UNIVERSIDAD NACIONAL DE LA PLATA

FACULTAD DE CIENCIAS EXACTAS

DEPARTAMENTO DE QUÍMICA

Trabajo de Tesis Doctoral

"ESTUDIO, DESARROLLO Y APLICACIÓN DE MODELOS DE LA TEORIA QSPR-QSAR EN PESTICIDAS"

Tesista: José Francisco Aranda

Director: Pablo Román Duchowicz

Codirector: Eduardo Alberto Castro

2018

"Actividad realizada en el marco del 70° Aniversario del INIFTA (1948-2018)"

El presente trabajo de Tesis se desarrolló en el Instituto de Investigaciones Fisicoquímicas Teóricas y Aplicadas (INIFTA), dependiente del Departamento de Química de la Facultad de Ciencias Exactas de la Universidad Nacional de la Plata (UNLP).

Se presenta en consideración de las autoridades de dicha Facultad para acceder al grado académico de Doctor de la Facultad de Ciencias Exactas Área Química.

Esta tesis está dedicada con profundo cariño a Irene y Julia, mi gran apoyo en la vida.

Y a la memoria de mi querido padre Julio Argentino Aranda.

"Si me caí, es porque estaba caminando. Y caminar vale la pena, aunque te caigas".

Eduardo Galeano

Agradecimientos

Quisiera expresar mi agradecimiento a todas las personas que hicieron posible el desarrollo de la presente tesis doctoral:

A mis directores Dr. Pablo R. Duchowicz y Dr. Eduardo A. Castro, por enseñarme a crecer en el ámbito científico y apoyarme en los momentos más complicados del desarrollo de la tesis.

A la Dra. Nieves C. Comelli, por su gran apoyo desde mis inicios en la decisión de trabajar en la presente tesis doctoral, hasta acompañarme en Catamarca para poder continuar con el trabajo.

A todas las personas que he conocido en el INIFTA, mi lugar de trabajo, particularmente a los integrantes del grupo "Estudio Teórico de Sistemas Químicos, Físicos y Biológicos", Dr. Eduardo A. Castro, Dr. Francisco M. Fernández, Dr. Pablo R. Duchowicz, Dr. Andrew G. Mercader, Dra. Ofelia B. Oña, Dr. Javier García y Dr. Cristian Rojas Villa, con quienes he compartido gratos momentos en el transcurso de estos años.

A la Dra. Beatriz Soria por haber permitido a Irene trabajar con ella los años que estuvimos en la Plata y por el gran apoyo de ella y todo el grupo del CEQUINOR, por acompañarnos en los momentos difíciles.

A la Facultad de Ciencias Agrarias de Catamarca, por haberme acompañado en el proceso de formación en mi carrera como Ingeniero Agrónomo, y durante el desarrollo de la tesis doctoral.

Al Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET) por otorgarme la Beca Doctoral con la cual he podido realizar mis estudios.

Finalmente, mis mayores agradecimientos son a mi esposa Irene, Julia y a mi familia: Gloria, Bety, Gaby, Cristina, Laura, Lucía, Gabriel, Araceli, Eugenia, Agustín, Francisco, Maxi, Paulina, Julieta, Camila, Pablo, Edith, Armando. Igualmente, a mi familia en España, Conchi, Paqui, Blanca, Pepe, Alberto, Laura, Alberto, Patricia, Alberto, Alejandro y Fernando.

A mis amigos de toda la vida Sebas, Fede, y todos los chicos del grupo de compañeros de la Escuela Preuniversitaria Fray Mamerto Esquiú.

Índice general

Lista de abreviaturas

2,4-D	*2,4-Ácido diclorofenoxiacético*
3D-MoRSE	*Representación molecular 3D de la estructura basada en difracción de electrones*
AEROWIN	*Programa que calcula la fracción de los compuestos absorbidos a partículas atmosféricas*
AFC	*Contribución de fragmentos de átomos*
ALS	*Acetolactato sintetasa*
ANN	*Red Neuronal Artificial*
AOs	*Orbitales atómicos*
BCF	*Factor de bioconcentración*
BCFBAFWIN	*Programa para la estimación del factor de bioconcentración y bioacumulación*
BCUT	*Autovalores de la matriz de Burden*
Blk	*Atributos estructurales bloqueados*
BOND	*Enlaces simples, dobles y estereoquímicos*
BP	*Punto de ebullición*
BSM	*Método de Subconjuntos Balanceados*
CAS	*Servicio de resúmenes químicos*
CAS-RN	*Número de registro del CAS*
CCC	*Coeficiente de correlación de concordancia*
CDB	*Base de datos de compuestos*
CE50	*Concentración efectiva en el 50% de la población*
CoMFA	*Análisis comparativo de campo molecular*
CoMSIA	*Análisis comparativo del índice de similitud molecular*
CORAL	*Programa correlación y lógica*
CW	*Peso de correlación*
DA	*Dominio de aplicación*
DBCP	*1,2 dibromo-3-cloropropano*
DCW	*Descriptor flexible basado en pesos de correlación*
DDT	*Diclorodifeniltricloroetano*
der	*Error relativo del coeficiente de regresión*
DG	*Energía libre de solvatación*
DMI	*Inhibidores de la desmetilación de esteroides*
EC	*Conectividad extendida de Morgan*
EDB	*Dibromuro de etileno*
ED50	*Dosis efectiva en el 50% de la población*
EEVA	*Descriptor electrónico de autovalores*
EFDB	*Base de datos basadas en medioambiente*
EPA	*Agencia de Protección Ambiental de EE.UU.*

EPI Suite	*Interface de programas de estimación*
EVA	*Descriptor de autovalores*
FAO	*Organización de las Naciones Unidas para la Alimentación y la Agricultura*
FIT	*Función de ajuste*
FN	*Falsos negativos*
FP	*Falsos positivos*
FS	*Búsqueda exacta o exhaustiva*
FWHM	*Anchura a media altura*
GA	*Algoritmos Genéticos*
GA-ANN	*Algoritmos Genéticos-Red Neuronal Artificial*
GAO	*Grafo de orbitales atómicos*
GETAWAY	*Ensamblado de geometría, topología y pesos atómicos*
G-media	*Media geométrica*
GFA	*Función de aproximación genética*
GHS	*Sistema globalmente armonizado*
G/PLS	*Cuadrados mínimos parciales genéticos*
GPC	*Cromatografía de permeación en gel*
GRAPH	*Grafo*
GRIND	*Descriptores independientes de cuadrícula*
HALO	*Átomos de compuestos halógenos*
HARD	*Atributo estructural del descriptor flexible HARD*
HENRYWIN	*Programa que estima la constante de Henry*
HFG	*Grafo con hidrógenos*
HLC	*Constante de la ley de Henry*
HODOC	*Base de datos del libro de compuestos orgánicos*
HSG	*Grafo sin hidrógenos*
HTS	*Barrido de alto rendimiento*
IT	*Índice topológico*
IUPAC	*Unión Internacional de Química Pura y Aplicada*
KOAWIN	*Programa que estima el coeficiente de partición octanol-aire*
Koc	*Coeficiente de sorción de suelo*
k-MCA	*Método de Análisis de Agrupamiento k-medias*
K-shape	*Descriptores del índice de forma de Kier*
LC50	*Concentración letal en el 50% de la población*
LDA	*Análisis discriminante lineal*
LD50	*Dosis letal en la mitad de la población expuesta*
LFER	*Relaciones lineales de energía libre*
lmo	*Técnica de Validación Cruzada dejar-varios-afuera*
Log Kow	*Logaritmo del coeficiente de partición octanol-agua*
Log P	*Logaritmo del coeficiente de partición octanol-agua*
Log S_w	*Logaritmo de la solubilidad en agua*
loo	*Técnica de Validación Cruzada dejar-uno-afuera*
LOVI	*Invariantes locales del vértice*
LSER	*Relaciones lineales de energía de solvatación*
LV	*Variable latente*
LWAPC	*Logaritmo del coeficiente de partición agua-aire*

MAE	*Error absoluto medio*
MAS	*Matrices de densidad electrónica grafo teóricas y ponderación atómica*
MC	*Método de simulación Monte Carlo*
MCI	*Índice de conectividad molecular de primer orden*
MDL MOL	*Formato de representación de la estructura molecular*
MIFs	*Campos de interacción molecular*
MLR	*Regresión lineal multivariable*
MLRA	*Análisis de regresión lineal multivariable*
Mold²	*Programa descriptores moleculares de estructuras 2D*
MTI	*Índice topológico molecular*
N^{iter}	*Número de iteración óptima del procedimiento Monte Carlo*
NNC	*Código del vecino más cercano a un determinado vértice*
NOSP	*Átomos de nitrógeno, oxígeno, azufre y fósforo*
NSc	*Atributos estructurales en el conjunto de calibración*
NSv	*Atributos estructurales en el conjunto de predicción*
NTRN	*Número (frecuencia) de SMILES que contienen SAk en el conjunto de calibración*
NTST	*Número (frecuencia) de SMILES que contiene SAk en el conjunto de predicción*
Nw	*Valores atípicos*
oL	*Número de compuestos salientes con residuo mayor que L veces el valor de Scal*
OCDE	*Organización para la Cooperación Económica y el Desarrollo*
OMS	*Organización Mundial de la Salud*
OPPT	*Subdivisión de Evaluación de la Exposición de la Agencia de Protección Ambiental de los EE.UU.*
Pa	*Pascales*
PaDEL	*Programa Laboratorio de Exploración de Datos Farmacéuticos*
PAIR	*Pares de átomos*
PCA	*Análisis de componentes principales*
pH	*Escala logarítmica para expresar la acidez o basicidad de un compuesto*
PHYSPROP	*Base de datos de compuestos*
pKa	*Constante de disociación ácida*
PLS	*Mínimos cuadrados parciales*
PMNs	*Prefabricación de productos químicos*
PPDB	*Base de Datos de Propiedades de Pesticidas*
PRESS	*Sumatoria de los residuos predichos al cuadrado*
PPR	*Método de relaciones propiedad-propiedad*
pt2	*Número de caminos de longitud 2*
pt3	*Número de caminos de longitud 3*
PTRN	*Probabilidad de encontrar SAk en SMILES del conjunto de predicción*
QSAR	*Relaciones Cuantitativas Estructura-Actividad*
QSPR	*Relaciones Cuantitativas Estructura-Propiedad*
QuBiLs	*Programa de mapas cuadráticos, bilineales y N-lineales*

REACH	*Registro, Evaluación, Autorización y Restricción de Compuestos Químicos*
RDF	*Función de distribución radial*
RF	*Bosques al azar*
RM	*Método del Reemplazo*
RMN	*Resonancia magnética nuclear*
RMS	*Raíz cuadrada media de la desviación estándar*
RMSP	*Error cuadrático medio de la predicción*
RS	*Representación estructural*
S	*Atributos SMILES*
SA	*Atributo estructural*
SAR	*Relación Estructura-Actividad*
SDEP	*Desviación estándar del error de predicción*
SE	*Exclusión de a pasos*
SI	*Inclusión de a pasos*
SMILES	*Sistema de entrada molecular lineal simplificado*
SMx	*Modelos de solvatación continua*
SOMFA	*Análisis de campo molecular de autoorganización*
SR	*Método de regresión paso a paso*
SRC	*Corporación de Investigación de Siracusa*
SVM	*Máquina de soporte vectorial*
SW	*Método de a pasos*
Sw	*Solubilidad en agua*
T	*Umbral (número entero) para el cálculo de DCW*
TAE	*Equivalentes atómicos transferibles*
TN	*Verdaderos negativos*
TP	*Verdaderos positivos*
TRPEV	*Teoría de repulsión de pares de electrones de valencia*
UNEP	*Programa del Medio Ambiente de la Naciones Unidas*
UNIFAC	*Coeficiente de actividad del grupo funcional quasi-químico universal*
USDA	*Departamento de Agricultura de EE.UU.*
VIF	*Factor de inflación de la varianza*
WATERNT	*Programa para la estimación de la solubilidad en agua*
WHIM	*Invariante holístico ponderado*
WHO	*Organización Mundial de la Salud*
WLN	*Sistema de notación lineal de Wiswesser*

Material Anexo

En el CD que se adjunta al ejemplar de tesis se encuentran disponibles todas las tablas que detallan las bases de datos utilizadas en las diversas aplicaciones QSPR-QSAR. Asimismo, se incluye una copia de todos los artículos científicos publicados.

Capítulo 1. Pesticidas y la Teoría QSPR-QSAR

1.1. Introducción

El objetivo principal de la ciencia de los pesticidas es ser capaz de predecir el impacto ambiental de un pesticida antes de ser liberado en el medio ambiente. Para ahorrar tiempo y dinero deberíamos ser capaces de hacer predicciones para cada pesticida, de ser posible con unos pocos ensayos experimentales y con cada vez menos experimentos en campo[1].

Los procesos ambientales, sin embargo, son enormemente complejos y a veces aparentemente aleatorios. Los sitios de mayor interés -campos agrícolas, bosques, lagos y arroyos, etc.- son ecosistemas delicados los cuales no son comprendidos en su totalidad y están sujetos a una gran variabilidad en tiempo y espacio. La diversidad y complejidad, indicadores de salud de tales ecosistemas, hace de la definición de lo que constituye un impacto significativo en tales sistemas una tarea difícil[2].

La mejor manera de desarrollar la capacidad de predicción es mejorar la comprensión de los procesos más básicos que impulsan la disipación y degradación de los pesticidas entre y dentro de los sitios ambientales, y aprender cómo estos procesos son controlados por las condiciones medioambientales. Por definición, los procesos básicos, una vez comprendidos, pueden ser extendidos a la descripción de cualquier situación.

Un ejemplo del enfoque actual para la predicción de la contaminación potencial del agua es estimar la tendencia inherente de los compuestos químicos a la lixiviación o la escorrentía a partir de sus propiedades físicas y químicas. Un índice numérico de esta tendencia es combinado con las condiciones de manejo y con las condiciones del lugar (clima, suelo, y modo

de aplicación) para determinar la contaminación potencial bajo aquellas condiciones[3].

El descubrimiento de nuevos pesticidas y su posterior desarrollo para su comercialización se ha convertido cada vez más en un proceso dificultoso y caro. Algunas pocas áreas han sido descubiertas y explotadas en las últimas décadas, con mucha mayor intensidad que en décadas anteriores, particularmente en el dominio de los insecticidas.

El criterio para que un pesticida sea aceptable ha cambiado dramáticamente. Además, las preocupaciones ambientales incrementaron la demanda en algunas nuevas áreas que han sido descubiertas. Como resultado de todo esto, los programas de descubrimiento exitosos deben ser muy eficientes para competir en el mercado actual de los pesticidas.

La eficiencia significa hacer una rápida identificación del potencial de nuevos compuestos descubiertos, y si el potencial de los mismos lo justifica, definir la rápida optimización de propiedades deseables tales como actividad, selectividad y seguridad ambiental.

En el pasado, estos procesos han sido llevados a cabo por expertos en los dominios específicos de pesticidas, insecticidas, herbicidas, fungicidas o nematicidas, quienes usaron mayoritariamente un enfoque basado en el arte y estudios intuitivos de estructura-actividad para el descubrimiento y optimización. Tales programas a menudo involucran la síntesis de miles de compuestos para descubrir un simple compuesto con las propiedades deseadas. A menudo, el compuesto encontrado ha sido descubierto más o menos por accidente y el compuesto elegido para la comercialización tiene propiedades que fueron un arreglo entre aquellos que están disponibles en el conjunto explorado.

La elección de compuestos para su síntesis ha estado basada largamente en la experiencia del pasado de los científicos de pesticidas o en base a hipótesis químicas sobre el mecanismo de acción de los compuestos. Los sustituyentes han sido utilizados en el pasado repetidamente en un intento de incrementar la actividad.

