccionados, independientemente de la técnica de modelado (Nantasenamat, 2020). Sin embargo, sabemos que a cambio de una mayor calidad de datos se redujo drásticamente el número de compuestos disponibles para utilizar durante el modelado, siendo este punto una de las debilidades de este trabajo. El reducido número de instancias de entrenamiento podría traducirse, a su vez, en un dominio de aplicación más reducido (aunque, como se ve en los mapas de calor, los conjuntos finales parecen exhibir una notable diversidad molecular). El modesto número de compuestos de los conjuntos de datos pudo ser el causante de varios de los aspectos que se analizarán a continuación.

Al ver los valores de *Acc* tanto en el conjunto de entrenamiento como en el conjunto de prueba para todas las bases de datos (primeras tres columnas -grises- en Tabla 5.6), podemos ver que usualmente la *Acc* en el conjunto utilizado para calibrar el modelo es mayor, excepto para el caso de *k*NN en los conjuntos MSH y MS. Si bien siempre se esperan mejores resultados en el conjunto de datos con el cual se ajusta el modelo (es decir, en el conjunto de entrenamiento), la marcada diferencia en el valor de *Acc* entre ambos conjuntos podría indicar que durante el proceso de modelado se realizó un perceptible sobreajuste, a pesar de haber aplicado diferentes técnicas para evitarlo, como se describió en el capítulo 3. En el caso del conjunto MSH, estas diferencias se encontraron en el rango [-4,8; 21,7]. Como ya se ha comentado previamente, el valor negativo corresponde al algoritmo *k*NN. Esto significaría que dicha técnica no presentó sobreajuste, sin embargo, se puede

observar que el desempeño de *k*NN en el conjunto de prueba fue bastante modesto. Esto podría llevar a pensar que, a pesar de evitar el sobreajuste, este algoritmo no alcanza a incorporar toda la información generalizable de los compuestos que componen el conjunto de entrenamiento.

Tabla 5.6. Diferencias encontradas en los valores de *Acc* (conjunto de entrenamiento – conjunto de prueba y promedio validación cruzada – conjunto de prueba) en los tres dataset considerados.

Algoritmo	**Acc conj de entrenamiento - Acc conj prueba**			**Acc prom validación cruzada – Acc conj de prueba**		
	MSH	**MS**	**MS Ref**	**MSH**	**MS**	**MS Ref**
SVM	5,8	0,2	18,8	-9,3	-18,2	-5,3
sGBM	18,2	8,2	17,2	-8,0	-18,1	-10,0
*k*NN	-4,8	-6,6	6,5	-7,4	-12,5	-5,1
cPLS	15,3	14,6	11,2	-8,9	-7,7	-8,1
RF	17,4	10,6	13,8	-9,7	-22,0	-12,7
XGBOOST	14,9	3,2	13,8	-10,0	-17,7	-14,2
DNN	21,7	8,5	17,2	-8,0	-23,7	-12,6
LDA – mejor modelo individual	12,2	11,3	20,7	-13,0	-17,1	17,8
LDA – Ensamblado mejores modelos indiv. *(operador mínimo)*	14,3	25,4	27,6	(-7,9; -6,0)	(-9,6; -4,6)	(16,1; 24,7)
LDA – Ensamblado mejores modelos indiv. (*operador promedio)*	11,5	20,7	20,7	(-21,4;-20,3;-16,1;-14,3; -12,5)	(-1,3; 3,7)	(9,2; 17,8)

Por otro lado, una técnica menos flexible, como lo es LDA, también evidenció sobreajuste en todos los conjuntos de datos. Aquí es donde se podría pensar que este comportamiento pudo haber sido independiente de la técnica de modelado. Siguiendo esta línea de pensamiento, tendríamos que pensar en aquellos aspectos o etapas que fueron idénticas para todos los algoritmos. Por ejemplo, la metodología de selección del punto de corte del *score* para todos los modelos fue de manera tal de incrementar el *MCC*, manteniendo el balance de *Se* y *Sp*, lo cual, en conjuntos de

datos no balanceados, como fueron los conjuntos de prueba de las diferentes bases de datos, puede no corresponder con el valor máximo de *Acc.*

Por otro lado, en el caso de los conjuntos de entrenamiento, donde había el mismo número de compuestos de ambas clases, el punto de corte para el *score* donde se maximizaba el *MCC* se correspondía con el mayor valor de *Acc*, pudiendo ser una manera de explicar la tendencia a obtener bajos valores de *Acc* en el conjunto de prueba (vs. el conjunto de entrenamiento).

Continuando el análisis de las diferencias, pero en este caso para el conjunto MS, se observa que el rango de las mismas es [-6,6; 25,4], el cual es aún más amplio que para el conjunto MSH. Sin embargo, si analizamos el rango, pero sacando los modelos desarrollados por la metodología *in-house* LDA, podemos ver que el mismo se reduce a [-6,6; 14,6]. Esto demuestra una mejora con respecto al conjunto MSH (cabe aclarar que el rango de diferencias para el conjunto MSH removiendo los modelos *in-house* LDA mantiene los mismos valores). Se podría pensar que esta mejora se debe al proceso de refinado de la base de datos, donde se eliminaron los valores de $K_{p,uu}$ obtenidos por una técnica de menor precisión. Esto pudo haber ayudado a disminuir el número de compuestos que generaban inestabilidad a los modelos (datos influyentes).

Sin embargo, si analizamos el rango para los modelos LDA, vemos que el mismo es [11,5; 14,3] para el conjunto MSH y [11,3; 25,4] para el MS, indicando un incremento del sobreajuste en el refinamiento del *dataset.* Una potencial explicación a este comportamiento puede encontrarse en cómo fueron elegidos los mejores modelos para cada técnica de modelado. Para los modelos desarrollados por la metodología *in-house* LDA, se determinó qué modelos eran los mejores utilizando el valor del área bajo la curva ROC en el conjunto de entrenamiento (cuanto mayor, mejor), mientras que para el resto de los algoritmos el mejor modelo se seleccionó en base a los resultados de la validación cruzada. Este hecho podría explicar el comportamiento de los modelos LDA, ya que esta manera de seleccionar el mejor modelo pondera en exceso al conjunto de entrenamiento, a la vez que algunos estudios sugieren que las métricas derivadas de la curva ROC presentan alta variabilidad para conjuntos de datos pequeños (Truchon et al., 2007).

Si continuamos el análisis previo, el rango de diferencias de *Acc* entre el conjunto de entrenamiento y el conjunto de prueba para los modelos no lineales -por un lado- e *in-house* LDA -por el otro- en el conjunto MS refinado, nos encontramos con los valores [6,5; 18,8] y [20,7; 27,6], respectivamente. La tendencia a mayor sobreajuste de los modelos LDA se mantiene y, comparando el rango de diferencias con los obtenidos para el conjunto MS para los modelos no lineales, incluso empeora, obteniendo nuevamente un mayor sobreajuste similar al que se describió previamente con el conjunto MSH. Posiblemente, la reducción del número de datos (de 109 a 67) resultó perjudicial desde este punto de vista. Como hemos analizado, el sobreajuste es un gran problema cuando se desarrollan modelos computacionales, particularmente relevante en metodologías de modelado muy flexibles, aplicadas a *datasets* con reducido número de compuestos (ya que contar con mayor número de instancias de entrenamiento reduce el riesgo de sobreajuste), incluso cuando se implementaron distintas estrategias para evitarlo, específicamente, vigilancia de la relación entre el número de casos de entrenamiento y de descriptores incorporados al modelo, y partición representativa de los conjuntos de datos.

Por otro lado, algo muy interesante para remarcar en los resultados de modelado de todos los conjuntos de datos es la tendencia a obtener valores de *Acc* en los conjuntos de prueba mayores a la *Acc* promedio obtenida por la validación cruzada (columnas a la derecha en Tabla 5.6). Esto llama la atención, dado que por lo general se esperarían valores similares, en tanto los resultados de la validación cruzada son predictivos de cómo será el desempeño en el conjunto de prueba, y en casos de discrepancia la validación cruzada tiende a ser optimista, esto es, a sugerir una capacidad predictiva mayor que la que se observa luego en el conjunto de prueba (Tropsha et al., 2003). Sin embargo, habría que remarcar que durante la validación cruzada los compuestos fueron particionados aleatoriamente, lo cual es un proceso diferente al realizado durante el armado del conjunto de prueba, en el que se utilizó una técnica de partición racional que, como se comentó en el capítulo 3, es más conveniente para el caso de base de datos pequeñas. Un punto a tener en cuenta es que los modelos basados en métodos de división racional del conjunto de datos tienden a generar mejores resultados estadísticos para el conjunto de prueba que los modelos basados en la división aleatoria (Martin et al., 2012). Al contrario, la

selección aleatoria funciona bien para conjuntos de datos grandes, pero provoca una significativa inestabilidad en los resultados para conjuntos de datos pequeños (Pirhadi et al., 2015). Esto está en concordancia con la repetición del proceso de las validaciones cruzadas, la cual tiene como objetivo robustecer las conclusiones y evitar un resultado sesgado por la técnica de muestreo. Entonces, cuando los conjuntos de datos son pequeños, la selección racional de los conjuntos de entrenamiento y de prueba se vuelve importante porque las diferentes formas de muestrear los conjuntos de datos pueden llevar a conclusiones diferentes sobre el poder predictivo y la solidez de los modelos (Hongmao, 2016).

Dado que dicha tendencia se observa en todos los conjuntos de datos por igual, independientemente de la composición y tamaño de los mismos (a pesar de que se podría pensar que cada conjunto de datos es un subconjunto del otro y que todos son de un tamaño modesto), se podría pensar que su origen radica en alguna etapa o elemento común a todos ellos (por ejemplo, alguna instancia de entrenamiento particularmente influyente, que al ser removida del conjunto de entrenamiento durante la validación cruzada es mal clasificada de manera sistemática). Los rangos de diferencias (*Acc* promedio de la validación cruzada – *Acc* conjunto de prueba) para los conjuntos de datos MSH, MS y MS refinado, son [-21,4; -6,0], [-23,7; 3,7] y [-14,2; 24,7], respectivamente (columnas a la derecha en Tabla 5.6). Idealmente, mientras más cercana a cero la diferencia, mejor. Si la diferencia es negativa significa que el proceso de validación cruzada subestimó la capacidad predictiva del modelo, y viceversa.

Como era de esperarse, el rango más estrecho pertenece al conjunto con mayor número de compuestos (MSH), mientras que el rango más amplio o de mayor variabilidad corresponde al conjunto MS refinado, de menor tamaño. Esta tendencia es lógica, ya que a menor número de compuestos la variabilidad resultante por la mala clasificación de uno de ellos es mayor.

Si los rangos antes mencionados se analizan sin tener en cuenta a los modelos LDA, los mismos se convierten en [-10,0; -7,4], [-23,7; -7,7] y [-14,2; -5,1] para los conjuntos MSH, MS y MS refinado, respectivamente. El signo negativo en todos los casos indica una tendencia sistemática a *subestimar* la capacidad predictiva de los modelos mediante validación cruzada. Como se comentó anteriormente, esto podría

deberse, al menos parcialmente, a la utilización de diferentes técnicas para la partición de los compuestos durante la formación del conjunto de entrenamiento y el conjunto de prueba, y durante la validación cruzada, o a que algunos datos excesivamente influyentes ("puntos palanca") hayan quedado localizados en los conjuntos de entrenamiento.

Se puede también observar que el rango más amplio corresponde al conjunto MS. Si observamos los valores de *Acc* promedio de la validación cruzada para todos los conjuntos de datos vemos que para los modelos no lineales suelen encontrarse alrededor de 70,0%, por lo que las variaciones entre los rangos de diferencias (*Acc* validación cruzada – *Acc* conjunto de prueba) entre conjuntos se deben mayormente a variaciones en los valores de *Acc* en los conjuntos de prueba; en el conjunto MS se obtuvieron los mejores resultados de modelado según el conjunto de prueba. En otras palabras, se trata de los modelos con mejor capacidad predictiva, y mayor subestimación por parte de la validación cruzada.

El mismo análisis, pero con los modelos desarrollados por la técnica *in-house* LDA, revela que los rangos de las diferencias (*Acc* validación cruzada – *Acc* conjunto de prueba) son [-21,4; -6,0], [-17,1; 3,7] y [9,2; 24,7] para los conjuntos MSH, MS y MS refinado, respectivamente.

Si observamos los resultados en el conjunto de prueba, los cuales representarían la capacidad predictiva de los modelos de forma más precisa, podemos ver que en el conjunto de datos MS fue donde los modelos de mejor desempeño presentaron, a su vez, la mejor performance, con valores de *Acc* y *MCC* mayores a 90,0% y 0,800, respectivamente. Creemos que este conjunto de datos representó el mejor compromiso entre calidad de los datos/número de instancias, es decir que la disminución del "ruido" (respecto al conjunto MSH) fue mayor que la pérdida de información por el menor número de compuestos debido al proceso de exclusión (respecto al conjunto MS refinado).

Por último, realizaremos una discusión de los **descriptores moleculares** seleccionados por los mejores modelos. En la Tabla 5.7 pueden observarse los 5 descriptores moleculares más relevantes para el mejor modelo en cada uno de los tres conjuntos de datos. En la Tabla 5.8 se presenta la definición de cada uno de los descriptores moleculares seleccionados por su importancia. Tres de los cinco

descriptores moleculares más importantes para el mejor modelo del conjunto MSH también fueron relevantes para el conjunto MS (MATS2m, P_VSA_LogP_4, ATSC1e). Sin embargo, cuando comparamos entre los descriptores moleculares más importantes para el conjunto MSH y el conjunto MS refinado, solo uno de los descriptores se repite (IC1), y ninguno cuando se realiza la comparación entre el conjunto MS y MS refinado. Este comportamiento de los descriptores moleculares seleccionados como importantes podría también visualizarse en los gráficos de PCA, los cuales son bastante similares para los conjuntos de datos MSH y MS (Figuras 5.3 y 5.6), pero para el caso del conjunto MS refinado (Figura 5.8) su forma es diferente. Era de esperarse que los descriptores moleculares más relevantes para el conjunto MS refinado difieran de los descriptores seleccionados por los otros conjuntos debido al criterio de exclusión que se utilizó (remoción de sustratos de P-gp y BCRP), lo cual generó un conjunto de datos que prioriza los compuestos que no tienen interacción con transportadores de membrana.