El estudio de las propiedades de interés agronómico en pesticidas tales como sus propiedades fisicoquímicas o bien sus actividades biológicas pueden realizarse mediante los estudios de la Teoría de las Relaciones Cuantitativas Estructura-Propiedad y Actividad (QSPR-QSAR), que serán analizados posteriormente.

1.2. Pesticidas

1.2.1. Definición de pesticidas

Conforme a la FAO[4] un pesticida es cualquier sustancia o mezcla de sustancias destinadas a prevenir, destruir o controlar cualquier plaga incluyendo los vectores de enfermedades en el ser humano o animal, especies no deseadas de plantas o animales que causan su perjuicio, o de otro modo interfiriendo con la producción, transformación, almacenamiento o comercialización de productos alimenticios, materias primas agrícolas, madera y productos de madera, o alimentos para animales, o que puede ser administrado a los animales para combatir insectos, arácnidos u otras plagas en o sobre sus cuerpos.

El término pesticida incluye los productos químicos utilizados como reguladores de crecimiento, defoliantes, desecantes, agentes para evitar la caída prematura de frutos, y las sustancias aplicadas a los cultivos antes o después de la cosecha para evitar el deterioro durante el almacenamiento o el transporte. El término, sin embargo, excluye a los productos químicos utilizados como fertilizantes, nutrientes de plantas y animales, aditivos alimentarios y medicamentos para animales. Un plaguicida se define también por la FAO en colaboración con la WHO/UNEP[5] como productos químicos diseñados para combatir los ataques de diversas plagas y vectores en los cultivos agrícolas, animales domésticos y los seres humanos.

Las definiciones anteriores implican que los pesticidas son agentes químicos tóxicos (compuestos orgánicos principalmente) que son deliberadamente liberados al medio ambiente para combatir las plagas en los cultivos y los vectores de enfermedades.

1.2.2. Antecedentes históricos del uso de pesticidas en agricultura y salud pública

Los antecedentes históricos del uso de pesticidas en agricultura se remontan al comienzo de la propia agricultura, y posteriormente se hicieron más pronunciados en el tiempo, debido al aumento de la población de plagas en paralelo con la disminución de la fertilidad del suelo[6]. Sin embargo, el uso de los pesticidas modernos en agricultura y salud pública procede del siglo XIX.

La primera generación de plaguicidas implicó el uso de compuestos altamente tóxicos, como arsénico (arseniato de calcio y arseniato de plomo) y un cianuro de hidrógeno, empleado como fumigante en 1860 para el control de plagas tales como hongos, insectos y bacterias. Otros compuestos incluyen al caldo bordelés (sulfato de cobre, cal y agua) y azufre. Su uso se abandonó debido a toxicidad e ineficacia.

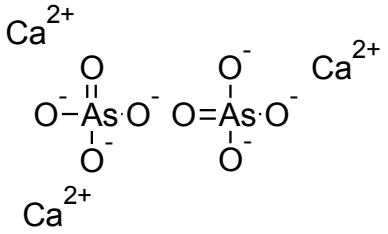


Figura 1.1. Arseniato de calcio

La segunda generación de plaguicidas implicó el uso de compuestos orgánicos de síntesis. El primero de ellos fue el 1,1,1-tricloro-2,2-bis(p-clorofenil) etano o diclorodifeniltricloroetano (DDT), primero sintetizado por el científico austríaco Othmar Zeidler, durante su tesis doctoral, y producido en los laboratorios de la Compañía Geigy en Alemania en 1874[7].

El efecto insecticida de DDT fue descubierto por el químico suizo Paul Müller en 1939. En sus primeros días, el DDT fue aclamado como un milagro debido a su amplio espectro de actividad, persistencia, insolubilidad, bajo costo económico y facilidad de aplicación[8].

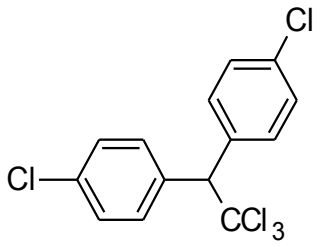


Figura 1.2. DDT

En particular, el DDT era muy eficaz en matar las plagas aumentando los rendimientos de los cultivos, y además, su bajo costo permitió extender rápidamente su uso por todo el mundo. El DDT se utilizó también para muchas aplicaciones no agrícolas. Por ejemplo, la eliminación de piojos en soldados durante la Segunda Guerra Mundial, mientras que en salud pública se lo utilizó para el control de los mosquitos vectores de la malaria. Tras el éxito del DDT, otros productos químicos fueron sintetizados durante esta era, lo que Rachel Carson[9] escribió en su libro como "Primavera silenciosa" se describe como la era de la "lluvia de productos químicos".

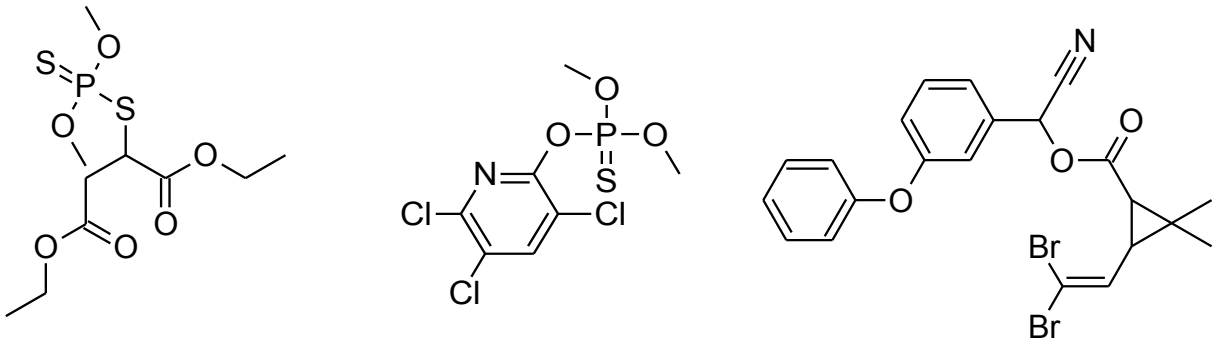
El uso intensivo de pesticidas en agricultura está unido a "la revolución verde". La revolución verde fue un movimiento agrícola mundial que comenzó en México en 1944 con el objetivo principal de aumentar los rendimientos de la producción de granos del mundo, que ya estaba en problemas para satisfacer la demanda creciente de alimentos de la población mundial.

La revolución verde implicó tres aspectos principales de las prácticas agrícolas, entre los cuales el uso de pesticidas era una parte integral. A raíz de su éxito en México, la revolución verde se extendió por todo el mundo. El control de plagas ha sido siempre importante en agricultura. Debido a la revolución verde, se hizo necesario el uso de plaguicidas en los sistemas agrícolas tradicionales, ya que la mayoría de las variedades de alto rendimiento no eran ampliamente resistentes a las plagas y enfermedades, y en parte esto era debido al sistema de monocultivo[10].

Las plagas de insectos y roedores también representan una gran pérdida de productos agrícolas almacenados. La alimentación de los insectos está basada principalmente en el endospermo del grano y el germen. El resultado es la pérdida en el peso del grano, la reducción en el valor nutritivo y el

deterioro de la calidad; por ende, afectan el aprovechamiento final del grano. Los insectos se alimentan de granos causando daños físicos y produciendo la contaminación de los mismos con excrementos, con los huevos vacíos, mudas larvales y cocones.

Un medio común de control de plagas en productos agrícolas almacenados siempre ha sido el uso de insecticidas tales como malatión, clorpirifos-metil o deltametrina, a través de impregnar dichos insecticidas sobre las superficies de los recipientes de almacenamiento[11].



(a) *(b)* *(c)*

Figura 1.3. Insecticidas: (a) malatión, (b) clorpirifos-metil, (c) deltametrina

Por otro lado, la malaria sigue siendo la principal enfermedad infecciosa transmitida por vectores en muchas partes de los trópicos. Se estima que se producen más de 300-500 millones de casos clínicos cada año, con casos como en África tropical que representan más del 90% de estas cifras[12]. Otras enfermedades transmitidas por vectores que representan un grave problema, especialmente en las zonas tropicales, incluyen la tripanosomiasis, la oncocercosis y la filariasis. Por lo tanto, es evidente que el descubrimiento de los pesticidas no es un lujo de una civilización técnica, sino más bien una necesidad para el bienestar de la humanidad.

1.2.3. La motivación del estudio de los pesticidas

El modo de acción de los pesticidas es extremadamente fascinante porque el tema abarca gran cantidad de campos de la biología y de la química y tiene muchas implicaciones prácticas.

Todas las disciplinas de la biología se han desarrollado enormemente desde el descubrimiento del DDT[13], y los demás pesticidas sintéticos se introdujeron inmediatamente después de la Segunda Guerra Mundial. En ese momento, el conocimiento de los procesos bioquímicos y fisiológicos normales en los organismos no estaba lo suficientemente estudiado como para comprender correctamente el modo de acción de los plaguicidas, su absorción, distribución y degradación en el medio ambiente.

El desarrollo de la resistencia de diversas especies de plagas a los plaguicidas no fue posible de predecir en aquel momento; qué tan rápido o en qué grado se desarrolla resistencia y qué mecanismos bioquímicos están involucrados en el proceso era una cuestión de experiencia e investigación de diversas áreas[14].

Ahora es conocida la manera en que se transmiten los impulsos nerviosos, las plantas sintetizan aminoácidos, o la manera en que los hongos invaden el tejido vegetal. La bibliografía en las diversas disciplinas biológicas se ha vuelto enorme, pero a pesar de esto, no revela la causa por la que los pesticidas interfieren con los procesos normales. Otros compuestos tóxicos se mencionan ocasionalmente, pero solamente cuando han sido herramientas para la exploración de los procesos normales.

Para poder comprender la ciencia de los pesticidas, es necesario tener conocimientos en bioquímica de las plantas, fisiología del sistema nervioso de insectos, etc., para obtener una explicación más completa de los procesos normales alterados por los pesticidas. Para comprender la toxicología de los plaguicidas, es necesario conocer la química orgánica, bioquímica, casi todas las disciplinas de fisiología vegetal y animal a nivel celular u orgánico, y ecología, así como las ciencias aplicadas en agricultura[15].

1.2.4. Producción de alimentos de baja tecnología

La mayoría de los pesticidas se han desarrollado en la industria química a gran escala, dado que el trabajo de persuasión necesario y todo el proceso de comercialización están ambos fuera del ámbito del pequeño empresario.

Por consiguiente, el desarrollo y producción de los plaguicidas puede clasificarse de alta tecnología, y el uso de tales productos no siempre se ajusta a las ideas cambiantes sobre una forma de vida mejor y más verdadera.

La agricultura orgánica y biodinámica, sin pesticidas ni antibióticos, son cada vez más populares dado que la comunidad no recibe con gran entusiasmo el uso de plantas transgénicas.

El conocimiento público de los posibles efectos colaterales negativos de la producción y uso de plaguicidas ha conducido definitivamente a un mayor requerimiento de responsabilidad por parte de la industria química, mayor prudencia por parte de los agricultores y una legislación más estricta.

Los pesticidas definitivamente han mejorado nuestras vidas al ser herramientas versátiles en la producción de alimentos y en el combate de enfermedades transmitidas por insectos[16].

1.2.5. Un gran mercado. El número de productos químicos usados como pesticidas

El Manual de Pesticidas de 1979[17] presenta 543 ingredientes activos. Aproximadamente 100 de estos son insecticidas organofosforados y 25 son carbamatos usados contra insectos. La publicación de El Manual de Pesticidas del año 2003[18] describe 812 plaguicidas y enumera 598 que son reemplazados.

Los 890 compuestos químicos sintetizados de hoy en día están aprobados como pesticidas en todo el mundo y la cantidad de productos comercializados se estima en 20700. Los insecticidas organofosforados siguen siendo el mayor grupo de insecticidas que, según El Manual de Pesticidas, contienen alrededor de 67 ingredientes activos en el mercado, pero los piretroides están aumentando en importancia, con 41 ingredientes activos.

Los inhibidores de la desmetilación de esteroides (DMI) constituyen el principal grupo de fungicidas (31). Los inhibidores de la fotosíntesis (triazinas 16, ureas 17 y otros grupos minoritarios) y los ácidos ariloxialcanoicos imitantes a las auxinas (20) continúan siendo muy populares como

herbicidas, pero muchos inhibidores de la síntesis de aminoácidos extremadamente potentes (por ejemplo, las sulfonilureas (27)) se han vuelto más importantes.

Es muy interesante estudiar las listas de pesticidas a la venta en 1945 o antes. El arseniato de plomo, las sales de mercurio y algunos compuestos orgánicos de mercurio, el arseniato de zinc, las sales de cianuro, la nicotina, el nitrocresol y el clorato de sodio se vendieron con pocas restricciones. Muy pocos de estos pesticidas tempranos se consideran ahora seguros. El mundo tenía una gran necesidad de plaguicidas seguros y eficientes como el DDT. Esta fantástica nueva sustancia comenzó a aparecer en ese momento en las listas de plaguicidas aprobados bajo varios nombres (Gesarol, Boxol S, pentaclorodifeniletano, etc.).

El herbicida 2,4-D obtuvo un estado similar al primer herbicida real efectivo que hizo posible la mecanización en agricultura. "El descubrimiento del 2,4-D como herbicida durante la Segunda Guerra Mundial precipitó el mayor avance individual en la ciencia del control de malezas y el adelanto más significativo en agricultura[19]".

1.2.6. Cantidad de pesticidas producidos

Los pesticidas exitosos se producen en cantidades masivas. Se ha estimado que, entre 1943 y 1974, la producción mundial de DDT solo alcanzó $2.8.10^9$ kg[20]. El DDT fue el primer pesticida sintético eficiente y tenía todas las buenas propiedades para un insecticida que en ese momento se pudiera imaginar. Es extremadamente estable y solo un tratamiento puede ser suficiente para un buen control de las plagas de insectos. Es económico en su producción y tenía (y aún tiene) una baja toxicidad para los humanos, pero es extremadamente activo para casi todos los insectos. Como herramienta en campañas antipalúdicas, fue extremadamente eficiente. Al final de la Segunda Guerra Mundial se utilizó para combatir enfermedades transmitidas por insectos y plagas agrícolas y domésticas, como moscas y chinches.

La producción de DDT como primer pesticida eficiente moderno alcanzó el máximo en 1963 con $8.13.10^7$ kg solo en Estados Unidos. Las prohibiciones y restricciones del uso de DDT han reducido su volumen de producción. Hoy se ha firmado un tratado internacional para restringir su uso a muy pocas aplicaciones en control de vectores.

Por lo tanto, el DDT ya no es tan importante como producto comercial. No hay protecciones de patentes, y debido a problemas ambientales, su utilidad es limitada. Además, la resistencia de los insectos al DDT en cualquier caso habría restringido su utilidad.

Otros pesticidas constituyen ahora una parte integral de la agricultura en todo el mundo y representan aproximadamente el 4.5% del costo total de la producción agrícola en Estados Unidos. El uso de pesticidas en ese país promedió más de $5.44.10^8$ kg de ingredientes activos en 1997, excediendo el precio de 11.9 mil millones de dólares, mientras que el consumo mundial de plaguicidas en 1995 se ha estimado en $2.6.10^9$ kg de ingredientes activos. Los plaguicidas superactivos más nuevos se pueden emplear a volúmenes muy bajos, incluidos los herbicidas como el glufosinato y el glifosato, e insecticidas como los piretroides sintéticos.

Los herbicidas dominan el mercado como se muestra en la Tabla 1.1.

(a) (b)

Figura 1.4. Herbicidas: (a) glufosinato de amonio, (b) glifosato

Tabla 1.1. Producción porcentual de pesticidas

Ventas	%
Herbicidas	47.6

Insecticidas	29.4
Fungicidas	17.4
Otros	5.5

Los herbicidas se aplican al 92%, y al 97%, de la superficie sembrada con maíz, algodón, soja y cítricos; tres cuartos de superficie vegetal; y dos tercios de la superficie cultivada con manzanas y otras frutas.