Tabla 5.7. Importancia de los descriptores moleculares según el mejor modelo para cada conjunto de datos.

Ranking de importancia	Conjunto MSH	Conjunto MS	Conjunto MS refinado (RF)
1	MATS2m	P_VSA_LogP_4	AMW
2	P_VSA_LogP_4	ATSC1e	IC0
3	ATSC1e	TPSA(Tot)	AAC
4	IC1	MATS2m	SpPosA_B(m)
5	P_VSA_s_6	ATSC2s	IC1

Los descriptores moleculares más importantes para los conjuntos MSH y MS están relacionados con la distribución de heteroátomos en la molécula (MATS2m), la polaridad (ATSC1e; TPSA(Tot)), la complejidad (IC1) y el estado intrínseco (P_VSA_s_6; ATSC2s) de los compuestos. El estado intrínseco se puede considerar como la relación de electrones π y pares de electrones libres y el total de enlaces sigma (σ) en el gráfico molecular para el átomo considerado. Esta propiedad refleja la influencia de electrones no-σ a lo largo de los caminos topológicos que parten del átomo considerado. Por lo tanto, cuanto menor sea la partición de la influencia de los electrones a lo largo de los caminos topológicos, más disponibles estarán los

electrones de valencia para la interacción intermolecular. En el caso del conjunto de datos MS refinado los descriptores moleculares más relevantes están relacionados con la masa (AMW; SpPosA_B(m)), la complejidad (IC0; IC1) y la diversidad atómica (ACC) de los compuestos. Puede observarse que tanto la polaridad y el estado intrínseco de los compuestos son características que aparecen como relevantes en el caso del conjunto de datos MS refinado, pero si para los conjuntos MSH y MS. Esto llevaría a pensar que estas dos características son influyentes cuando hay interacción con transportadores de membrana, indicando que se producen mayormente a través de interacciones de cargas moleculares.

Tabla 5.8. Definición de los descriptores moleculares seleccionados por los mejores modelos.

Descriptor	**Definición**
AAC	Índice promedio de información sobre la composición atómica
AMW	Peso molecular promedio
ATSC1e	Autocorrelación centrada de Broto-Moreau. Propiedad: electronegatividad de Sanderson. Distancia topológica: 1.
ATSC2s	Autocorrelación centrada de Broto-Moreau. Propiedad: estado intrínseco. Distancia topológica: 2
IC0	Índice de contenido de información (simetría de vecindad de orden 0)
IC1	Índice de contenido de información (simetría de vecindad de orden 1)
MATS2m	Autocorrelación de Moran. Propiedad: masa. Distancia topológica: 2
P_VSA_LogP_4	Descriptor tipo P_VSA. Propiedad: LogP. Rango: 4
P_VSA_s_6	Descriptor tipo P_VSA. Propiedad: estado intrínseco. Rango: 6
SpPosA_B(m)	Suma positiva espectral normalizada de la matriz de carga ponderada por masa
TPSA(Tot)	Área de superficie polar topológica considerando las contribuciones de N, O, S y P.

Los descriptores de autocorrelación (AC) se basan en la función del mismo nombre, la cual, para una secuencia ordenada de *n* valores de una función dada *f(x)*, se calcula como:

$$AC_d = \sum_{i=1}^{n-d} f(x_i) \cdot f(x_{i+d}) \quad (5.1)$$

Donde *d* es la distancia topológica entre dos átomos dados (la cual determina el "orden" de la autocorrelación: 0, 1, 2…), *n* es el número de átomos del grafo[2] y *f(x)* es la contribución de cada átomo a una propiedad fisicoquímica dada. Entonces, si por ejemplo *d*=2, y la propiedad de interés es la electronegatividad, el descriptor se calcula como la suma de los productos de la contribución atómica a la electronegatividad de todos los pares de átomos de la molécula separados por una distancia topológica igual a 2. Si bien las correlaciones de Broto- Moreau (ATSC) y de Moran (MATS) presentan otras particularidades de cálculo que no discutiremos aquí, ambas se basan en la función AC_d descripta en la ecuación (5.1), y describen la distribución espacial de la propiedad molecular considerada, a lo largo de la estructura topológica (Mauri et al., 2017).

Los descriptores tipo P_VSA, por su parte, cuantifican o dan idea de la cantidad de área superficial de van der Waals (VSA) con una propiedad fisicoquímica dada (*P*) dentro de cierto rango *k* de valores. Es decir:

$$P\_VSA_k = \sum_{i=1}^{n} V_i \cdot \delta(P_i \in [a_{k-1}; a_k]) \qquad k = 1, 2, 3 \dots K \tag{5.2}$$

Donde V_i representa la contribución del átomo *i* al VSA de la molécula, y $\delta(P \dots)$ es una función delta generalizada, la cual adopta el valor 1 si se satisface la condición $(P_i \in [a_{k-1}; a_k])$, o el valor 0, en caso contrario. Acorde a los autores que introdujeron su uso, este tipo de descriptores es útil no sólo para el modelado de propiedades físicas, sino también para el modelado de afinidad de un compuesto por su receptor (Labute, 2000).

Los índices de información, por último, combinan la teoría de grafos con la teoría de la información estadística de Shannon, según la cual la cantidad de información se define en términos probabilísticos. El grado de incertidumbre de un resultado determinado *i* se expresa mediante su entropía H_i, la cual es una función de la probabilidad p_i de dicho resultado:

$$H_i = -\log_2 p_i \tag{5.3}$$

[2] El grafo es la representación 2D de la molécula, en la cual los átomos se toman como vértices, y los enlaces covalentes como ejes que conectan los vértices. La distancia topológica es el número de ejes que separan dos vértices dados.

Cuando el resultado carece de incertidumbre (es decir, cuando está totalmente determinado), $p_i = 1$ y la entropía vale cero. Por el contrario, la total incertidumbre ($p_i = 0$) se corresponde con un valor de entropía infinito (Bonchev, 1983).

Para su aplicación al cálculo de descriptores, la ecuación anterior se expresa en términos de entropía media H(P) de la distribución de probabilidad P= {$p_1, p_2, \ldots, p_k$} de todos los posibles resultados, es decir:

$$H(p) = -\sum_{i=1}^{k} p_i \cdot \log_2 p_i \qquad (5.4)$$

Existen múltiples índices de información topológicos, los cuales se basan en la ecuación anterior, pero difieren en términos del grafo considerado (con hidrógenos explícitos o implícitos), si se calculan para vértices o enlaces, si tienen en cuenta o no la naturaleza química de los átomos, de la distancia topológica considerada para evaluar la equivalencia, etc.

Los índices de contenido de información (IC_d), por ejemplo, se basan en evaluar la simetría de las vecindades o entornos de orden *d*, para cada vértice considerado. Por ejemplo, para el primer compuesto presentado en la Tabla 5.9, se tiene que el conjunto de vecindades o subgrafos de primer orden corresponde a los especificados en la Fig. 5.9.