Los países nórdicos tienen menos plagas de insectos en su agricultura y muy pocas enfermedades humanas o veterinarias que se transmiten por insectos. Existen restricciones contra el uso de aviones para la fumigación con insecticidas en las áreas de silvicultura y agricultura. Los insecticidas son en volumen mucho menos significativos que los herbicidas en esos casos.

El 87% del uso mundial de plaguicidas está en agricultura, y Europa, EE.UU., y Japón constituyen el mercado más grande, especialmente para herbicidas, mientras que los insecticidas dominan Asia, África y América Latina.

El mercado mundial de pesticidas químicos está en torno a un valor de 31 mil millones de dólares, y está aumentando en torno del 1% al 2% por año. El costo para desarrollar un pesticida nuevo se estimó en alrededor de 80 millones de dólares en 1999, y la alta demanda de investigación toxicológica sobre cada sustancia nueva es la razón más importante del alto costo.

Resulta mucho más económico desarrollar un nuevo pesticida cuando se conoce su modo de acción. Por lo tanto, no es sorprendente que los nuevos insecticidas organofosforados y los herbicidas derivados de la urea se comercialicen cada vez más cada año. Los piretroides constituyen un nuevo grupo de reputación similar. El modo exacto de acción no se entendió durante mucho tiempo, pero en Rothamstead Experimental Station y otros institutos se llevaron a cabo estudios básicos de las relaciones estructura-actividad, lo que permitió desarrollar compuestos más activos.

1.2.7. Mercado e investigación

Muy pocas compañías multinacionales de agroquímicos dominan el mercado. Debido a la integración vertical y horizontal, el número de empresas disminuye cada año. Por ejemplo, las empresas suizas CIBA y Geigy se fusionaron para convertirse en CIBA-Geigy, que se fusionó con Sandoz para formar Novartis, que se fusionó con AstraZeneca para formar Syngenta. AgroEvo se fusionó con Rhône-Poulenc para formar Aventis.

La nueva era de la biotecnología que acaba de comenzar acelerará este proceso. Las empresas tratarán de hacerse con el mercado de semillas de cultivos transgénicos resistentes a plagas de insectos y enfermedades, o que sean tolerantes a los herbicidas.

Vale la pena mencionar que muchos países como India, Brasil, China y Sudáfrica tienen grandes productores de pesticidas. A menudo toman la producción de pesticidas más viejos sin protección de patente, y producen pesticidas que por diversas razones ya no están aprobados en EE.UU. o Europa. Un ejemplo de esto, es el insecticida organofosforado muy tóxico monocrotofos, que se canceló en EE.UU. en 1988 pero de todas maneras se produce y utiliza en Asia[21].

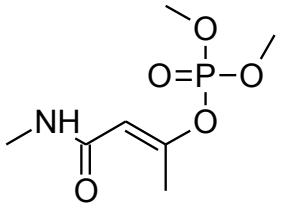


Figura 1.5. Monocrotofos

1.2.8. Lista de compuestos incluidos en el Convenio de Estocolmo

La tasa de rendimiento de un producto químico tenderá a disminuir a lo largo de los años debido a que se desarrollan nuevos compuestos competitivos, sea porque la resistencia puede restringir su utilidad, porque aparecen nuevos datos sobre su actividad ecotoxicológica, o bien por su toxicidad en la salud humana.

Muchas organizaciones involucradas en problemas ambientales intentan acelerar el proceso y promover la producción agrícola sin el uso de pesticidas[22]. Es muy común configurar listas de compuestos que poseen efectos sobre la salud humana y el medio ambiente.

Muy a menudo, estas sustancias ya han sido reemplazadas y no tienen ninguna protección de patente. Por ejemplo, la siguiente lista es producida por Pesticide Action Network. Se agrega el año de comercialización o patente. Todas las sustancias tienen más de 30 años, y muchas de ellas ya están en la lista de pesticidas reemplazados de acuerdo con El Manual de Pesticidas[23] (1994 o posterior), por lo tanto, son de menor interés.

Tabla 1.2. Lista de compuestos pesticidas retirados del mercado

Compuesto	año de entrada en mercado/patente
Aldicarb	1965
Aldrin	1948
Amitrol	1955
Binapacryl	1960
Camphechlor	1947
Clordano	1945
Chlordimeform	1966
Clorobencilato	1952
Chlorpropham	1951
DBCP[1]	1955
DDT	1942
Dieldrin	1948
Dinoseb	1945
EDB[2]	1946
Endrin	1951

Óxido de etileno	1935
Fluoroacetamida	1955
Heptacloro	1951
Hexaclorobenceno	1945
Hexaclorociclohexano (isómeros mixtos)	1940
Isobenzan	1957
Lindano	1942
Compuestos de mercurio	?
Methamidophos	1970

Fuente: http://www.pan-uk.org/briefing/SIDA_FIL/Chap1.html

[1] 1,2 dibromo-3-cloropropano

[2] dibromuro de etileno

Se alienta el cambio de las empresas agroquímicas hacia el desarrollo de plaguicidas de bajo riesgo, y los procedimientos de aprobación más eficientes son un instrumento al que puede recurrirse para acelerar el cambio. Desde 1993, la Agencia de Protección Ambiental (EPA) de EE.UU. comenzó un programa de revisión expeditivo de lo que podría clasificarse como plaguicidas de bajo riesgo.

Las revisiones aceleradas pueden reducir el tiempo de registro en más de la mitad. Resulta de interés estudiar los criterios establecidos por la EPA para plaguicidas de bajo riesgo, porque se tratan de principios importantes en el desarrollo de nuevos plaguicidas.

El pesticida:

- Debe tener un impacto reducido en la salud humana y muy baja toxicidad para los mamíferos
- Puede tener una toxicidad menor que las alternativas
- Puede desplazar sustancias químicas que plantean problemas potenciales para la salud humana, o reducir la exposición a

mezcladores, cargadores, aplicadores o en el reingreso de los trabajadores

- Puede reducir los efectos en organismos no-objetivo (como las abejas melíferas, las aves y los peces
- Puede exhibir un menor potencial de contaminación de aguas subterráneas
- Puede disminuir o implicar menos aplicaciones que en otras alternativas
- Puede tener un menor potencial de resistencia a plagas (tiene un nuevo modo de acción)
- Puede tener una alta compatibilidad con el manejo integrado de plagas
- Tiene una mayor eficacia

Aproximadamente una veintena de tales plaguicidas de riesgo reducido están ahora registrados en EE.UU., que comprenden herbicidas (5), insecticidas (8), fungicidas (5), un repelente de aves y un activador de plantas. Su modo de acción se basa en nuevos principios activos. La mayoría de ellos se enumeran a continuación, junto con el año de registro.

Tabla 1.3. Pesticidas, modo de acción y año de registro

pesticida	modo de acción	año de registro
	Herbicidas	
Imazapic	Inhibidor de ALS [1] (1997)	1997
Imazamox	Inhibidor de ALS (1997)	1997
Carfentrazona	Inhibe la protoporfirinógeno oxidasa, causando la disrupción de la membrana (1996)	1996
Dioffezopir	Inhibe el mecanismo de transporte de auxinas (1999)	1999
Dimethenamide	Inhibidor de la división celular	1999
	Insecticidas	
Diflubenzuron	Inhibidor de la síntesis de quitina	1998
Hexaflumurón	Inhibidor de la síntesis de quitina	1994
Pymetrocina	Detiene la alimentación	1999
Pyriproxyfen	Inhibe la embriogénesis	1998
Spinosad	Activa el receptor nicotínico de acetilcolina	1997
	Fungicidas	
Azoxystrobin	Bloquea la transferencia de electrones entre el citocromo b y el citocromo C_1 en la mitocondria	1997
Cyprodinil	Inhibe la síntesis de metionina	1994
Fludioxonil	Puede inhibir la fosforilación de glucosa	1993
Metalaxil-M	Inhibe la síntesis de ARN ribosómico en hongos	1996

[1] ALS: Acetolactato sintetasa

1.2.9. Toxicología, ecotoxicología y toxicología ambiental

La palabra griega τοξικον (*toxicon*) se usó para líquidos venenosos en los que se sumergían las puntas de flecha. La palabra toxicología, derivada de esta palabra, se ha usado como el nombre de la ciencia dentro de la medicina humana que describe el efecto de los venenos en los humanos.

La definición incluye captación, excreción y metabolismo de venenos (toxicocinética), así como los síntomas y la manera en que se desarrollan (toxicodinámica). Podemos decir que la toxicodinámica nos dice lo que los tóxicos hacen a los organismos; y la toxicocinética, lo que el organismo hace con la sustancia.

La toxicología también incluye la legislación que se aplica para proteger el medio ambiente y la salud humana, y las evaluaciones de riesgos necesarias para este fin. Hoy en día, un toxicólogo no trabaja exclusivamente con la especie *Homo sapiens* u organismos modelo como las ratas, sino con todo tipo de organismos[24].

El término ecotoxicología se define como "la ciencia que se ocupa de la acción de las sustancias químicas y agentes físicos sobre organismos, poblaciones y sociedades dentro de ecosistemas definidos". Incluye transferencia de sustancias e interacciones con el medio ambiente[25].

La ecotoxicología a veces se usa como sinónimo de toxicología ambiental; sin embargo, este último también abarca los efectos de los compuestos químicos ambientales y otros agentes en los seres humanos. Debido a que los procesos fisicoquímicos básicos detrás de la interacción entre las biomoléculas y los productos químicos son independientes del tipo de organismo, no es necesario tener una división demasiado rígida entre las diversas ramas de la toxicología.

1.2.10. Pesticidas, biocidas, nombres comunes, nombres químicos y nombres comerciales

Los plaguicidas son sustancias químicas específicamente desarrolladas y producidas para su uso en el control de plagas agrícolas y de salud pública, para aumentar la producción de alimentos y fibra, y para facilitar los métodos agrícolas modernos. No están incluidos los antibióticos que permiten controlar a los microorganismos como las bacterias.

Por lo general, los plaguicidas se clasifican de acuerdo con el tipo de plaga que van a controlar (fungicidas, alguicidas, herbicidas, insecticidas, nematicidas y molusquicidas). Cuando la palabra pesticida se usa sin modificación, implica un material sintetizado. El pesticida vegetal es una sustancia producida naturalmente por las plantas que la defiende contra los insectos y los microbios patógenos.

El término biocida no se emplea demasiado en la literatura científica. Se puede usar para una sustancia que es tóxica y mata varias formas de vida diferentes. Las sales de mercurio (Hg^{2+}) pueden llamarse biocidas porque son tóxicas para microorganismos, animales y muchos otros organismos, mientras que el DDT no es un biocida debido a su especificidad para los organismos con un sistema nervioso (tales como animales).

La palabra biocida también se considera un término colectivo para denotar sustancias desarrolladas intencionalmente contra organismos dañinos. En una directiva de la Comunidad Europea[26], encontramos la siguiente definición:

Los biocidas son preparados químicos que contienen una o más sustancias activas destinadas a controlar organismos nocivos por medios químicos o biológicos, pero por inferencia, no por medios físicos. La clasificación de los biocidas se divide en cuatro grupos principales: desinfectantes y biocidas generales, conservantes, control de plagas y otros biocidas, y estos se dividen en 23 categorías diferentes[27].

Los plaguicidas tienen uno o más nombres estándares y uno o más nombres químicos. Las diferentes compañías fabrican productos con nombres

comerciales registrados. Deben ser diferentes de los nombres estándares, pero también deben estar aprobados. La industria química también utiliza con frecuencia un número de código para sus productos.

En Alemania, por ejemplo, los viejos agricultores todavía conocen al paratión por el número E-605, que fue utilizado por Bayer Chemie antes de que se le otorgara un nombre estándar y se le diera una marca comercial o un nombre químico como O,O'-dietil paranitrofenil fosforotioato. El nombre químico es a menudo muy complicado e incluso difícil de interpretar para un químico. La fórmula química, sin embargo, es a menudo mucho más simple y puede decir algo sobre la propiedad del compuesto, incluso a una persona con un nivel de conocimiento moderado de la química.

Nombres comunes

Institución Británica de Estándares	Captan
Organización Internacional de Estandarización (ortografía francesa)	Captan
Ministerio Japonés de Agricultura, Silvicultura y Pesca	Captan
Sudáfrica	Captan
Norsk språkråd (norma noruega)	Kaptan

Nombres químicos

Chemical Abstracts (CA)

3a, 4,7,7a-tetrahidro-2 - [(triclorometil) tio] -1H-isoindol-1,3 (2H) -diona

Unión Internacional de Química Pura y Aplicada (IUPAC)

N- (triclorometiltio) ciclohex-4-eno-1,2-dicarboximida

Nombres comerciales

Captan, Captec, Merpan, Orthocide, Phytocape, etc. (se han registrado hasta 38 nombres comerciales y nombres químicos diferentes solo para esta sustancia).

Número de registro de Chemical Abstracts Services (CAS-RN)

133-06-2

Varios códigos

SR 406, ENT 26538

Estructura química

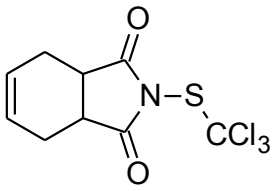


Figura 1.6. Captan

1.3. Fundamentos de la Teoría QSPR-QSAR

Hoy día encuentra verdadero interés la predicción teórica del gran número de propiedades fisicoquímicas y biológicas de las sustancias, en diversas áreas de la Química. En muchas circunstancias las medidas experimentales de tales propiedades permanecen desconocidas por trabajarse con compuestos que resultan ser nuevos, tóxicos o que demandan demasiado tiempo de medición experimental. A su vez, si la cuestión se resolviera con el simple procedimiento de síntesis y testeo de sustancias sin otra guía que la posterior prueba y error, ello constituiría una metodología extremadamente laboriosa, costosa y nada científica.

Las distintas formulaciones de la Teoría de las Relaciones Cuantitativas Estructura Propiedad (QSPR) y Actividad (QSAR)[28-30] son capaces de lograr dicho objetivo, por medio de la búsqueda de una relación hipotética entre la estructura molecular y la propiedad/actividad de las sustancias químicas.

Desde que Corwin Hansch y Toshio Fujita realizaron su trabajo fundamental en la Teoría QSPR-QSAR en 1964[31], la metodología ha evolucionado a lo largo de los años a partir de un modelo de regresión simple con pocas variables, hasta transformarse en una herramienta aplicable a un amplio rango de problemas químicos, biológicos, y farmacológicos[29]. Los avances de la Teoría QSPR-QSAR no han cesado y las aportaciones significativas se suceden continuamente.

Los estudios QSPR-QSAR son capaces de sugerir paralelismos entre la estructura y la propiedad, asisten al descubrimiento, diseño, y optimización molecular de nuevos compuestos químicos, y si son construidos racionalmente pueden hasta lograr suministrar alguna información inherente

a los mecanismos de acción molecular[30]. Una ventaja de poseer mejores descripciones de la estructura molecular es que resulta posible transferir información de una serie de moléculas a otra serie distinta.

Por ejemplo, el Factor de Bioconcentración (BCF) de un herbicida, funguicida o de cualquier pesticida tiene gran interés cuando se clasifica su potencial tóxico como resultado de una exposición crónica al mismo. La determinación experimental de BCF no es simple, por lo cual resulta de interés disponer de modelos QSPR aplicables en sustancias aún no medidas. Lo mismo vale para la acción insecticida: un estudio QSAR de la actividad de dosis letal media en ratas podría revelar aquellas estructuras que poseen mayor influencia sobre dicha actividad.