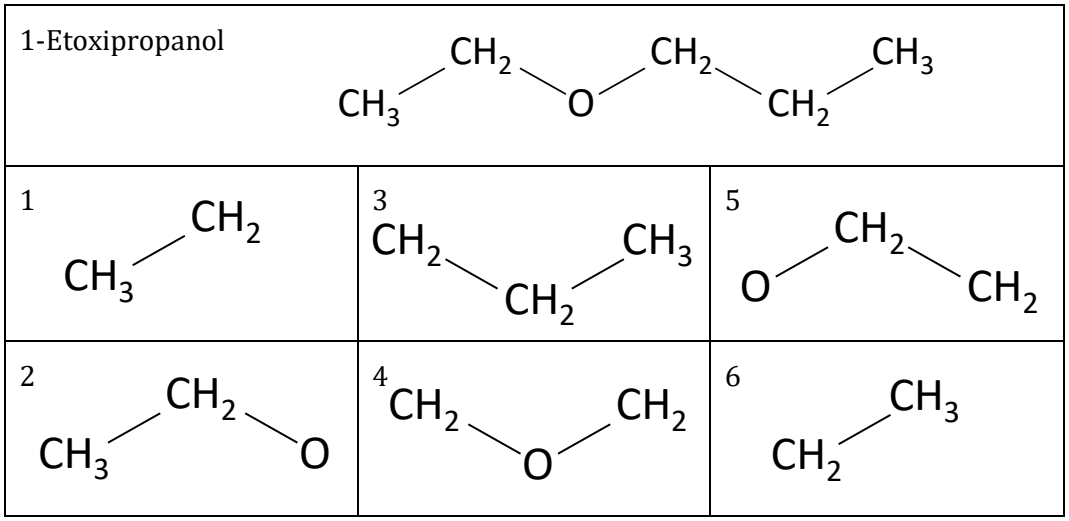


Figura 5.9. Subgrafos de orden 1 (distancia topológica *d* = 1) del 1-Etoxipropanol.

Considerando que las subestructuras 1 y 6 son iguales, el cálculo del índice de contenido de información de orden 1 (IC_1) resulta, en este caso:

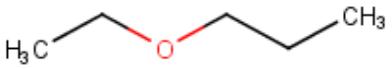
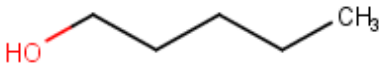
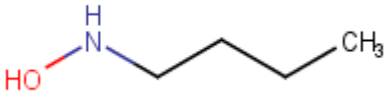
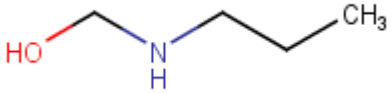
$$IC_1 = -\left\{\frac{2}{6}\log_2\left(\frac{2}{6}\right) + 4 \cdot \left[\frac{1}{6}\log_2\left(\frac{1}{6}\right)\right]\right\} = 2{,}25 \qquad (5.5)$$

donde p_i de la ecuación 4 se ha reemplazado aquí por la probabilidad de encontrar al subgrafo de tipo i, extrayendo un elemento al azar del conjunto de subgrafos del orden considerado. No es difícil demostrar que repitiendo el cálculo para el segundo compuesto de la Tabla 5.9 (1-Pentanol), se llega al mismo resultado.

Este ejemplo ilustra un aspecto importante de los índices de información, los cuales, si bien son particularmente útiles para cuantificar el grado de redundancia y heterogeneidad en las inmediaciones de algún elemento del grafo (vértices, enlaces), no son particularmente sensibles a la distribución de los heteroátomos dentro de la molécula.

Por lo tanto, el cambio en el patrón de descriptores relevantes (desaparición de autocorrelaciones 2D y descriptores tipo P_VSA, junto con el enriquecimiento en índices de información) que se observa al pasar de los conjuntos de datos MSH y MS al MS refinado parecería indicar que la disposición de los átomos adquiere mayor relevancia cuando se tiene en cuenta la interacción con los transportadores de la BHE, lo cual tiene sentido pensando que tal interacción depende de eventos de reconocimiento específico, por lo cual vale aquí la noción de farmacóforo, que habitualmente comprende las distancias y disposiciones relativas entre átomos que pueden participar de interacciones de tipo puente de hidrógeno y/o de tipo electrostático.

Tabla 5.9. Se presentan dos pares de compuestos, que sólo difieren en la posición relativa de un heteroátomo (oxígeno en el primer caso, nitrógeno en el segundo). Se observa que mientras que los índices de información (AAC, IC0 e IC1) permanecen invariantes, las autocorrelaciones 2D (ATS1e, ATS2s y MATS2m) y los descriptores de tipo P_VSA son capaces de detectar e informar sobre las diferencias estructurales de ambas series.

Nombre	Estructura	AMW	TPSA (Tot)	AAC	IC0	IC1	ATS 1e	ATS 2s	MATS 2m	P_VSA_ s_6	P_VSA_ LogP_4
1-Etoxipropanol		12,7	9,23	0,65	0,65	2,25	1.895	3.829	-0,40	77,6	NA
1- Pentanol		12,7	17,07	0,65	0,65	2,25	1.845	3.773	-0,10	59,9	NA
N-Butilhidroxilamina		13,0	31,17	1,25	1,25	2,58	1.902	3.801	-0,21	80,3	NA
(Propilamino)Metanol		13,0	31,17	1,25	1,25	2,58	1.894	3.876	0,43	78,0	NA

Referencias

Bonchev, D. (1983). *Information theoretic indices for characterization of chemical structures*. Research Studies Press, Chichester, West Sussex, UK.

Chen, H., Winiwarter, S., Fridén, M., Antonsson, M., & Engkvist, O. (2011). In silico prediction of unbound brain-to-plasma concentration ratio using machine learning algorithms. *Journal of Molecular Graphics and Modelling*, *29*(8), 985–995. https://doi.org/10.1016/j.jmgm.2011.04.004

DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. *Biometrics*, *44*(3), 837-845. https://doi.org/10.2307/2531595

Doan, K. M., Wring, S. A., Shampine, L. J., Jordan, K. H., Bishop, J. P., Kratz, J., et al. (2004). Steady-State Brain Concentrations of Antihistamines in Rats. *Pharmacology*, *72*(2), 92–98. https://doi.org/10.1159/000079137

Hongmao, S. (2016). *Practical Guide to Rational Drug Design*. Woodhead Publishing, Sawston, UK. https://doi.org/10.1016/c2014-0-02348-9

Kanamitsu, K., Nozaki, Y., Nagaya, Y., Sugiyama, Y., & Kusuhara, H. (2017). Quantitative prediction of histamine H1 receptor occupancy by the sedative and non-sedative antagonists in the human central nervous system based on systemic exposure and preclinical data. *Drug Metabolism and Pharmacokinetics*, *32*(2), 135–144. https://doi.org/10.1016/j.dmpk.2016.11.007

Labute, P. (2000). A widely applicable set of descriptors. *Journal of Molecular Graphics and Modelling*, *18*(4–5), 464–477. https://doi.org/10.1016/S1093-3263(00)00068-1

Martin, T. M., Harten, P., Young, D. M., Muratov, E. N., Golbraikh, A., Zhu, H., & Tropsha, A. (2012). Does Rational Selection of Training and Test Sets Improve the Outcome of QSAR Modeling? *Journal of Chemical Information and Modeling*, *52*(10), 2570–2578. https://doi.org/10.1021/ci300338w

Mauri, A., Consonni, V., & Todeschini, R. (2017). Molecular descriptors. In: Leszczynski J., Kaczmarek-Kedziera A., Puzyn T., G. Papadopoulos M., Reis H., K. Shukla M. (Eds.) *Handbook of Computational Chemistry*. Springer, Cham. https://doi.org/10.1007/978-3-319-27282-5_51

Nakazono, T., Murakami, T., Higashi, Y., & Yata, N. (1991). Study on Brain Uptake of Local Anesthetics in Rats. *Journal of Pharmacobio-Dynamics*, *14*(11), 605–613. https://doi.org/10.1248/bpb1978.14.605

Nantasenamat, C. (2020). Best practices for constructing reproducible QSAR models. In: Roy K. (Ed.) *Ecotoxicological QSARs. Methods in Pharmacology and Toxicology*. Humana, New York, NY. https://doi.org/10.1007/978-1-0716-0150-1_3

Pirhadi, S., Shiri, F., & Ghasemi, J. B. (2015). Multivariate statistical analysis methods in QSAR. *RSC Advances*, *5*, 104635–104665. https://doi.org/10.1039/c5ra10729f

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J. C., & Müller, M. (2011). pROC: An open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, *12*, 77. https://doi.org/10.1186/1471-2105-12-77

Tamai, I., & Tsuji, A. (2000). Transporter-mediated permeation of drugs across the blood-brain barrier. *Journal of Pharmaceutical Sciences*, *89*(11), 1371–1388. https://doi.org/10.1002/1520-6017(200011)89:11<1371::AID-JPS1>3.0.CO;2-D

Tropsha, A., Gramatica, P., & Gombar, V. K. (2003). The importance of being earnest: Validation is the absolute essential for successful application and interpretation of QSPR models. *QSAR and Combinatorial Science*, *22*(1), 69–77. https://doi.org/10.1002/qsar.200390007

Truchon, J. F., & Bayly, C. I. (2007). Evaluating virtual screening methods: Good and bad metrics for the “early recognition” problem. *Journal of Chemical Information and Modeling*, *47*(2), 488–508. https://doi.org/10.1021/ci600426e

Yasuhara, M., & Levy, G. (1988). Kinetics of drug action in disease states XXVI: Effect of fever on the pharmacodynamics of theophylline-induced seizures in rats. *Journal of Pharmaceutical Sciences*, *77*(7), 569–570. https://doi.org/10.1002/jps.2600770704

Young, R. C., Mitchell, R. C., Brown, T. H., Griffiths, R., Jones, M., Rana, K. K., … Ganellin, C. R. (1988). Development of a New Physicochemical Model for Brain Penetration and Its Application to the Design of Centrally Acting H2receptor Histamine Antagonists. *Journal of Medicinal Chemistry*, *31*(3), 656–671. https://doi.org/10.1021/jm00398a028

Yusof, S. R., Mohd Uzid, M., Teh, E. H., Hanapi, N. A., Mohideen, M., Mohamad Arshad, A. S., … Hammarlund-Udenaes, M. (2019). Rate and extent of mitragynine and 7-hydroxymitragynine blood–brain barrier transport and their intra-brain distribution: the missing link in pharmacodynamic studies. *Addiction Biology*, *24*(5), 935–945. https://doi.org/10.1111/adb.12661

Capítulo 6

Conclusiones

En el presente trabajo de tesis se obtuvieron modelos computacionales con el objetivo de poder predecir, en función de la estructura molecular de un compuesto químico, su parámetro farmacocinético $K_{p,uu}$. Dicho parámetro es aceptado hoy en día como el de mayor biorrelevancia para la evaluación de la biodisponibilidad de un fármaco en el SNC.

El proceso para la generación de los modelos se estructuró en dos etapas bien diferenciadas. En una primera instancia, se compiló, a partir de fuentes bibliográficas, un conjunto de datos de compuestos con valores de $K_{p,uu}$ obtenidos en estado estacionario por cualquiera de las tres metodologías experimentales disponibles para dicho fin: Microdialisis, S*lice* y Homogenato (conjunto de datos MSH). Dicha base de datos MSH se utilizó para generar modelos clasificatorios mediante el uso de distintos algoritmos, algunos de los cuales lograron discernir adecuadamente entre compuestos de baja y alta biodisponibilidad de fármaco libre en el SNC. Todos los modelos fueron validados computacionalmente y demostraron un buen poder predictivo. Se validó también experimental y prospectivamente el modelo individual seleccionado como el mejor en base a los resultados de la

validación *in silico* (modelo XGBOOST). Para ello se determinó experimentalmente el parámetro $K_{p,uu}$ mediante la técnica de homogenato de cinco compuestos que no habían formado parte del conjunto de datos original. Los cinco nuevos compuestos demostraron pertenecer al dominio de aplicación del mejor modelo y, entre ellos, tres eran de baja biodisponibilidad de fármaco libre en el SNC (Teofilina, Ranitidina y Ácido p-aminobenzoico o PABA) y dos de alta biodisponibilidad de fármaco libre en el SNC (Clorfeniramina y Lidocaína). Todos los compuestos fueron bien clasificados por el mejor modelo excepto Teofilina. Dicho compuesto fue predicho como de alta biodisponibilidad de fármaco libre en el SNC, no obstante, su valor experimental de $K_{p,uu}$ (0,119) indica que pertenece a la clase de compuestos de baja biodisponibilidad de fármaco libre en el SNC. A pesar del acotado número de nuevos compuestos utilizados para la validación experimental prospectiva, la *Acc* fue de 80,0% (4/5), similar al obtenido durante la validación computacional con el conjunto de prueba (85,1%).

Con el objetivo de aumentar el poder predictivo del modelo desarrollado, una estrategia que se podría implementar sería aumentar el número de compuestos del conjunto de datos inicial, de manera de tener mayor información sobre la relación entre la estructura molecular y el parámetro a modelar. Otra posible estrategia sería, por el contrario, intentar refinar más el conjunto de datos, removiendo los compuestos que fueron obtenidos por la técnica de homogenato, dado que dicho método suele presentar mayor variabilidad que los demás, la cual se traslada directamente a los modelos obtenidos. En la presente tesis se decidió explorar esta última estrategia para intentar mejorar los modelos obtenidos, construyendo un conjunto de datos más refinado.

Por lo tanto, para la segunda etapa de desarrollo de modelos se consideraron únicamente aquellos valores de $K_{p,uu}$ obtenidos por las técnicas de Microdialisis y/o *Slice.* Se conformó así un nuevo conjunto de datos (conjunto MS), con el objetivo de disminuir la variabilidad/ruido de la base de datos a utilizar en la obtención de los modelos, y mejorar así el poder predictivo de los mismos. Como cabía esperar, la misma metodología (algoritmos clasificatorios) de modelado generó, partiendo de este nuevo conjunto de datos refinado, modelos con mejor poder predictivo que los

obtenidos con el conjunto MSH, en función de los resultados de la validación computacional de los mismos.

A continuación, se seleccionó, de entre todos los modelos generados a partir del conjunto MS, el que mejores resultados obtuvo en la validación computacional (modelo sGBM). En esta oportunidad la validación prospectiva se llevó a cabo desde un enfoque diferente al descripto para el conjunto de datos MSH. Se utilizaron compuestos que no formaron parte del conjunto de datos MS, y cuyos datos observados del parámetro modelado fueron obtenidos de diferentes maneras: datos publicados con posterioridad al armado del conjunto de datos MS y datos proporcionados por la Escuela de Medicina de la Universidad de Indiana (EE. UU.). De esta forma, el conjunto de datos para esta nueva validación prospectiva quedó conformado por 10 compuestos, 6 con alta y 4 con baja biodisponibilidad de fármaco libre en el SNC. De éstos, 9 fueron bien clasificados por el mejor modelo, lo que representó un 90,0% de *Acc*, valor que está en concordancia al obtenido en el conjunto de prueba del mejor modelo (91,8%). El compuesto mal clasificado fue Aripiprazol, el cual fue predicho como de alta biodisponibilidad de fármaco libre en el SNC, cuando los datos proporcionados por la Universidad de Indiana indicaban lo contrario.