Es posible clasificar a los estudios QSPR/QSAR de dos maneras: tradicionales o semiempíricos. El trabajo fundamental que desarrollaron Hansch y Fujita se basó en un modelado tradicional, en el cual presenta especial interés establecer relaciones empíricas propiedad-propiedad. Por ejemplo, se pueden explicar propiedades "complicadas" como lo son los efectos biológicos de sustancias en términos de propiedades fisicoquímicas más simples de entender, como es el caso de la solubilidad acuosa, el coeficiente de partición entre las fases n-octanol/agua, la refracción molar, puntos de ebullición, volúmenes molares o calores de vaporización. Todas estas propiedades que son medida del carácter lipofílico, forma molecular y propiedades electrónicas dependen de la estructura, si bien en un modo indirecto, y se pueden medir experimentalmente de manera más fácil que las propiedades que buscan modelarse.

Sin embargo, en la práctica presenta mayor utilidad proponer modelos basados únicamente en información teórica deducida de la estructura molecular, requerimiento que los modelos empíricos de Hansch no cumplen.

Por tal motivo, en el contexto de la Teoría QSPR-QSAR debe representarse la estructura molecular, una propiedad genérica del sistema, a través de números denominados descriptores moleculares, cantidades teórica- o empíricamente definidas y que reflejan alguna característica global o subestructural de cada molécula[32,33].

Los descriptores se deducen de diferentes teorías, tales como la Mecánica Cuántica, la Teoría de la Información, la Teoría de Grafos[34-36], u otras, siendo los más elementales aquellos que corresponden a la cuenta de átomos y tipos de enlaces presentes en la molécula.

En la actualidad, se dispone de miles de definiciones de descriptores asequibles de la literatura, y un problema principal a resolver en QSPR/QSAR es la adecuada selección de un conjunto reducido y representativo de descriptores moleculares, para diseñar un modelo que sea capaz de explicar y predecir lo mejor posible a la propiedad bajo estudio[32].

Es factible desarrollar el trabajo QSPR-QSAR cotidiano gracias a la gran disponibilidad de métodos aproximados existentes en la literatura y que posibilitan establecer la relación estructura-propiedad/actividad. El más simple y muy conocido de todos ellos es el análisis de Regresión Lineal Multivariable (MLR)[37].

En aquellas situaciones en que la relación estructura-propiedad resulta más complicada y presenta un alto carácter no-lineal, como es el modelado de actividades biológicas de sustancias, es usual recurrir a otras técnicas estándares más elaboradas, tales como los Algoritmos Genéticos (GA)[38], o las Redes Neuronales Artificiales (ANN)[39].

Los Algoritmos Genéticos se basan en las reglas de evolución biológica de los organismos vivientes, mientras que las Redes Neuronales fueron originalmente definidas como modelo de la actividad del cerebro humano: son capaces de reconocer relaciones altamente no-lineales entre la propiedad y la estructura durante el procesamiento de los datos.

También es posible establecer modelos no-lineales y no-estadísticos por medio de la Teoría del Orden[40,41], que permite ordenar datos y realizar interpolaciones lineales sin necesidad de emplear los métodos clásicos dependientes de parámetros ajustables. La Teoría del Orden se basa en elementos básicos de la Matemática Discreta y no requiere conocer la función matemática del modelo en la relación estructura-propiedad.

Con el fin de elucidar el poder predictivo de los modelos desarrollados, la etapa de validación de los modelos QSPR-QSAR suele recurrir al uso de un conjunto externo de moléculas de validación[42], a la técnica de Validación Cruzada[43], la técnica de Aleatorización-Y [44], o a la definición del dominio de aplicabilidad del modelo[45]. No obstante, el desarrollo de los métodos de validación es un área de investigación activa y en desarrollo[46–48].

Es importante señalar que la hipótesis fundamental de QSPR-QSAR no es predecir el mecanismo de acción molecular sino la propiedad/actividad, el resultado final del mecanismo que una estructura química produce. Sin embargo, la predicción acertada de la propiedad/actividad de una sustancia permite inferir alguna información del fenómeno involucrado, como cuando una estructura es predicha activa/inactiva. Toda formulación QSPR-QSAR trata de establecer paralelismos que seleccionen los factores estructurales microscópicos preponderantes que afectan la propiedad/actividad[49-51].

1.4. Objetivos específicos

Entre los objetivos específicos del presente trabajo de tesis doctoral se citan los siguientes:

- desarrollar modelos matemáticos que resulten capaces de cuantificar relaciones hipotéticas entre la estructura química y la propiedad/actividad de pesticidas, a través de la técnica del análisis de regresión lineal multivariable aplicada a diferentes bases de datos de propiedades de interés agronómico extraídas de la literatura actualizada. Para ello, se utilizarán los mejores descriptores moleculares que surjan del análisis de miles de descriptores estructurales, obtenidos de programas computacionales de libre acceso

- investigar el comportamiento de los descriptores flexibles u óptimos en los estudios QSPR-QSAR de pesticidas, e incorporarlos en los modelos en caso que resulten adecuados. Para ello, uno debe ser capaz de definir la construcción matemática del descriptor flexible, y debe elegir el procedimiento de ajuste de sus partes variables para alcanzar las mejores predicciones de la propiedad, evitando el sobreajuste del conjunto de calibración para así poder

alcanzar una calidad predictiva aceptable y el modelo supere su validación externa

- abordar el tratamiento de grandes conjuntos moleculares de alta diversidad estructural y que incluyan pesticidas

- demostrar a través de los resultados encontrados que un enfoque basado en la representación estructural independiente de la conformación molecular permite alcanzar predicciones confiables de la propiedad/actividad estudiada

La calidad de las predicciones conseguidas con estos estudios QSPR-QSAR de pesticidas se comparará con la información experimental disponible y a través de las predicciones alcanzadas por metodologías teóricas alternativas de la literatura.

La presente tesis se enfoca en la construcción de modelos predictivos y que sean de utilidad como herramienta para asistir la búsqueda de estructuras químicas con valores favorables de la propiedad/actividad. La habilidad de predecir las propiedades fisicoquímicas y actividades biológicas de las sustancias químicas permite analizar de antemano las propiedades de compuestos nuevos, tóxicos o que demandan demasiado tiempo de evaluación experimental. Así, los modelos pueden ser utilizados para la predicción de las propiedades fisicoquímicas/actividades biológicas de los nuevos compuestos químicos sintetizados en el laboratorio y carentes de datos experimentales.

Bibliografía

1. Casida, J. E. & Quistad, G. B. Why insecticides are more toxic to insects than people: the unique toxicology of insects. *J. Pestic. Sci.* **29,** 81–86 (2004).
2. Matson, P. A., Parton, W. J., Power, A. G. & Swift, M. J. Agricultural intensification and ecosystem properties. *Science (80-.).* **277,** 504–509 (1997).
3. Wauchope, R. D., Buttler, T. M., Hornsby, A. G., Augustijn-Beckers, P. W. M. & Burt, J. P. The SCS/ARS/CES pesticide properties database for environmental decision-making. in *Reviews of environmental contamination and toxicology* 1–155 (Springer, 1992).
4. FAO. International Code of Conduct on the Distribution and Use of Pesticides, Italy, Rome. (1989).
5. Organization, W. H. Public health impact of pesticides used in agriculture. (1990).
6. Muir, P. The History of Pesticides Use. *Oregon State Univ. Press. USA* (2002).
7. Othmer, K. Encyclopedia of Chemical Technology, New York, USA. (1996).
8. Kenneth, M. The DDT Story. *London, UK* (1992).
9. Carson, R. Silent spring. 1962. (2009).
10. Vocke, G. The green revolution for wheat in developing countries. *Staff Rep.* (1986).
11. McFarlane, J. A. Guidelines for pest management research to reduce stored food losses caused by insects and mites (ODNRI Bulletin No. 22). (1989).
12. Organization, W. H. Vector control for malaria and other mosquito-borne diseases: Report of a WHO study group. (1995).
13. Beard, J., Marshall, S., Jong, K., Newton, R., Triplett-McBride, T., Humphries, B., & Bronks, R. 1, 1, 1-trichloro-2, 2-bis (p-chlorophenyl)-ethane (DDT) and reduced bone mineral density. *Arch. Environ. Heal. An Int. J.* **55,** 177–180 (2000).
14. Anber, H. A. I. Studies on Pesticide Resistance: The Biochemical Genetics of Resistance to Organophosphates and Carbamates in the Predacious Mite, Amblyseius Potentillae (Garman). (1989).
15. Ware, G. W. *The pesticide book.* (Thomson Publications, 2000).
16. Casida, J. E. & Quistad, G. B. Golden age of insecticide research: past, present, or future? *Annu. Rev. Entomol.* **43,** 1–16 (1998).
17. British Crop Protection Council & Worthing, C. R. *The pesticide manual: a world compendium.* (British crop protection council London, 1979).
18. Tomlin, C. The Pesticide Manual, A World Compendium, British Crop. *Prot. Counc. Farnham, United Kingdom* (2003).
19. Peterson, G. E. The discovery and development of 2, 4-D. *Agric. Hist.* **41,** 243–254 (1967).
20. Woodwell, G. M., Craig, P. P. & Johnson, H. A. DDT in the biosphere: where does it go? *Science (80-.).* **174,** 1101–1107 (1971).
21. Agrahari, S., Pandey, K. C. & Gopal, K. Biochemical alteration induced by monocrotophos in the blood plasma of fish, Channa punctatus (Bloch). *Pestic.*

Biochem. Physiol. **88,** 268–272 (2007).

22. de Estocolmo, C. Convenio de Estocolmo sobre Contaminantes Orgánicos Persistentes (COPs). *PNUMA Prod. Químicos, Ginebra. Texto a descasrgar http//www. pops. int Febrero* (2005).
23. Tomlin, C. The Pesticide Manual, British Crop Protection Council. *Farnham, Surrey, UK* **1250,** (1994).
24. Hayes, A. W. *Principles and methods of toxicology*. (Crc Press, 2007).
25. Hodgson, E., Mailman, R. B., Chambers, J. E. & Dow, R. E. *Dictionary of toxicology*. (Macmillan Reference London, UK, 1998).
26. Europeas, D. O. de las C. Directiva 98/8/CE relativa a la comercialización de biocidas.
27. Jørgen, S. Chemical Pesticides: Mode of Action and Toxicology. (2004).
28. Hansch, C. Quantitative approach to biochemical structure-activity relationships. *Acc. Chem. Res.* **2,** 232–239 (1969).
29. Mihalić, Z. & Trinajstić, N. A graph-theoretical approach to structure-property relationships. (1992).
30. Kier, L. B. & Hall, L. H. *Molecular connectivity in structure-activity analysis*. (Research Studies, 1986).
31. Katritzky, A. R., Lobanov, V. S. & Karelson, M. QSPR: the correlation and quantitative prediction of chemical and physical properties from structure. *Chem. Soc. Rev.* **24,** 279–287 (1995).
32. Randic, M. Molecular profiles novel geometry-dependent molecular descriptors. *New J. Chem.* **19,** 781–791 (1995).
33. Golbraikh, A. & Tropsha, A. Beware of q2! *J. Mol. Graph. Model.* **20,** 269–276 (2002).
34. Konovalov, D. A., Llewellyn, L. E., Vander Heyden, Y. & Coomans, D. Robust cross-validation of linear regression QSAR models. *J. Chem. Inf. Model.* **48,** 2081–2094 (2008).
35. Wold, S., Eriksson, L. & Clementi, S. Statistical validation of QSAR results. *Chemom. methods Mol. Des.* 309–338 (1995).
36. Gramatica, P. Principles of QSAR models validation: internal and external. *Mol. Inform.* **26,** 694–701 (2007).
37. Schüürmann, G., Ebert, R.-U., Chen, J., Wang, B. & Kühne, R. External Validation and Prediction Employing the Predictive Squared Correlation Coefficient — Test Set Activity Mean vs Training Set Activity Mean. *J. Chem. Inf. Model.* **48,** 2140–2145 (2008).
38. Roy, K. & Mitra, I. On various metrics used for validation of predictive QSAR models with applications in virtual screening and focused library design. *Comb. Chem. High Throughput Screen.* **14,** 450–474 (2011).
39. Chirico, N. & Gramatica, P. Real external predictivity of QSAR models: how to evaluate it? Comparison of different validation criteria and proposal of using the concordance correlation coefficient. *J. Chem. Inf. Model.* **51,** 2320–2335 (2011).

Capítulo 2. Descriptores moleculares

2.1. Introducción

La geometría o estructura molecular se refiere a la disposición tridimensional de los átomos en la molécula, y determina muchas de las propiedades de las sustancias químicas, tales como reactividad, polaridad, fase, color, magnetismo, actividad biológica, etc.

Actualmente, el principal modelo estructural es la teoría de repulsión de pares de electrones de valencia (TRPEV), empleada internacionalmente por su gran predictividad. Dicho modelo es uno de los conceptos más importantes en el desarrollo científico. El razonamiento basado en la estructura molecular ha jugado un papel importante en el desarrollo de la fisicoquímica, la física molecular, la química orgánica, química cuántica, síntesis química, química médica, etc.[1]

Las geometrías moleculares se determinan experimentalmente mejor cuando las muestras están próximas al cero absoluto de temperatura, porque a temperaturas más altas las moléculas presentarán un movimiento rotacional considerable. En el estado sólido la geometría molecular puede ser determinada por difracción de rayos X. Las geometrías se pueden calcular teóricamente por procedimientos mecanocuánticos *ab initio* o por métodos semiempíricos de modelado molecular[2].

La posición de cada átomo se determina por la naturaleza de los enlaces químicos con los que se conecta a sus átomos vecinos. La geometría molecular puede describirse por las posiciones de estos átomos en el espacio, mencionando la longitud de enlace de dos átomos unidos, ángulo de enlace de tres átomos conectados y ángulo de torsión de tres enlaces consecutivos.

La geometría o estructura molecular es un sistema complejo. Las ciencias de la complejidad estudian fenómenos, sistemas, o comportamientos

de complejidad creciente; esto es, fenómenos y sistemas que aprenden y se adaptan[3]. Los sistemas de complejidad creciente se caracterizan principalmente porque presentan dinámicas irreversibles, súbitas, imprevisibles, aperiódicas, además de varios otros rasgos característicos.

Los sistemas complejos pueden y deben ser simulados, y la simulación apunta a la importancia de la computadora, de las ciencias computacionales y las herramientas y enfoques propios de la computación[4,5].

En complejidad, los grados de libertad de un fenómeno o sistema consisten en el número de parámetros necesarios o posibles, según el caso, para comprender o explicar un fenómeno.[6] La complejidad de un fenómeno es directamente proporcional a los grados de libertad que exhibe un sistema, de manera que a mayores grados de libertad mayor complejidad.

Es evidente que una molécula, con su concepto integrado de estructura molecular, cumple exactamente estas condiciones. Las propiedades de una molécula no dependen sólo de las propiedades de los átomos componentes, sino también de sus conexiones mutuas. Por lo tanto, son inherentes a la organización molecular en su conjunto y estabilidad y no se pueden derivar como la suma de las propiedades de los átomos componentes. En principio, tenemos un sistema holístico.

Debido a su complejidad, la estructura molecular no puede ser representada por un modelo formal único. Dependiendo del nivel de la aproximación teórica subyacente, varias representaciones moleculares pueden representar la misma molécula. Estas representaciones, sin embargo, a menudo no son derivables la una de la otra.

Se han propuesto representaciones estructurales diferentes, por ejemplo, la representación tridimensional euclidiana, la representación bidimensional en base a la teoría de grafos, o las representaciones vectoriales llamadas 'indicadores' donde se almacenan las frecuencias de varios fragmentos moleculares.

La historia de la Teoría QSPR-QSAR comenzó un siglo antes que la historia de los descriptores moleculares, estando estrechamente relacionada

con el desarrollo de las teorías de la estructura molecular. El modelado QSPR-QSAR nació en el campo de la toxicología. Los intentos de cuantificar las relaciones entre la estructura química y la toxicidad aguda potencial han formado parte de la literatura toxicológica durante más de 100 años. En defensa de su tesis titulada "Acción del alcoholismo sobre el organismo" en la Facultad de Medicina de la Universidad de Estrasburgo en Francia (1863), Cros señaló que existía una relación entre la toxicidad de los alcoholes alifáticos primarios y su solubilidad en agua.