Por último, comparando nuestros modelos computacionales con los modelos previamente publicados para predecir el parámetro $K_{p,uu}$, podemos visualizar que el desempeño como clasificador de nuestros modelos es similar o levemente superior. El mejor modelo de clasificación desarrollado hasta el momento de la escritura de esta tesis (que en total fueron seis, como se discutió en el Capítulo 2) es el reportado por Chen *et al.*, el cual presentó, en el conjunto de prueba, un *MCC* de 0,72 y una *Acc* de 85,0. El mismo fue desarrollado utilizando una base de datos de 246 compuestos totales (no disponibles públicamente), 173 en el conjunto de entrenamiento y 73 en el conjunto de prueba. Nuestro mejor modelo (modelo sGBM obtenido en el conjunto de datos MS), presentó valores superiores, con un *MCC* de 0,812 y una *Acc* de 91,8%. En base a esta comparación de resultados podríamos concluir que la metodología que ha sido utilizada para el desarrollo de nuestros modelos *in silico* es al menos competitiva con respecto a la utilizada por los mencionados grupos de investigación, destacándose como puntos positivos el cuidadoso curado de la base

de datos, y la diversidad de descriptores y algoritmos clasificatorios empleados, los cuales nos permitieron obtener resultados superadores aun partiendo de un conjunto de datos más pequeño (109 compuestos en el caso del conjunto MS, 60 en el conjunto de entrenamiento y 49 en el conjunto de prueba).

Sin embargo, debemos mencionar también que el modesto número de compuestos de los conjuntos de datos hacen que, necesariamente, los modelos posean un dominio de aplicación más restringido, el cual deberá ser siempre verificado antes de realizar una predicción. A futuro, la actualización de nuestros modelos computacionales a partir de la ampliación de los conjuntos de datos sería el camino a seguir para mejorar el poder predictivo de los mismos y ampliar su dominio de aplicación.

Nuestros modelos serán incorporados como filtros *in silico* en las etapas tempranas de los proyectos de descubrimiento de nuevos fármacos destinados al SNC en nuestro laboratorio (LIDeB, Facultad de Ciencias Exactas, UNLP). En dichas etapas de desarrollo y optimización de moléculas candidatas, la mejora de las características farmacocinéticas es uno de los pilares esenciales para el éxito futuro de los proyectos de búsqueda de nuevos fármacos. Si, por otra parte, esto se logra mediante el uso de herramientas computacionales, las cuales son económicas y rápidas, se podría lograr un mejor uso de los recursos utilizados para el desarrollo de nuevos medicamentos para el tratamiento de las enfermedades del SNC, pudiendo tal vez disminuir el impacto de la baja en la inversión sufrida en los últimos tiempos.

Anexos

Publicaciones realizadas, becas obtenidas y presentaciones a congresos durante el período en el que se desarrolló la presente tesis doctoral

Durante la realización de este trabajo de tesis se han realizado 29 presentaciones a congresos. También se han generado 8 artículos científicos, 7 de ellos ya publicados en revistas científicas con referato.

En el año 2015 se obtuvo una beca de viaje y estadía de parte de *International Brain Research Organization Latin America Regional Committee* (IBRO LARC) para realizar una estadía corta de investigación de un mes en el Centro de Investigación y de Estudios Avanzados (CINVESTAV) de la Ciudad de México, México; En el año 2016 se obtuvo una beca de viaje, estadía e inscripción de parte de *The World Academy of Sciences for the Developing Countries* (TWAS) para asistir a *Biovision 2016 – The World Life Sciences* en Lyon, Francia; En ese mismo año se obtuvo otra beca de viaje, estadía e inscripción de parte de la *International League Against Epilepsy* (ILAE) para asistir a la *Latin American Summer School on Epilepsy* (LASSE X) en San Pablo, Brasil; En el año 2018 se obtuvo una beca de viaje y estadía de parte de la Comisión

Fulbright Argentina para realizar una estadía corta de investigación de 3 meses en la División de Farmacología Clínica, Departamento de Medicina, Escuela de Medicina de la Universidad de Indiana, Indianápolis, Estados Unidos.

En el año 2018, nuestro trabajo denominado *Application of machine learning for predicting bioavailability of drugs in the CNS* fue premiado en el encuentro anual 2018 del Indiana CTSI celebrado en Indianapolis, Estados Unidos.

A continuación, se detallan los trabajos publicados/enviados durante el transcurso de la tesis doctoral:

1. Current State and Future Perspective in QSAR Models to Predict Blood Brain Barrier Penetration in Central Nervous System Drug R&D
   Morales, JF; Scioli Montoto, S; Fagiolino, P; Ruiz, ME.
   Mini Reviews in Medicinal Chemistry, 17(3): 247 – 257 (2017)
   DOI: 10.2174/1389557516666161013110813

2. Development and validation of a computational model ensemble for the early detection of BCRP/ABCG2 substrates during the drug design stage
   Gantner, ME; Peroni, RN; **Morales, JF;** Villalba, ML; Ruiz, ME; Talevi, A.
   Journal of Chemical Information and Modeling, 57(8): 1868-1880 (2017)
   DOI: 10.1021/acs.jcim.7b00016

3. Cascade Ligand- and Structure-Based Virtual Screening to Identify New Trypanocidal Compounds Inhibiting Putrescine Uptake
   Alberca, LN; Sbaraglini, ML; **Morales, JF;** Dietrich, R; Ruiz, MD; Pino Martínez, AM; Miranda, CG; Fraccaroli, L; Alba Soto, CD; Carrillo, C; Palestro, PH; Talevi, A.
   Frontiers in Cellular and Infection Microbiology, 8, 173 (2018)
   DOI: 10.3389/fcimb.2018.00173

4. Molecular topology and other promiscuity determinants as predictors of therapeutic Class? A theoretical framework to guide drug repositioning
   Morales, JF; Alberca, LN; Di Ianni, ME; Chuguransky, S; Talevi, A; Ruiz, ME.
   Current Topics in Medicinal Chemistry, 8: 1110-1122 (2018)
   DOI: 10.2174/1568026618666180801091642

5. Application of machine learning approaches to identify new anticonvulsant compounds active in the 6 Hz seizure model
   Goicoechea, S; Sbaraglini, ML; Chuguransky, S; **Morales, JF;** Ruiz, ME; Talevi, A; Bellera, CL.
   Communications in Computer and Information Science, 1068: 3-19 (2019).
   DOI: 10.1007/978-3-030-36636-0_1

6. Positivity Predictive Value surfaces as a complementary tool to assess the performance of virtual screening methods
   Morales, JF; Chuguransky, S; Alberca, LN; Alice, JI; Goicoechea, S; Ruiz, ME; Bellera, CL; Talevi, A.
   Mini-Reviews in Medicinal Chemistry, 20(14):1447-1460 (2020)
   DOI: 10.2174/1871525718666200219130229

7. Machine learning in drug discovery and development part 1 – a primer
   Talevi, A; **Morales, JF**; Hather, G; Podichetty, J; Kim, S; Bloomingdale, PC; Kim, S; Burton, J; Brown, JD; Winterstein, AG; Schmidt, S; White, JK; Conrado, DJ.
   CPT: Pharmacometrics & Systems Pharmacology, 9(3):129-142 (2020)
   DOI: 10.1002/psp4.12491

8. Application of machine learning to predict unbound drug bioavailability in the brain
   Morales, JF; Ruiz, ME; Stratford, RE; Talevi, A.
   CPT: Pharmacometrics & Systems Pharmacology (enviado)

Detalle de la base de datos

Tabla S1. Estructuras de los compuestos que formaron parte de las bases de datos de la presente tesis. Se utilizó el formato de especificación de introducción lineal molecular simplificada, o SMILES (del inglés, *Simplified Molecular Input Line Entry Specification*), el cual es una especificación para describir sin ambigüedades la estructura de una determinada molécula.

Nombre del compuesto	**SMILES del compuesto**
6-Mercaptopurina	Sc1[n]c[n]c2[n]c[nH]c21
Acetaminofeno	CC(=O)Nc1ccc(O)cc1
Alprenolol	CC(C)NCC(O)COc1ccccc1CC=C
Amitriptilina	CN(C)CCC=C1c2ccccc2CCc2ccccc21
Anastrozol	CC(C)(C#N)c1cc(C[n]2c[n]c[n]2)cc(c1)C(C)(C)C#N
Antipirina	C[n]1c(C)cc(=O)[n]1-c1ccccc1
Apomorfina	CN1CCc2cccc3c2C1Cc1ccc(O)c(O)c1-3
Atenolol	CC(C)NCC(O)COc1ccc(CC(N)=O)cc1
Atomoxetina	Cc1ccccc1OC(CCNC)c1ccccc1
Baclofeno	NCC(CC(O)=O)c1ccc(Cl)cc1
Bencilpenicilina	CC1(C)SC2C(NC(=O)Cc3ccccc3)C(=O)N2C1C(O)=O
Bupropión	CC(C)(C)NC(C)C(=O)c1cccc(Cl)c1
Buspirona	O=C1CC2(CC(=O)N1CCCCN1CCN(CC1)c1[n]ccc[n]1)CCCC2
Cafeína	C[n]1c(=O)c2c([n]c[n]2C)[n](C)c1=O
Carbamazepina	NC(=O)N1c2ccccc2C=Cc2ccccc12
Carisoprodol	CCCC(C)(COC(N)=O)COC(=O)NC(C)C
Cefadroxilo	CC1CSC2C(NC(=O)C(N)c3ccc(O)cc3)C(=O)N2C=1C(O)=O
Cefalexina	CC1CSC2C(NC(=O)C(N)c3ccccc3)C(=O)N2C=1C(O)=O
Cetirizina	OC(=O)COCCN1CCN(CC1)C(c1ccc(Cl)cc1)c1ccccc1

Clorpromazina	CN(C)CCCN1c2ccccc2Sc2ccc(Cl)cc12
Cimetidina	Cc1[nH]c[n]c1CSCCNC(NC#N)=NC
Citalopram	CN(C)CCCC1(OCc2cc(ccc12)C#N)c1ccc(F)cc1
Clozapina	CN1CCN(CC1)C1=Nc2cc(Cl)ccc2Nc2ccccc21
Codeína	COc1ccc2CC3C4C=CC(O)C5Oc1c2C45CCN3C
Colchicina	CC(=O)NC1CCc2cc(OC)c(OC)c(OC)c2-c2ccc(OC)c(=O)cc21
Daidzeína	Oc1ccc(cc1)-c1coc2cc(O)ccc2c1=O
DAMGO	CC(NC(=O)C(N)Cc1ccc(O)cc1)C(=O)NCC(=O)N(C)C(Cc1ccccc1)C(=O)NCCO
Dantroleno	[O-][N+](=O)c1ccc(cc1)-c1ccc(C=NN2CC(=O)NC2=O)o1
Delavirdina	CS(=O)(=O)Nc1cc2cc([nH]c2cc1)C(=O)N1CCN(CC1)c1[n]cccc1NC(C)C
Desloratadina	Clc1cc2CCc3ccc[n]c3C(c2cc1)=C1CCNCC1
Desmetilclozapina	Clc1ccc2Nc3ccccc3C(=Nc2c1)N1CCNCC1
Dexametasona	CC1CC2C3CCC4=CC(=O)C=CC4(C)C3(F)C(O)CC2(C)C1(O)C(=O)CO
Diazepam	CN1C(=O)CN=C(c2cc(Cl)ccc12)c1ccccc1
Digoxina	CC12CCC(CC1CCC1C2CC(O)C2(C)C(CCC21O)C1COC(=O)C=1)OC1CC(O)C(OC2CC(O)C(OC3CC(O)C(O)C(C)O3)C(C)O2)C(C)O1
Difenhidramina	CN(C)CCOC(c1ccccc1)c1ccccc1
E2074	C[n]1c(=O)[n](CC(F)CN2C3CC(CC2CC3)OCc2cc(F)ccc2)[n]c1C
EAB 515	NC(Cc1cc(cc(CP(O)(O)=O)c1)-c1ccccc1)C(O)=O
Etil-2-fenilmalonamida	CCC(c1ccccc1)(C(N)=O)C(N)=O
Etopósido	CC1OC2C(CO1)OC(OC1C3COC(=O)C3C(c3cc4OCOc4cc31)c1cc(OC)c(O)c(c1)OC)C(O)C2O
Fexofenadina	CC(C)(c1ccc(cc1)C(O)CCCN1CCC(CC1)C(O)(c1ccccc1)c1ccccc1)C(O)=O
Flavopiridol	CN1CC(O)C(CC1)c1c2oc(cc(=O)c2c(O)cc1O)-c1ccccc1Cl
Fleroxacina	CN1CCN(CC1)c1c(F)c2c(cc1F)c(=O)c(c[n]2CCF)C(O)=O
Fluoresceína	Oc1cc2Oc3cc(O)ccc3C3(OC(=O)c4ccccc34)c2cc1
Fluoxetina	CNCCC(Oc1ccc(cc1)C(F)(F)F)c1ccccc1
Gabapentina	NCC1(CC(O)=O)CCCCC1
Ganciclovir	Nc1[nH]c(=O)c2[n]c[n](COC(CO)CO)c2[n]1
Genisteína	Oc1cc(O)cc2occ(c(=O)c12)-c1ccc(O)cc1