Crum-Brown y Fraser (1868-69)[7–9] propusieron la existencia de una correlación entre la actividad biológica de diferentes alcaloides y su constitución molecular.

Más específicamente, la acción fisiológica de una sustancia $\mathbf{\Phi}$ en un determinado sistema biológico se definió como una función f de su constitución química (**C**):

$$\mathbf{\Phi} = f(\mathbf{C}) \tag{2.1}$$

Por lo tanto, una alteración en la constitución química se reflejaría por un efecto sobre la actividad biológica. Esta ecuación puede considerarse la primera formulación general de un modelo QSPR-QSAR.

Años más tarde, una hipótesis sobre la existencia de correlaciones entre la estructura molecular y las propiedades fisicoquímicas se informó en el trabajo de Körner (1874)[10], que se ocupó de la síntesis de bencenos disustituidos y el descubrimiento de los derivados orto, meta y para. Se pensó que los diferentes colores de los bencenos disustituidos estaban relacionados con sus diferencias en la estructura molecular. Diez años más tarde, Mills (1884)[11] publicó un estudio sobre el punto de fusión y el punto de ebullición relacionado con la composición.

Los modelos cuantitativos de propiedad y actividad, comúnmente referidos como el inicio de los estudios sistemáticos de modelado QSPR-QSAR,[12] han surgido de la búsqueda de relaciones entre la potencia de los anestésicos locales y el coeficiente de partición aceite/agua[13] entre la narcosis y la longitud de la cadena[14], y la narcosis y la tensión superficial[15]. En

particular, los conceptos desarrollados por Meyer y Overton son a menudo referidos como la Teoría Meyer-Overton de la acción narcótica[13,14]. Los primeros enfoques teóricos de modelos QSPR-QSAR datan de finales de la década de 1940 y son aquellos que relacionan actividades biológicas y propiedades fisicoquímicas con índices numéricos teóricos derivados de la estructura molecular.

Sobre la base de la teoría de grafos, el índice de Wiener[16] y el número de Platt[17] propuestos en 1947 para modelar el punto de ebullición de los hidrocarburos, fueron los primeros descriptores moleculares teóricos basados en la teoría de grafos.

A principios de la década de 1960, se propusieron nuevos descriptores moleculares, dando inicio a estudios sistemáticos sobre los descriptores moleculares, principalmente basados en la teoría de grafos[18–27].

El uso de descriptores químico-cuánticos en los estudios QSPR-QSAR se remonta a principios de la década de 1970[24], aunque tales descriptores se definieron y utilizaron mucho tiempo atrás en el marco de la química cuántica. Durante 1930-60, los hitos fueron los trabajos de Pauling[28,29] y Coulson[30] en el enlace químico, Sanderson[31] en electronegatividad, y Fukui *et al.*[32] y Mulliken[33] en la distribución electrónica.

Una vez que el concepto de estructura molecular se consolidó definitivamente por los éxitos de las teorías químico-cuánticas y se aceptaron los enfoques para el cálculo de índices numéricos que codifican información de la estructura molecular, todos los elementos constitutivos para el despegue de las estrategias del modelado QSPR-QSAR estuvieron disponibles.

Basado en la ecuación de Hammett[34,35] el trabajo seminal de Hammett dio lugar a la cultura $'\sigma-\rho'$ en la delineación de los efectos de los sustituyentes sobre las reacciones orgánicas, cuyo objetivo fue la búsqueda de relaciones lineales de energía libre (LFER)[36]: se definieron las constantes estéricas, electrónicas e hidrofóbicas, convirtiéndose en una herramienta básica para modelar las propiedades de las moléculas.

En la década de 1950, los trabajos fundamentales de Taft[37–39] en fisicoquímica orgánica fueron el fundamento de las relaciones entre las propiedades fisicoquímicas y las energías de interacción soluto-solvente (relaciones lineales de energía de solvatación, LSER), basadas en parámetros estéricos, polares y de resonancia para grupos sustituyentes en compuestos congenéricos.

A mediados de la década de 1960, liderado por los trabajos pioneros de Hansch[18,40,41], el enfoque QSPR-QSAR comenzó a asumir su aspecto moderno.

En 1962, Hansch *et al.*[40] publicaron su estudio sobre las relaciones estructura-actividad de los reguladores del crecimiento de las plantas y su dependencia de las constantes de Hammett y la hidrofobicidad. Usando el sistema octanol/agua, se midieron una serie completa de coeficientes de partición, y por lo tanto, se introdujo una nueva escala hidrofóbica para describir la actitud de las moléculas para moverse a través de entornos caracterizados por diferentes grados de hidrofilia tales como la sangre y las membranas celulares. La delineación de los modelos de Hansch condujeron a un desarrollo explosivo en el análisis QSPR-QSAR y los enfoques relacionados[42].

En los mismos años, Free y Wilson desarrollaron un modelo de contribución de sustituyentes aditivos para actividades biológicas, dando un empuje adicional para el desarrollo de estrategias QSPR-QSAR. Ellos propusieron un modelo para respuestas biológicas en base a la presencia/ausencia de grupos de sustituyentes en un esqueleto molecular común[43,44]. Este enfoque llamado "enfoque *de novo*" cuando se presentó en 1964, estaba basado en la asunción de que cada sustituyente proporciona un efecto aditivo y constante a la actividad biológica, independientemente de los otros sustituyentes del resto de la molécula.

Hacia finales de 1960, muchas relaciones estructura-actividad fueron propuestas basadas no sólo en los efectos de los sustituyentes, sino también, en los índices que describen la estructura molecular completa. Estos índices teóricos fueron derivados de la representación topológica de las moléculas,

principalmente, aplicando conceptos de la teoría de grafos, y posteriormente, refiriéndose a ellos como descriptores 2D.

Los trabajos fundamentales de Balaban,[45,46], Randiç,[47,48], y Kier *et al.*[49], llevó a futuros desarrollos en el enfoque QSPR-QSAR basados en índices topológicos (ITs).

Como una extensión natural de la representación topológica de una molécula, los aspectos geométricos fueron tomados en cuenta desde la mitad de 1980s, llevando al desarrollo de los modelos QSAR-3D, los cuales aprovechan la información de la geometría molecular.

Los descriptores geométricos fueron derivados de las coordenadas espaciales en 3D de una molécula, y entre ellos, había índices de sombras[50], descriptores de carga parcial de área superficial[51], descriptores invariantes moleculares holísticamente pesados (WHIM)[52], índices gravitacionales[53], descriptores de autovalores (EVA)[54], descriptores de representación molecular-3D de la estructura basada en difracción de electrones (3D-MoRSE)[55], descriptores electrónicos de autovalores (EEVA)[56], descriptores topológico-geométricos, y descriptores de ensamblado de geometría, topología y pesos atómicos (GETAWAY)[57].

A fines de 1980 fue propuesta una nueva estrategia para describir las características moleculares basada en campos de interacción molecular (MIFs), los cuales están compuestos por la energía de interacción entre una molécula y pruebas específicas en puntos espaciales en el espacio 3D.

Diferentes pruebas tales como agua, grupos metilos, e hidrógeno, fueron usadas para evaluar la energía de interacción en miles de puntos de la grilla donde la molécula es embebida. Como resultado final de este enfoque, se obtiene un campo escalar (enrejado) de valores de energía de interacción que caracterizan a la molécula.

La primera formulación de un modelo de grilla o red para comparar moléculas por alineación de ellas en el espacio 3D y extraer la información química del MIF, fue propuesta por Goodford[58] en el método de grilla, y luego por Cramer *et al.*[59] en el análisis comparativo de campo molecular (CoMFA).

Aún basados en campos de interacción molecular, numerosos otros métodos fueron propuestos satisfactoriamente, entre ellos se pueden nombrar al análisis comparativo del índice de similitud molecular (CoMSIA)[60], el método del compás[61], descriptores G-WHIM[62], análisis de campo de Voronoi[63], el enfoque VolSurf[64], y los descriptores GRIND[65].

Finalmente, en años reciente la comunidad científica ha aumentado su interés en la química combinatoria, en el barrido de alto rendimiento (HTS), el análisis sub-estructural, y la búsqueda por similaridad, para los cuales numerosos enfoques de similaridad/diversidad han sido propuestos principalmente basados en descriptores sub-estructurales como los descriptores indicadores[66–68].

2.2. Descriptores moleculares

En las últimas décadas, un gran número de investigaciones científicas se han enfocado en capturar y convertir -por medio de un camino teórico-la información codificada en la estructura molecular, dentro de un número o muchos números, usados para establecer relaciones cuantitativas entre estructuras y propiedades o actividades biológicas experimentales.

Los descriptores moleculares son formalmente representaciones matemáticas de una molécula obtenidas por un algoritmo bien determinado y aplicado para definir una representación molecular o un procedimiento experimental bien especificado: "El descriptor molecular es el resultado final de un procedimiento lógico y matemático el cual transforma la información química codificada dentro de una representación simbólica de una molécula en un número útil o el resultado de algún experimento estandarizado"[69].

Los descriptores juegan un papel fundamental en química, en ciencias farmacéuticas, en políticas de protección medioambiental, toxicología, ecotoxicología, investigación en salud y en el control de calidad. La evidencia del interés de la comunidad científica en los descriptores moleculares, está augurada por el enorme número de descriptores propuestos hoy en día: más de 3000 descriptores[69], derivados de diferentes teorías y enfoques que están

actualmente definidos y calculados mediante el uso de herramientas computacionales.

Cada descriptor molecular tiene en cuenta una pequeña parte de toda la información química contenida en la molécula real, y como consecuencia de esto, el número de descriptores aumenta continuamente con la solicitud creciente de investigaciones profundas en sistemas químicos y biológicos. Los diferentes descriptores son diferentes maneras o perspectivas de ver una molécula, tomando en cuenta varias características de su estructura química.

Los descriptores moleculares se han convertido en una de entre las más importantes variables usadas en los estudios computacionales y, consecuentemente, manejados por herramientas estadísticas, quimiométricas y quimioinformáticas.

La disponibilidad de descriptores moleculares no ha sido sólo una nueva oportunidad para el desarrollo de nuevos estudios QSPR-QSAR, sino también ha sido un gran cambio en el paradigma de la investigación en esta área de investigación. En efecto, el uso de los descriptores moleculares ha permitido por primera vez, unir el conocimiento experimental con la información teórica surgida de la estructura molecular.

Mientras que en la década de 1960-70 el modelado molecular consistía en la búsqueda de relaciones matemáticas entre las cantidades experimentales medidas, hoy en día, consiste principalmente en la búsqueda de modelos entre propiedades/actividades medidas y descriptores moleculares teóricos o semiempíricos, capaces de capturar la información química estructural.

Una consideración general sobre el uso de descriptores moleculares en problemas de modelado tiene que ver con el contenido de la información. Esto depende del tipo de representación molecular usada y el algoritmo definido para su cálculo.

Existen descriptores moleculares simples derivados del conteo de algún tipo de átomo o fragmento estructural, así como propiedades fisicoquímicas y

propiedades de cantidad. Algunos ejemplos son el peso molecular, el número de enlaces dadores/aceptores de hidrógenos, y el número de grupos-OH.

Otros descriptores moleculares son derivados de algoritmos aplicados a la representación topológica y usualmente llamados topológicos o descriptores 2D. A su vez, están también los descriptores moleculares derivados de las coordenadas espaciales (x, y, z) de la molécula, usualmente llamados descriptores geométricos o 3D; otra clase de descriptores moleculares llamados descriptores 4D derivan de las energías de interacción entre la molécula embebida en una grilla, y alguna prueba específica.

A pesar de su gran diversidad, los descriptores moleculares se consideran para describir solamente tres aspectos fundamentales de una molécula -sus propiedades hidrofóbicas, electrónicas y estéricas-, las cuales son sin duda responsables de la actividad biológica en un organismo.

El campo de los descriptores moleculares es interdisciplinario y abarca muchas teorías y disciplinas diferentes. Se requieren conocimientos de álgebra, teoría de grafos, teoría de la información, química computacional, fisicoquímica, química cuántica, así como las teorías de la reactividad orgánica para la definición de descriptores moleculares.

Desde el principio con Wiener, los descriptores moleculares numéricos llamados índices topológicos (IT) por Hosoya, han ganado gradual aceptación junto con otros descriptores nombrados anteriormente para ser usados en los estudios QSPR-QSAR.

En las próximas secciones de este capítulo se seguirá un desarrollo histórico, pero también una clasificación dentro de una primera, segunda y tercera generación de IT, de acuerdo a la naturaleza del descriptor ya sea local o global (número entero o real), o bien, a partir de un solo átomo o una molécula completa, respectivamente. Las intercorrelaciones entre algunos de estos IT hacen a algunos de ellos redundantes.

Los IT son por un lado aceptados como herramienta legítima en estudios QSPR-QSAR, pero por otro lado su número no se ha visto incrementado a la misma velocidad como en décadas previas.

Los IT son números asociados a fórmulas integradas por operaciones matemáticas en los grafos moleculares. La necesidad de utilizar tales herramientas como IT se origina en el hecho de que las propiedades fisicoquímicas o actividades biológicas son expresadas como números, lo que permite así de alguna manera, tener una medida científica para hacer comparaciones y correlaciones. Por el contrario, las estructuras químicas, incluso expresadas en la forma matemática de grafos, son entidades discretas.

Con el fin de evaluar cuantitativamente el grado de similaridad o disimilaridad de las estructuras químicas o de encontrar correlaciones entre estructuras y propiedades uno necesita trasladar las estructuras dentro de números.

En el caso de factores electrónicos, la química cuántica o las relaciones lineales de energía libre proveen tales datos numéricos. Para factores estéricos o de hidrofobicidad/hidrofilicidad, existen datos numéricos bien establecidos. Para la forma molecular, sin embargo, los IT proveen una solución simple, ya que no requieren de tiempos largos en los cálculos computacionales, a diferencia de los métodos promovidos por Mezey[70], Pearlman[71], Burden[72], o por Cramer[73].

2.3. Representaciones de la estructura molecular

La representación molecular es la manera en la cual una molécula, un cuerpo real fenomenológico, es simbólicamente representado por un procedimiento formal específico y reglas convencionales. La cantidad de información química disponible depende del tipo de representación molecular empleada[74,75].

La representación molecular más simple es la fórmula molecular, la cual es considerada como una lista de los diferentes tipos de átomos, cada elemento acompañado por un subíndice que representa el número de ocurrencias de los átomos en la molécula. Por ejemplo, la fórmula química del p-clorotolueno es C_7H_7Cl, lo que indica la presencia en la molécula de 15 átomos distinguiéndolos en $N_C = 7$, $N_H = 7$, y $N_{Cl} = 1$.

Esta representación es independiente de cualquier conocimiento concerniente de la estructura molecular, por lo tanto, los descriptores moleculares obtenidos de la fórmula química pueden ser llamados descriptores 0D. Ejemplos de estos son el número de átomos, el peso molecular, el conteo del tipo de átomos, es decir, en general descriptores constitucionales y cualquier función de las propiedades atómicas.

Las propiedades atómicas están constituidas por el peso molecular usado para caracterizar a los átomos de la molécula; las propiedades atómicas más comunes como: la masa atómica, la carga atómica, el radio covalente y de van der Waals, la polaridad atómica, y la constante de hidrofobicidad atómica.

La representación de la sub-estructura puede ser considerada como una representación 1D de la molécula y está formada por una lista de fragmentos estructurales; la lista puede ser sólo una lista parcial de fragmentos, tales como grupos funcionales, o sustituyentes de interés, pero que no requieren un conocimiento completo de toda la estructura molecular.

Los descriptores derivados de este tipo de representación son referidos como descriptores 1D y son típicamente usados en análisis de sub-estructura y búsquedas de sub-estructuras, con un nombre común de descriptores moleculares conocidos como “descriptores indicadores”.

La representación 2D de una molécula considera la manera en que los átomos están conectados y la naturaleza de los enlaces químicos. Los enfoques basados en el grafo molecular permiten una representación 2D, generalmente conocida como representación topológica.

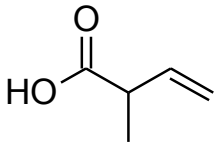


Figura 2.1. Representación 2D

El grafo molecular representa la conectividad atómica, independientemente de parámetros geométricos como las distancias

interatómicas de equilibrio, los ángulos de enlace y los ángulos de torsión. Por lo tanto, un grafo molecular es una representación topológica de la molécula de la que se derivan muchos descriptores moleculares. Estos son descriptores 2D y generalmente son invariantes del grafo conocidos con el nombre de IT.

Las representaciones bidimensionales alternativas al grafo molecular son los sistemas de notación lineal, como la propuesta por el sistema de notación lineal de Wiswesser (WLN)[76] y la notación del sistema de entrada molecular lineal simplificado (SMILES)[77].

La representación molecular 3D considera una molécula como un objeto geométrico rígido en el espacio y permite una representación dependiente no solo de la naturaleza y conectividad de los átomos, sino también de la configuración espacial general de la molécula. Esta representación geométrica define a una molécula en término de tipos de átomos que constituyen la misma y el conjunto de coordenadas (x, y, z) asociadas a cada átomo (Figura 2.2)

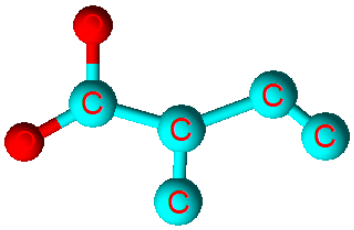


Figura 2.2. Representación 3D

Los descriptores moleculares derivados de esta representación se denominan descriptores 3D, y ejemplos de ellos son los descriptores geométricos, varios descriptores estéricos y descriptores del tamaño.

La representación masiva de una molécula la describe en términos de un objeto físico con atributos tridimensionales, tales como propiedades estéricas, el área y su volumen.

La representación estéreo electrónica (o representación de grilla) de una molécula es una descripción molecular relacionada con aquellas propiedades moleculares que surgen de la distribución de electrones, la interacción de la

molécula con diferentes tipos de pruebas caracterizando el espacio que los rodea (por ejemplo, MIF). Esta representación es típica de las técnicas QSAR basadas en grilla. Los descriptores en este nivel se pueden considerar descriptores 4D, que se caracterizan por un campo escalar, es decir, un entramado de números escalares, derivados de la geometría molecular tridimensional.

Finalmente, la representación estereodinámica de una molécula es una representación dependiente del tiempo que agrega propiedades estructurales a las representaciones 3D, como la flexibilidad, el comportamiento conformacional y las propiedades de transporte. El QSAR dinámico es un ejemplo de enfoque multiconformacional[78,79].

2.3.1. Descriptores 0D o de conteo

Todos los descriptores moleculares para los cuales no hay información sobre la estructura molecular y los enlaces entre sus átomos, pertenecen a la clase de descriptores 0D. Los recuentos de átomos y enlaces, así como la suma o el promedio de los valores atómicos, son las propiedades típicas de esta clase de descriptores.

Estos descriptores pueden ser siempre fáciles de calcular, son interpretados naturalmente, no requieren optimización de la estructura molecular, y son independientes de cualquier problema conformacional. Por lo general, muestran una degeneración muy alta, es decir, tienen valores iguales para varias moléculas, como es el caso de los isómeros.

Su contenido de información es bajo, pero sin embargo pueden desempeñar un papel importante en el modelado de varias propiedades fisicoquímicas o formar parte de modelos más complejos.

2.3.2. Descriptores 1D o indicadores

Todos los descriptores moleculares que pueden calcularse a partir de la información sub-estructural de la molécula pertenecen a los descriptores 1D.

El recuento de los grupos funcionales y los fragmentos de la sub-estructura, así como los descriptores centrados en el átomo, son los más conocidos 1D.

Estos descriptores a menudo se presentan como descriptores indicadores (ya que codifican una información distintiva de cada molécula), es decir, un vector binario donde 1 indica la presencia de la sub-estructura definida y 0 su ausencia.

Una ventaja relevante en la descripción de las moléculas mediante descriptores indicadores es la posibilidad de realizar cálculos rápidos en problemas de similitud/diversidad en conjuntos grandes de moléculas.

Al igual que los descriptores 0D, estos descriptores pueden calcularse fácilmente, se interpretan de forma natural, no requieren la optimización de la estructura molecular y son independientes de cualquier problema de conformación. Por lo general, muestran una degeneración media-alta y a menudo son muy útiles para modelar propiedades fisicoquímicas y biológicas.

2.3.3. Descriptores 2D o topológicos

Los IT son descriptores moleculares basados en la representación de un grafo de la molécula y representan propiedades de grafos teóricos que se conservan por isomorfismo, es decir, propiedades con valores idénticos para grafos isomórficos.

Un invariante del grafo puede ser un polinomio característico, una secuencia de números o un único índice numérico obtenido mediante la aplicación de operadores algebraicos a matrices que representan grafos moleculares y cuyos valores son independientes de la numeración o del etiquetado de los vértices.

Los IT generalmente se derivan de un grafo molecular suprimido de hidrógenos. Pueden ser sensibles a una o más características estructurales de la molécula, como el tamaño, la forma, la simetría, la ramificación y la ciclicidad, y también pueden codificar información química relacionada con el tipo de átomo y la multiplicidad de enlaces.

De hecho, los IT se suelen dividir en dos categorías: los índices topológicos y los índices estructurales[80]. Los índices topo-estructurales codifican solo información sobre la adyacencia y la distancia de los átomos en la estructura molecular; mientras que, los índices topo-químicos cuantifican la información en cuanto a topología, pero también a las propiedades químicas específicas de los átomos, como su identidad química y el estado de hibridación.

Los índices de información topológica son invariantes de grafos, basados en la teoría de la información y calculados como contenido de información de las relaciones de equivalencia especificadas en el grafo molecular.

En general, los IT no caracterizan de forma exclusiva la topología molecular; diferentes estructuras pueden tener iguales valores de IT. Una consecuencia de la falta de uniformidad de los IT es que, en general, no permiten la reconstrucción de las moléculas.

Hay varias formas de obtener descriptores topológicos. Los IT simples consisten en el conteo de algunos elementos del grafo específicos; ejemplos de estos son; el índice Z de Hosoya [81], conteo de caminos[82], conteos de los pasos, recuentos de pasos autorregresibles[83], descriptores de forma de Kier[84], índices de capa de caminos/pasos[85].

Sin embargo, los IT más comunes se obtienen aplicando algunos operadores algebraicos (p. ej. el operador de Wiener) a una representación matricial de la estructura molecular, como las matrices de adyacencia y distancia. Entre ellos se encuentran el índice de Wiener[86], índices espectrales[87], e índices de Harary[88].

Las matrices moleculares son las herramientas matemáticas más comunes para codificar información estructural de la molécula. Las más populares son las matrices teóricas de grafos, una gran cantidad de las cuales fueron propuestas en las últimas décadas para derivar IT y describir moléculas desde un punto de vista topológico.

Jánezic *et al.*[89] reportan una amplia colección de matrices teóricas de grafos. Las matrices de vértices son indudablemente las más frecuentemente

utilizadas para caracterizar un grafo molecular. Las entradas de dicha matriz codifican información diferente sobre pares de vértices, como sus conectividades, distancias topológicas, sumas de los pesos de los átomos a lo largo de los caminos de conexión. Las entradas diagonales pueden codificar información química sobre los vértices. A partir de las matrices de vértices, se propuso una gran cantidad de IT.

Se pueden obtener otros descriptores moleculares topológicos mediante el uso de funciones adecuadas aplicadas a invariantes de vértices locales (LOVI), que son representaciones numéricas de los átomos derivados de los grafos moleculares.

Las funciones más comunes son la suma de átomos y/o enlaces, lo que da como resultado a los descriptores que correlacionan bien las propiedades fisicoquímicas de los propios átomos y/o sumas de enlace. Por ejemplo, los índices de Zagreb[26], el índice de conectividad de Randiç[90], índices de conectividad de orden superior relacionados[91], y los índices de conectividad de distancia de Balaban[92] se derivan de acuerdo con este enfoque.

Los IT particulares se derivan de grafos moleculares ponderados donde los vértices y/o aristas se pesan por cantidades que representan características en 3D de la molécula. Los invariantes del grafo obtenidos de esta manera codifican tanto la información sobre topología molecular como de la geometría molecular. Los descriptores autovalores de la matriz de Burden (BCUT)[93] son un ejemplo de tales descriptores topológicos.

Los invariantes del grafo se han aplicado con éxito en la caracterización de similitud/diversidad en grandes conjuntos estructurales de moléculas y en el modelado QSPR-QSAR.

2.3.4. Descriptores 3D o geométricos

Otra clase de descriptores moleculares, llamados geométricos o 3D, se derivan de la representación geométrica de la molécula, es decir, de las coordenadas cartesianas (x, y, z) de los átomos de la molécula. La mayoría de los descriptores geométricos conocidos están aquí brevemente presentados.

Los descriptores invariantes holísticos ponderados (WHIM)[52] están basados en índices estadísticos calculados sobre las proyecciones de los átomos a lo largo de los ejes principales de la molécula. Están construidos de tal manera que capturan información molecular 3D relevante con respecto al tamaño molecular, forma, simetría y distribución de los átomos con respecto a los marcos de referencia invariantes. El algoritmo consiste en realizar un Análisis de Componentes Principales (PCA) en las coordenadas cartesianas centradas de una molécula, mediante el uso de una matriz de covarianza ponderada obtenida por diferentes esquemas de ponderación para los átomos. Para cada esquema de ponderación, se calcula un conjunto de índices estadísticos sobre los átomos proyectados en cada componente principal, es decir, la puntuación de cada uno.

Los índices gravitacionales[53] son descriptores geométricos que reflejan la distribución de masa en una molécula, definida como:

$$G_1 = \sum_{i=1}^{A-1} \sum_{j=i+1}^{A} \frac{m_i m_j}{r_{ij}^2} \tag{2.2}$$

$$G_2 = \sum_{b=1}^{B} \left(\frac{m_i m_j}{r_{ij}^2} \right)_b \tag{2.3}$$

donde m_i y m_j son las masas atómicas de los átomos considerados, r_{ij} las correspondientes distancias interatómicas, *A* y *B* el número de átomos y enlaces de la molécula, respectivamente. El índice G_1 tiene en cuenta todos los pares de átomos en la molécula, mientras que el índice G_2 está restringido a pares de átomos enlazados. Estos índices están relacionados con el cálculo de la cohesión del volumen de las moléculas.

Los descriptores de autovalores (EVA)[54] se propusieron para extraer información química estructural de los espectros del infrarrojo medio y cercano. El enfoque consiste en utilizar las frecuencias vibratorias de una molécula, una propiedad molecular fundamental caracterizada de forma fiable y sencilla a partir de la función de energía potencial. Las frecuencias se

pueden calcular utilizando métodos mecanocuánticos o métodos moleculares estándares de química computacional.

Los descriptores electrónicos de autovalores (EEVA)[56] son análogos a los descriptores EVA, pero en ellos se utilizan las energías de los orbitales moleculares semiempíricos, es decir, los autovalores de la ecuación de Schrödinger, en lugar de las frecuencias vibratorias de la molécula.

Los descriptores de representación molecular 3D de la estructura basada en difracción de electrones (3D-MoRSE)[55] se basan en la idea de obtener información de las coordenadas atómicas tridimensionales mediante la transformación usada en los estudios de difracción de electrones para preparar curvas de dispersión teóricas. La expresión derivada es la siguiente:

$$I(s) = \sum_{i=1}^{A-1} \sum_{j=i+1}^{A} \varpi_i \varpi_j \frac{\sin(sr_{ij})}{sr_{ij}} \qquad (2.4)$$

donde $I(s)$ es la intensidad de electrones dispersos; ϖ una propiedad atómica, por ejemplo, el número atómico; r_{ij} es la distancia interatómica entre los átomos *i*-ésimo y *j*-ésimo; y *A* es la cantidad de átomos.

Los descriptores de la función de distribución radial (RDF)[94] se basan en la distribución de distancias en la representación geométrica de una molécula y constituyen un código RDF que muestra ciertas características en común con los descriptores 3D-MoRSE.

Los descriptores de ensamblado de geometría, topología y pesos atómicos (GETAWAY)[57] se derivan de la matriz de influencia molecular (**H**), que es una representación de la estructura molecular, definida como:

$$\mathbf{H} = \mathbf{M}\left(\mathbf{M}^{\mathrm{T}}\mathbf{M}\right)^{-1}\mathbf{M}^{\mathrm{T}} \qquad (2.5)$$

donde **M** es la matriz molecular que presenta las coordenadas cartesianas centradas (x, y, z) de los átomos de la molécula en una conformación elegida.

Se supone que las coordenadas atómicas se calculan con respecto al centro de geometría de la molécula para obtener la invarianza traslacional. La matriz de influencia molecular es una matriz simétrica *A*x*A*, donde *A*

representa el número de átomos, y muestra la invarianza rotacional con respecto a las coordenadas de la molécula, lo que resulta independiente de las reglas de alineación.

Los elementos diagonales h_{ii} de la matriz tienen un rango que va de 0 a 1, y codifican información atómica relacionada con la "influencia" de cada átomo en la molécula y en la determinación de la forma molécular; en efecto, los átomos de la capa exterior siempre tienen valores h_{ii} más altos que los átomos más cercanos al centro de la molécula.

Los descriptores de GETAWAY se obtienen utilizando funciones de autocorrelación de doble ponderación, donde un esquema de ponderación es el parámetro de influencia y el otro una propiedad atómica (por ejemplo, la masa atómica).

Una representación geométrica implica el conocimiento de las posiciones relativas de los átomos en el espacio 3D. Por tanto, en el caso de estructuras químicas similares, los descriptores geométricos suelen proporcionar mayor información y capacidad de discriminación en comparación a los descriptores topológicos.

A pesar de su alto contenido de información, los descriptores geométricos generalmente muestran algunos inconvenientes. Requieren la optimización de la geometría y, por lo tanto, una sobrecarga en los procedimientos de cálculo. Además, para moléculas flexibles existen varias conformaciones de moléculas disponibles: por un lado, hay una gran cantidad de información disponible y que puede explotarse, pero, por otro lado, la complejidad del problema de cálculo aumenta significativamente.

Además, es importante recordar que rara vez se conoce la conformación biológica activa de los compuestos químicos estudiados. Algunos autores superan este problema utilizando un enfoque dinámico de multiconformación [78,79].

En conclusión, los descriptores indicadores basados en recuentos de fragmentos y otros descriptores simples suelen preferirse para el barrido de grandes bases de datos de conjuntos moleculares. Por el contrario, la

búsqueda de relaciones cuantitativas entre las estructuras moleculares y sus actividades biológicas, a menudo puede llevarse a cabo de manera más eficiente mediante el uso de descriptores geométricos, explotando su gran contenido de información.

2.3.5. Descriptores 4D o basados en grillas

Los enfoques GRID[58] y CoMFA[59] fueron los primeros métodos basados no solo en la estructura molecular sino en el cálculo de la energía de interacción entre la molécula y las diferentes pruebas bajo estudio. El objetivo de estos enfoques es identificar y caracterizar cuantitativamente las interacciones entre la molécula y el sitio activo del receptor.

Las moléculas se sitúan en una red tridimensional constituida por varios miles de puntos de una cuadrícula, espaciados uniformemente, y se utiliza una prueba de tipo estérica, electrostática, hidrofílica, etc., para mapear la superficie de la molécula sobre la base de la interacción de la molécula con dicha prueba.