Haloperidol	OC1(CCN(CCCC(=O)c2ccc(F)cc2)CC1)c1ccc(Cl)cc1
Hidroxizina	OCCOCCN1CCN(CC1)C(c1ccc(Cl)cc1)c1ccccc1
Indinavir	CC(C)(C)NC(=O)C1CN(Cc2c[n]ccc2)CCN1CC(O)CC(Cc1ccccc1)C(=O)NC1C(O)Cc2ccccc21
Indometacina	Cc1c(CC(O)=O)c2cc(ccc2[n]1C(=O)c1ccc(Cl)cc1)OC
Lamotrigina	Nc1[n]c(N)c([n][n]1)-c1cccc(Cl)c1Cl
Letrozol	N#Cc1ccc(cc1)C(c1ccc(cc1)C#N)[n]1c[n]c[n]1
Levetiracetam	CCC(C(N)=O)N1CCCC1=O
Levofloxacina	CN1CCN(CC1)c1c(F)cc2c3c1OCC(C)[n]3cc(C(O)=O)c2=O
Loperamida	CN(C)C(=O)C(CCN1CCC(O)(CC1)c1ccc(Cl)cc1)(c1ccccc1)c1ccccc1
Loratadina	CCOC(=O)N1CCC(CC1)=C1c2[n]cccc2CCc2cc(Cl)ccc21
L-triptófano	NC(Cc1c[nH]c2ccccc21)C(O)=O
Manitol	OC(CO)C(O)C(O)C(O)C(O)CO
Memantina	CC12CC3CC(N)(CC(C)(C3)C1)C2
Metotrexato	CN(Cc1c[n]c2[n]c(N)[n]c(N)c2[n]1)c1ccc(cc1)C(=O)NC(CCC(O)=O)C(O)=O
Metoclopramida	COc1cc(N)c(Cl)cc1C(=O)NCCN(CC)CC
Metoprolol	CC(C)NCC(O)COc1ccc(CCOC)cc1
Midazolam	Cc1[n]cc2CN=C(c3ccccc3F)c3cc(Cl)ccc3-[n]21
Morfina	CN1CCC23C4Oc5c2c(CC1C3C=CC4O)ccc5O
Morfina-3-glucurónido	CN1CCC23C4C=CC(O)C2Oc2c3c(CC14)ccc2OC1OC(C(O)C(O)C1O)C(O)=O
Morfina-6-glucurónido	CN1CCC23C4Oc5c2c(CC1C3C=CC4OC1OC(C(O)C(O)C1O)C(O)=O)ccc5O
Moxalactama	C[n]1[n][n][n]c1SCC1COC2N(C=1C(O)=O)C(=O)C2(NC(=O)C(c1ccc(O)cc1)C(O)=O)OC
Nadolol	CC(C)(C)NCC(O)COc1cccc2CC(O)C(O)Cc21
Naltrexona	OC12CCC(=O)C3Oc4c5c(CC1N(CC1CC1)CCC253)ccc4O
Nelfinavir	Cc1c(cccc1O)C(=O)NC(CSc1ccccc1)C(O)CN1CC2CCCCC2CC1C(=O)NC(C)(C)C
Nitrofurantoína	[O-][N+](=O)c1ccc(C=NN2CC(=O)NC2=O)o1
Norfloxacina	CC[n]1cc(C(O)=O)c(=O)c2cc(F)c(cc12)N1CCNCC1
Nortriptilina	CNCCC=C1c2ccccc2CCc2ccccc21
Ondansetrón	C[n]1c2ccccc2c2c1CCC(C[n]1cc[n]c1C)C2=O

Oxprenolol	CC(C)NCC(O)COc1ccccc1OCC=C
Oxicodona	COc1ccc2CC3N(C)CCC45C(Oc1c42)C(=O)CCC53O
Oximorfona	CN1CCC23C4Oc5c2c(CC1C3(O)CCC4=O)ccc5O
Paclitaxel	CC1(C)C2C(OC(C)=O)C(=O)C3(C)C(C(OC(=O)c4ccccc4)C1(O)CC(OC(=O)C(O)C(NC(=O)c1ccccc1)c1ccccc1)C=2C)C1(COC1CC3O)OC(C)=O
Paliperidona	Cc1[n]c2C(O)CCC[n]2c(=O)c1CCN1CCC(CC1)c1[n]oc2cc(F)ccc21
Paroxetina	Fc1ccc(cc1)C1CCNCC1COc1cc2OCOc2cc1
Pefloxacina	CN1CCN(CC1)c1cc2c(cc1F)c(=O)c(c[n]2CC)C(O)=O
Pemetrexed	Nc1[nH]c2[nH]cc(CCc3ccc(cc3)C(=O)NC(CCC(O)=O)C(O)=O)c2c(=O)[n]1
Fenitoína	O=C1NC(=O)C(N1)(c1ccccc1)c1ccccc1
Pindolol	CC(C)NCC(O)COc1cccc2[nH]ccc12
Probenecid	CCCN(CCC)S(=O)(=O)c1ccc(cc1)C(O)=O
Propranolol	CC(C)NCC(O)COc1cccc2ccccc12
Pirilamina	CN(C)CCN(Cc1ccc(cc1)OC)c1cccc[n]1
Quinidina	COc1ccc2[n]ccc(C(O)C3CC4CCN3CC4C=C)c2c1
Rifampicina	CN1CCN(CC1)N=Cc1c2NC(=O)C(C)=CC=CC(C)C(O)C(C)C(O)C(C)C(OC(C)=O)C(C)C(C=COC3(C)Oc4c(c(c(c2O)c(O)c4C)c1O)C3=O)OC
Risperidona	Cc1[n]c2CCCC[n]2c(=O)c1CCN1CCC(CC1)c1[n]oc2cc(F)ccc21
Ácido salicílico	Oc1ccccc1C(O)=O
Saquinavir	CC(C)(C)NC(=O)C1CC2CCCCC2CN1CC(O)C(Cc1ccccc1)NC(=O)C(CC(N)=O)NC(=O)c1ccc2ccccc2[n]1
Sertralina	CNC1CCC(c2ccccc21)c1cc(Cl)c(Cl)cc1
Estavudina	Cc1c[n](C2C=CC(CO)O2)c(=O)[nH]c1=O
Sulpirida	CCN1CCCC1CNC(=O)c1cc(ccc1OC)S(N)(=O)=O
Sumatriptán	CNS(=O)(=O)Cc1cc2c(cc1)[nH]cc2CCN(C)C
Tacrina	Nc1c2ccccc2[n]c2CCCCc21
Tiopental	CCC1(C(C)CCC)C(=O)NC(=S)NC1=O
Tioridazina	CN1CCCCC1CCN1c2cc(ccc2Sc2ccccc12)SC
Topiramato	CC1(C)OC2C3OC(C)(C)OC3COC2(COS(N)(=O)=O)O1
Tramadol	COc1cc(ccc1)C1(O)CCCCC1CN(C)C
Trifluoperazina	CN1CCN(CCCN2c3ccccc3Sc3ccc(cc23)C(F)(F)F)CC1

Verapamilo	CN(CCc1cc(OC)c(cc1)OC)CCCC(C#N)(C(C)C)c1cc(OC)c(cc1)OC
Vinblastina	CN1C2C3(CCN4CC=CC(CC)(C34)C(OC(C)=O)C2(O)C(=O)OC)c2cc(c(cc12)OC)C1(CC2CC(O)(CN(C2)CCc2c1[nH]c1ccccc21)CC)C(=O)OC
Vorozol	C[n]1[n][n]c2ccc(cc12)C(c1ccc(Cl)cc1)[n]1c[n]c[n]1
YM992	Fc1ccc(OCC2CNCCO2)c2CCCc21
Zidovudina	Cc1c[n](C2CC(N=[N+]=[N-])C(CO)O2)c(=O)[nH]c1=O
Zolpidem	CN(C)C(=O)Cc1c([n]c2ccc(C)c[n]21)-c1ccc(C)cc1