Los modelos QSAR se obtienen mediante la aplicación del método de regresión de cuadrados mínimos parciales (PLS) a la matriz del campo de interacción. Debe notarse que el uso de los puntos de la grilla como descriptores moleculares requiere el cuidadoso paso de alinear las moléculas consideradas, de tal manera que cada uno de los miles de puntos de la grilla represente, para todas las moléculas, el mismo tipo de información y no información espuria debido a la falta de invarianza en la rotación de las moléculas en la red.

Además de los dos métodos más conocidos, GRID y CoMFA, otros métodos conocidos basados en este enfoque son CoMSIA[60], Compás[61], descriptores G-WHIM[62], Campo de análisis de Voronoi[63], SOMFA[95], descriptores VolSurf[64], y GRIND[65].

Aunque estos descriptores a menudo se denominan descriptores tridimensionales, se los puede denominar descriptores 4D (o descriptores basados en cuadrículas) porque a la información geométrica se agrega otra

fuente de información dada por la energía de interacción con una prueba específica. Los descriptores moleculares son los MIFs generados por las pruebas. Dichos campos escalares pueden visualizarse de manera eficiente y permiten pensar visualmente en nuevos candidatos a fármacos, lo que resulta muy útil en el proceso de descubrimiento de fármacos[96,97].

Una ventaja de estos enfoques es que los resultados finales muestran dónde y cómo modificar los compuestos para alcanzar los valores deseables de la propiedad molecular estudiada. Por el contrario, un inconveniente es la necesidad de una alineación molecular correcta para lograr la comparación molecular y la selección de la conformación más apropiada.

La alineación determina en qué medida los descriptores difieren de una molécula a la siguiente, por lo tanto, influye sustancialmente en los resultados de la evaluación.

Solo se pueden esperar resultados significativos y relevantes si la alineación se llevó a cabo de manera adecuada y sin ambigüedades. A menudo, la necesidad de una alineación limita la aplicación de ciertos descriptores a conjuntos de datos homogéneos, e incluso entonces la alineación no siempre se realiza fácilmente.

Como consecuencia, diferentes grupos de investigación comenzaron a desarrollar descriptores moleculares independientes de la alineación. El primer conjunto de descriptores basados en campos escalares pero independientes de la alineación fueron los descriptores G-WHIM[62], basados en los principios teóricos de los descriptores WHIM[52] pero aplicados a los MIF. Los descriptores de VolSurf[64] y GRIND[65] también son independientes de cualquier alineación previa de las moléculas.

2.4. Un breve repaso de la Teoría de Grafos

Las fórmulas químicas pueden ser vistas como grafos[98,99], es decir, como dos conjuntos de elementos no vacíos V y A. Los elementos V son llamados vértices y representan a los átomos. Los elementos A son los elementos denominados aristas; hay relaciones binarias entre los elementos

V, es decir, pares desordenados, y ellos simbolizan enlaces covalentes entre átomos.

A menos que se indique lo contrario, los átomos de hidrógeno serán ignorados, como hacen los químicos orgánicos cuando usualmente escriben un anillo de benceno como un hexágono; es decir, realizar la representación mediante un grafo sin hidrógenos (HSG).

Dos vértices conectados por una arista son llamados adyacentes, y un grafo singular puede ser descripto por su matriz de adyacencia **A**. Esta matriz tiene elementos a_{ij} iguales a 1 para vértices adyacentes i y j, y cero el caso contrario. Inicialmente, se llamó a esta "matriz de Hückel".

Un camino es una sucesión de aristas no repetidas en el cual no hay cortes entre las aristas. Mientras que un paso puede tener aristas repetidas. Los grafos químicos (molecular o constitucional) son grafos conectados, donde hay al menos un camino de un vértice a otro vértice del grafo.

La distancia (topológica) d_{ij} entre dos vértices i y j, es el número de aristas a lo largo del camino más corto entre estos vértices. El grafo singular es también determinado por su matriz de distancia **D**, cuyos elementos son distancias d_{ij}. Es fácil ver que las matrices **A** y **D** son simétricas con respecto a su diagonal principal, y que ellas comparten 1 en las respectivas posiciones adyacentes y ceros en la diagonal principal, pero todos los otros ceros de la matriz **A** son reemplazados en **D** por enteros mayores que 1.

La suma de las entradas sobre las filas o columnas de **A** son los grados del vértice v_i; ellos representan el número de vértices adyacentes al vértice i. Los químicos orgánicos usan esta terminología y llaman a estos, átomos de carbono primario, secundario, terciario y cuaternario, cuando ellos corresponden a vértices de grado 1, 2, 3, y 4, respectivamente. Similarmente, a los grados del vértice, uno puede definir el grado de la arista (a_i) como el número de todas las aristas adyacentes a una arista i dada.

Por supuesto, uno tiene que contar puntos finales de las aristas, y a estos conducirlos a las relaciones entre los grados de vértice y grado de aristas v_i y v_j de estos puntos finales:

$$a_i = v_i + v_j - 2 \tag{2.6}$$

La suma de cada elemento de una fila o columna de **D** es la llamada suma-distancia (s_i). A menor suma-distancia, el vértice i es más cercano al centroide del grafo.

Un circuito es un camino cuyo vértice inicial y final coinciden. Un grafo sin circuitos es llamado un árbol y se corresponde con un compuesto acíclico en química. El menor número de circuitos (es decir, el número más pequeño de enlaces que tienen que ser borrados para llegar a una estructura acíclica) es el número ciclomático (μ) del grafo.

Se puede mostrar que si a se refiere al número de aristas y v al número de vértices, se obtiene:

$$\mu = a - v + 1 \tag{2.7}$$

Los vértices de un grafo pueden ser etiquetados con números enteros de 1 a *n*, números de vértices (este número es también llamado el orden del grafo). Hay muchas posibles etiquetas, por lo tanto, muchas maneras diferentes en las que pueden aparecer las matrices **A** y **D**. Dos grafos son llamados isomorfos si hay una etiqueta de un grafo que preserva todas las adyacencias del otro.

2.5. Primera generación de índices topológicos

Harold Wiener[100,101] propuso entre 1947 y 1948 uno de los primeros descriptores moleculares de naturaleza topológica para hidrocarburos acíclicos saturados (alcanos). Wiener llamó a la suma del número de enlaces unidos a todos los pares de átomos como el número de caminos.

Este nombre tan original estaba equivocado ya que los caminos pueden tener longitudes variables en los grafos (una arista es un camino de longitud

unitaria). Hoy en día este descriptor molecular fue renombrado como el índice de Wiener (W). Hosoya demostró que W es la mitad de la suma de todas las entradas en la matriz de distancia (suma de valores suma-distancia):

$$W = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} d_{ij} = \frac{1}{2}\sum_{i=1}^{n} s_i \tag{2.8}$$

Numerosas fórmulas analíticas fueron avanzando para calcular W para varias clases de grafos como alcanos lineales, bencenoides, etc.[102–104]. Además, Wiener propuso el llamado número de polaridad p de un grafo, como el número de pares de vértices separados por tres aristas, o en otras palabras, el número de distancias de longitud tres en **D**. Ambos índices W y p fueron aplicados por Wiener y Platt para correlacionar la estructura de los alcanos con propiedades termodinámicas tales como el punto de ebullición, el calor de formación y vaporización, el volumen molecular y la refracción molar.

2.5.1. Hosoya y su índice topológico

En los años setenta, Haruo Hosoya propuso otro descriptor molecular denominado Z[105]. Su definición es:

$$Z = \sum_{k=0}^{[n/2]} p(G,k) \tag{2.9}$$

donde $p(\mathrm{G},\mathrm{k})$ es el número de maneras en las cuales las k aristas del grafo pueden ser elegidas de modo que dos de ellas no sean adyacentes, y el cuadrado medio de Gauss entre corchetes indica el menor entero que no excede el número incluido en estos corchetes. El número de a aristas iguales es $p(\mathrm{G},1)$ y por definición $p(\mathrm{G},0)$ es igual a 1.

Para árboles existe una relación simple entre Z y el valor absoluto de los coeficientes del polinomio característico del grafo (introduciendo la variable x en la diagonal principal de la matriz de adyacencia **A**, e igualando el determinante a cero).

Se demostró que el índice Z posee una excelente capacidad de correlación con muchas propiedades físicas, como el punto de ebullición.

2.5.2. Índice del grupo Zagreb

Alrededor de 1975, los químicos que habían estado interesados en las aplicaciones químicas de la teoría de grafos en la capital de Croacia, Zagreb, propusieron[106] dos ITs derivados de la matriz de adyacencia:

$$M_1 = \sum_{i=2}^{n} v_i^2 \tag{2.10}$$

$$M_2 = \sum_{todasaristas} v_i v_j \tag{2.11}$$

2.5.3. Índice Céntrico

El índice céntrico C desarrollado por Balaban en 1979[107] es medida de la forma molecular. El mismo utiliza un procedimiento conocido como partición de poda de átomos terminales, y se calcula de la siguiente manera:

$$C = \sum_{i} m_i^2 \tag{2.12}$$

donde m_i es el número de átomos eliminados en la etapa i. Este índice ha sido ampliamente utilizado en los estudios QSPR-QSAR.

2.5.4. Índice de Schultz

A pesar de que apareciera más tarde[108,109], el índice topológico molecular propuesto por H. P. Schultz (MTI) pertenece a la primera generación de índices topológicos:

$$MTI = \sum_{i=1}^{n} E_i \tag{2.13}$$

Donde E_i, son los componentes del vector $\mathbf{E} = \mathbf{c}(\mathbf{A} + \mathbf{D})$, con $\mathbf{A}$ la matriz de adyacencia, $\mathbf{D}$ la matriz de distancia, $\mathbf{c}$ es el vector de grados del vértice del grafo. El índice topológico molecular está bien definido solo para grafos

conectados, y es indeterminado para grafos desconectados con nodos aislados e infinito para todos los demás grafos desconectados.

El índice *MTI*, tiene una degeneración bastante baja. Varios grupos de investigación encontraron que el índice de Schultz y el índice de Wiener están linealmente correlacionados en muchos casos[110].

Pudo demostrarse que el índice de Schultz correlaciona bien con el calor de formación de hidrocarburos bencenoides[111].

2.6. Segunda generación de índices topológicos

2.6.1. Índice de conectividad molecular de Randić

Un índice topológico muy popular y ampliamente utilizado hoy día, es el índice de conectividad molecular (χ) introducido por Randić[112] en 1975, el cual da inicio a la segunda generación de ITs:

$$\chi = \sum_{todasaristas} \left(v_i v_j\right)^{-1/2} \qquad (2.14)$$

El índice χ no sólo posee baja degeneración, sino que su capacidad de correlación es bastante buena para muchas propiedades físicas y bioquímicas. Debido a su definición, el grafo más compacto es el que presenta el menor valor de χ.

2.6.2. Índices de la Teoría de la Información

Tomando en cuenta que la degeneración de los ITs de primera y segunda generación es originada en parte a que los números son obtenidos como resultado de sumatorias, Bonchev y Trinajstić[113] aplicaron la fórmula de Shannon para los sumandos, obteniendo por lo tanto ITs de menor grado de degeneración.

Así, en lugar de adicionar todas las distancias d_{ij} del grafo para obtener el índice de Wiener, uno ordena primero estas distancias de acuerdo a sus valores dentro de grupos de sumandos g_i que tienen el mismo valor i. El índice de Wiener queda como:

$$W = \sum_{i=1}^{l} ig_i \qquad (2.15)$$

donde l es el elemento más grande de la matriz de distancia **D**, o el diámetro molecular del grafo.

2.6.3. Índice de conectividad de Balaban

Un inconveniente de muchos ITs a remediar es principalmente el hecho de que, con el incremento del número de vértices o ciclos, los ITs también se incrementan. Por consiguiente, se incluye dentro del IT no sólo la información de la "forma topológica" de la molécula, sino también de su tamaño o el número de ciclos al mismo tiempo.

Por dicho motivo Balaban[114] modificó la fórmula del índice de Randić, reemplazando los grados de vértices por suma-distancias, e introdujo un factor de corrección que incluye el número ciclomático para obtener el índice de conectividad J:

$$J = \frac{e}{\mu + 1} \sum_{todasaristas} (s_i s_j)^{-1/2} \qquad (2.16)$$

donde e es el número de aristas del grafo y la sumatoria es sobre todos los pares de aristas adyacentes i, j. La degeneración de este índice es mucho menor que los otros ITs[115]. El grafo más compacto es el que posee mayor valor de J.

A diferencia de los ITs basados en la matriz de adyacencia, los ITs basados en la matriz de distancia pueden codificar información de la multiplicidad de enlaces.

2.6.4. Índice topológico del cuadrado medio de la distancia

Un grafo compacto tiene menos distancias que un grafo alargado con el mismo número de vértices. En base a esto, Balaban[107] propuso que el promedio de la distancia llevaba a ITs razonables::

$$D^2 = \left[\frac{\sum_{i=1}^{l} i^2 g_i}{\sum_{i=1}^{l} g_i} \right]^{1/2} \qquad (2.17)$$

Para grafos acíclicos, en lugar de promediar todas las distancias en **D**, se puede considerar en la fórmula anterior sólo los puntos finales (vértices de grado uno). Ambos índices decrecen con el incremento de la ramificación.

Estos índices dan buenas correlaciones entre el número de octanos de alcanos y su estructura. El grafo más compacto posee el menor valor del cuadrado medio de la distancia.

2.6.5. Índice de Rücker

Rücker y Rücker[116] presentaron una familia de descriptores topológicos: el recuento de todos los caminos de mayor longitud (todos los caminos a diferencia de los caminos que regresan hacia ellos mismos).

Estos descriptores, aunque se calculan de manera extremadamente fácil mediante un procedimiento de suma de tipo Morgan, exhiben un alto poder de discriminación entre los isómeros y entre la situación estructural de los átomos.

El índice topológico *twc* (recuento total de caminos) es de menor degeneración que el índice de conectividad de Balaban, y un código para grafos basado en todos los caminos parece ser más discriminante que el índice OSC de Trinajstić (código estructural ordenado). Se utiliza con éxito como variable

alternativa en la descripción de datos de ^{13}C en RMN para alcanos usuales y altamente ramificados.

2.6.6. Índice del estado electrotopológico

Kier y Hall introdujeron nuevos ITs llamados índices de estado electrotopológico (estado-E) basados en invariantes del grafo para cada átomo en la molécula. El valor del estado-E codifica el estado electrónico intrínseco del átomo influenciado por el ambiente electrónico de todos los otros átomos dentro de la estructura molecular[117–119].

2.7. Tercera generación de índices topológicos

La tercera generación de ITs puede ser considerada como el inicio de la definición del "número real invariante de vértice local" (LOVI) aplicado a la resolución de ecuaciones lineales obtenidas de una matriz simétrica, tales como **A** o **D**, por sustitución de ceros en la diagonal principal con un vector, y por interpretación de la matriz resultante como coeficientes de un sistema de ecuaciones lineales teniendo como valor desconocido el LOVI[120-122].

2.7.1. Índice de Diudea

Mircea V. Diudea introdujo[123] un índice basado en la matriz de Cluj, una matriz no simétrica $n \times n$ la cual puede ser usada para construir nuevos ITs. Cuando se usan valores de distancia recíproca aumenta el número de invariantes altamente discriminantes.

2.7.2. Índice Hiper-Wiener

El índice Hiper-Wiener fue definido por Randić para grafos acíclicos[124], considerando los caminos en los grafos. Se debe notar, sin embargo, que los tiempos de cómputo requeridos para caminos y pasos son considerablemente mayores que para calcular distancias.

Diudea[125–127] empleó este índice Hiper-Wiener para dendrímeros. Posteriormente, se hicieron intentos para extender esta definición a grafos cíclicos: Klein, Lukovits y Gutman[128–131] dieron una definición para el índice Hiper-Wiener *R* en términos de distancia d_{ij} entre vértices i y j que puede ser aplicable a cualquier grafo:

$$R = \frac{1}{2}\sum_{i \le j} d_{ij}\left(d_{ij}+1\right) \qquad (2.18)$$

2.8. Índices con información de estereoisomerismo

Randić introdujo un índice de conectividad 3D; análogamente, un número de Wiener 3D fue definido por Trinajstić y colaboradores, basado en distancias geométricas entre pares de átomos[132–134]. Las geometrías son optimizadas por ejemplo mediante cálculos de mecánica molecular. La flexibilidad molecular puede ser expresada numéricamente por un índice propuesto por Kier, el índice de forma-k[135].

Tal vez, el reto más difícil para los químicos y matemáticos de la teoría de grafos es tratar con enantiómeros y con compuestos poliquirales diasteroisómeros, porque la distancia entre átomos no-enlazados ya no puede usarse para distinguir tales estereoisomerismos. Sin embargo, algunos avances han sido llevados a cabo con la propuesta de Schultz, Schultz y Schultz[136,137], por inclusión de la información de ambos diasteroisomerismo y enantiomerismo usando matrices pesadas.

La multiplicación de estos índices topológicos causó preocupación en una parte de la comunidad científica, por el hecho de que el significado físico de estos descriptores no estaba del todo claro. Sin embargo, cuando se demostró que muchos ITs estaban interrelacionados, y que ellos estaban también correlacionados con el volumen molar, la situación se apaciguó.

Varios investigadores incluyeron ITs entre sus descriptores numéricos en estudios QSPR-QSAR que correlacionan propiedades fisicoquímicas. Cuando una actividad biológica es estudiada, otros descriptores contribuyen sustancialmente (junto con ITs)[138].

2.9. Descriptores flexibles

CORAL (Correlación y lógica, SMILES, Electrones, Átomos)[139] es un programa de libre acceso empleado para la construcción de modelos de la Teoría QSPR-QSAR, y que permite el cálculo de descriptores flexibles, es decir, descriptores moleculares cuyos valores numéricos son dependientes no solo de la estructura química sino también de la propiedad/actividad estudiada[140-142].

De hecho, la gran mayoría de los descriptores convencionales no suelen correlacionar bien con una cierta propiedad/actividad bajo estudio. Sin embargo, el descriptor flexible u óptimo puede ser dirigido para correlacionar con cualquier propiedad/actividad a través de la optimización de sus partes variables, que se vuelven específicas a la base de datos analizada[143-145].

Este tipo de descriptores ajustables podrían combinarse de manera plausible con los descriptores topológicos antes mencionados, de manera de mejorar la calidad predictiva de los modelos.

Un descriptor flexible de CORAL se calcula a través de la optimización del peso de correlación *CW* (parte variable) correspondiente a cada invariante local o atributo estructural (SA) que lo define, para lo cual se emplea el método de simulación Monte Carlo que maximiza la correlación lineal entre el descriptor flexible y la propiedad/actividad estudiada.

CORAL permite emplear como representación estructural las siguientes opciones: HSG, HFG y SMILES, a partir de las cuales es posible obtener diferentes invariantes locales. Hay varios atributos estructurales que están disponibles en el programa CORAL, entre los cuales se encuentran los vértices y los índices de conectividad extendida de Morgan. En el caso de HSG y HFG, los vértices son la representación de los elementos químicos presentes en la estructura molecular, tales como carbono, nitrógeno, oxígeno, etc.

A continuación se brinda un ejemplo de construcción del descriptor flexible.

$$DCW(SA) = \sum_i CW(SA_i) \qquad (2.19)$$

donde SA_i es el atributo estructural correspondiente al vértice i. El descriptor flexible así obtenido depende además de 2 parámetros de optimización: i) el umbral T (número entero que suele variar entre 0 y 6) y el número de iteraciones óptimas N^{iter} del método de simulación Monte Carlo (MC). La selección apropiada de los valores de T y N^{iter} evita el sobreajuste del modelo en el conjunto de calibración.

Cabe destacar que la principal ventaja en el empleo de un descriptor de naturaleza flexible basado en la Teoría de Grafos es que este no dependerá de los aspectos conformacionales de las estructuras químicas.

Bibliografía

1. Baran, E. J. Mean amplitudes of vibration of the pentagonal pyramidal XeOF5− and IOF52− anions. *J. Fluor. Chem.* **101,** 61–63 (2000).
2. Nicholis, G. & Prigogine, I. La estructura de lo complejo. En el camino hacia una nueva comprensión de las ciencias. *Editor. Alianza SA Madrid, España* (1994).
3. Rada, E. La polémica Leibniz-Clarke. (1980).
4. Mitchell, M. *Complexity: A guided tour*. (Oxford University Press, 2009).
5. Maldonado, C. E. & Gómez-Cruz, N. A. Synchronicity among biological and computational levels of an organism: quantum biology and complexity. *Procedia Comput. Sci.* **36,** 177–184 (2014).
6. Maldonado, C. E. ¿ Qué es un sistema complejo? *Rev. Colomb. Filos. la Cienc.* **14,** 71–93 (2014).
7. Crum-Brown, A. On the Theory of Isomeric Compounds. *Trans. R. Soc. Edinb.* **23,** 707–719 (1864).
8. Crum-Brown, A. On an Application of Mathematics to Chemistry. *Proc. R. Soc.* **VI (73),** 89–90 (1866).
9. Crum-Brown, A; Fraser, T. R. On the Connection between Chemical Constitution and Physiological Action. Part 1. On the Physiological Action of Salts of the Ammonium Bases, Derived from Strychnia, Brucia, Thebia, Codeia, Morphia and Nicotia. *Trans. R. Soc. Edinb.* **25,** 151–203 (1868).
10. Körner, W. Studi sulla Isomeria delle Cosi Dette Sostanze Aromatiche a Sei Atomi di Carbonio. *Gazz. Chim. Ital.* **4,** 242 (1874).
11. Mills, E. J. On Melting Point and Boiling Point as Related to Composition. *Philos. Mag.* **17,** 173–187 (1884).
12. Richet, M. C. Notè sur la Rapport entre la Toxicitè et les Propriètès Physiques des Corps. *Compt. Rend. Soc. Biol.* **45,** 775–776. (1893).
13. Meyer, H. Zur Theorie der Alkoholnarkose. *Arch. Exp. Pathol. Pharmacol.* **42,**

109–118 (1899).

14. Overton, E. Studien über die Narkose, zugleich ein Beitrag zur allgemeinen Pharmakologie. *Verlag Gustav Fischer Jena, Ger.* 141 (1901).
15. Traube, I. Theorie der Osmose und Narkose. *Arch. für die ges. Physiol.* **105,** 541–558 (1904).
16. Wiener, H. J. Influence of Interatomic Forces on Paraffin Properties. *Chem. Phys.* **15,** 766. (1947).
17. Platt, J. R. Influence of Neighbor Bonds on Additive Bond Properties in Paraffins. *J. Chem. Phys.* **15,** 419–420 (1947).
18. Fujita, T.; Iwasa, J.; Hansch, C. A New Substituent Constant, π, Derived from Partition Coefficients. *J. Am. Chem. Soc.* **86,** 5175–5180. (1964).
19. Gordon, M.; Scantlebury, G. R. Non-Random Polycondensation: Statistical Theory of the Substitution Effect. *Trans. Faraday Soc.* **60,** 604–621 (1964).
20. Smolenskii, E. A. Application of the Theory of Graphs to Calculations of the Additive Structural Properties of Hydrocarbons. *Russ. J. Phys. Chem.* **38,** 700–702 (1964).
21. Spialter, L. The Atom Connectivity Matrix (ACM) and Its Characteristic Polynomial (ACMCP). *J. Chem. Doc.* **4,** 261–269 (1964).
22. Balaban, A. T.; Harary, F. Chemical Graphs. V. Enumeration and Proposed Nomenclature of Benzenoid Catacondensed Polycyclic Aromatic Hydrocarbons. *Tetrahedron* **24,** 2505–2516 (1968).
23. Harary, F. Graph Theory. *Addison-Wesley Reading, MA* 1969
24. Kier, L. B. Molecular Orbital Theory in Drug Research. *Acad. Press New York, NY* (1971).
25. Cammarata, A. Interrelationship of the Regression Models Used for Structure–Activity Analyses. *J. Med. Chem.* **15,** 573–577 (1972).
26. Gutman, I.; Trinajstic, N. Graph Theory and Molecular Orbitals. Total π-Electron Energy of Alternant Hydrocarbons. *Chem. Phys. Lett.* **17,** 535–538 (1972).
27. Hosoya, H. Topological Index as a Sorting Device for Coding Chemical Structures. *J. Chem. Doc.* **12,** 181–183 (1972).
28. Pauling, L. The Additivity of the Energies of Normal Covalent Bonds. *Proc. Natl. Acad. Sci. USA* **14,** 414–416 (1932).
29. Pauling, L. The Nature of the Chemical Bond. *Cornell Univ. Press Ithaca, NY* (1939).
30. Coulson, C. A. The Electronic Structure of Some Polyenes and Aromatic Molecules. VII. Bonds of Fractional Order by the Molecular Orbital Method. *Proc. R. Soc. London A* **169,** 413–428 (1939).
31. Sanderson, R. T. Electronegativity I. Orbital Electronegativity of Neutral Atoms. *J. Chem. Educ.* **29,** 540–546 (1952).
32. Fukui, K.; Yonezawa, Y.; Shingu, H. Theory of Substitution in Conjugated Molecules. *Bull. Chem. Soc. Jpn.* **27,** 423–427 (1954).
33. Mulliken, R. S. Electronic Population Analysis on LCAO-MO Molecular Wave Functions. *I. J. Chem. Phys.* **23,** 1833–1840 (1955).

34. Hammett, L. P. Reaction Rates and Indicator Acidities. *Chem. Rev.* **17,** 67–79

35. Hammett, L. P. The Effect of Structure upon the Reactions of Organic Compounds. Benzene Derivatives. *J. Am. Chem. Soc.* **59,** 96–103. (1937).

36. Hammett, L. P. Linear Free Energy Relationships in Rate and Equilibrium Phenomena. *Trans. Faraday Soc.* **34,** 156–165 (1938).

37. Taft, R. W. Polar and Steric Substituent Constants for Aliphatic and o-Benzoate Groups from Rates of Esterification and Hydrolysis of Esters. *J. Am. Chem. Soc.* **74,** 3120–3128 (1952).

38. Taft, R. W. The General Nature of the Proportionality of Polar Effects of Substituent Groups in Organic Chemistry. *J. Am. Chem. Soc.* **75,** 4231–4238 (1953).

39. Taft, R. W. Linear Steric Energy Relationships. *J. Am. Chem. Soc.* **75,** 4538–4539 (1953).

40. Hansch, C.; Maloney, P. P.; Fujita, T.; Muir, R. M. Correlation of Biological Activity of Phenoxyacetic Acids with Hammett Substituent Constants and Partition Coefficients. *Nature* **194,** 178–180 (1962).

41. Hansch, C.; Muir, R. M.; Fujita, T.; Maloney, P. P.; Geiger, F.; Streich, M. The Correlation of Biological Activity of Plant Growth Regulators and Chloromycetin Derivatives with Hammett Constants and Partition Coefficients. *J. Am. Chem. Soc.* **85,** 2817–2824 (1963).

42. Hansch, C.; Leo, A. *Exploring QSAR. Fundamentals and Applications in Chemistry and Biology*. *American Chemical Society: Washington, DC,* (1995).

43. Free, S. M.; Wilson, J. W. A. Mathematical Contribution to Structure–Activity Studies. *J. Med. Chem.* **7,** 395–399 (1964).

44. Kubinyi, H. Free Wilson Analysis. Theory, Applications and Its Relationship to Hansch Analysis. *Quant. Struct. -Act. Relat.* **7,** 121–133 (1988).

45. Balaban, A. T.; Harary, F. The Characteristic Polynomial Does Not Uniquely Determine the Topology of a Molecule. *J. Chem. Doc.* **11,** 258–259 (1971).

46. Balaban, A. T. Ed. Chemical Applications of Graph Theory. *Acad. Press New York, NY* 390 (1976).

47. Randic, M. On the Recognition of Identical Graphs Representing Molecular Topology. *J. Chem. Phys.* **60,** 3920–3928 (1974).

48. Randic, M. On Characterization of Molecular Branching. *J. Am. Chem. Soc.* **97,** 6609–6615 (1975).

49. Kier, L. B.; Hall, L. H.; Murray, W. J.; Randic, M. Molecular Connectivity. I: Relationship to Nonspecific Local Anesthesia. *J. Pharm. Sci.* **64,** 1971–1974 (1975).

50. Rohrbaugh, R. H.; Jurs, P. C. Descriptions of Molecular Shape Applied in Studies of Structure/Activity and Structure/Property Relationships. *Anal. Chim. Acta* **199,** 99–109 (1987).

51. Stanton, D. T.; Jurs, P. C. Development and Use of Charged Partial Surface Area Structural Descriptors in Computer-Assisted Quantitative Structure–Property Relationship Studies. *Anal. Chem.* **62,** 2323–2329 (1990).

52. Todeschini, R.; Lasagni, M.; Marengo, E. New Molecular Descriptors for 2D- and 3D-Structures, Theory. *J. Chemom.* **8,** 263–273 (1994).

53. Katritzky, A. R.; Mu, L.; Lobanov, V. S.; Karelson, M. Correlation of Boiling Points with Molecular Structure. 1. A Training Set of 298 Diverse Organics and a Test Set of 9 Simple Inorganics. *J. Phys. Chem.* **100,** 10400–10407 (1996).

54. Ferguson, A. M.; Heritage, T. W.; Jonathon, P.; Pack, S. E.; Phillips, L.; Rogan, J.; Snaith, P. J. EVA: A New Theoretically Based Molecular Descriptor for Use in QSAR/QSPR Analysis. *J. Comput. Aided Mol. Des.* **11,** 143–152 (1997).

55. Schuur, J.; Selzer, P.; Gasteiger, J. The Coding of the Three-Dimensional Structure of Molecules by Molecular Transforms and Its Application to Structure-Spectra Correlations and Studies of Biological Activity. *J. Chem. Inf. Comput. Sci.* **36,** 334–344 (1996).

56. Tuppurainen, K. EEVA (Electronic Eigenvalue): A New QSAR/QSPR Descriptor for Electronic Substituent Effects Based on Molecular Orbital Energies. *SAR QSAR Environ. Res.* **10,** 39–46 (1999).

57. Consonni, V.; Todeschini, R.; Pavan, M. Structure/Response Correlations and Similarity/Diversity Analysis by GETAWAY Descriptors. Part 1. Theory of the Novel 3D Molecular Descriptors. *J. Chem. Inf. Comput. Sci.* **42,** 682–692 (2002).

58. Goodford, P. J. A Computational Procedure for Determining Energetically Favorable Binding Sites on Biologically Important Macromolecules. *J. Med. Chem.* **28,** 849–857 (1985).

59. Cramer, R. D. III; Patterson, D. E.; Bunce, J. D. Comparative Molecular Field Analysis (CoMFA). 1. Effect of Shape on Binding of Steroids to Carrier Proteins. *J. Am. Chem. Soc.* **110,** 5959–5967 (1988).

60. Klebe, G.; Abraham, U.; Mietzner, T. Molecular Similarity Indices in a Comparative Analysis (CoMSIA) of Drug Molecules to Correlate and Predict Their Biological Activity. *J. Med. Chem.* **37,** 4130–4146 (1994).

61. Jain, A. N.; Koile, K.; Chapman, D. Compass: Predicting Biological Activities from Molecular Surface Properties. Performance Comparisons on a Steroid Benchmark. *J. Med. Chem.* **37,** 2315–2327 (1994).

62. Todeschini, R.; Moro, G.; Boggia, R.; Bonati, L.; Cosentino, U.; Lasagni, M.; Pitea, D. Modeling and Prediction of Molecular Properties. Theory of Grid-Weighted Holistic Invariant Molecular (G-WHIM) Descriptors. *Chemom. Intell. Lab. Syst.* **36,** 65–73 (1997).

63. Chuman, H.; Karasawa, M.; Fujita, T. A Novel 3-Dimensional QSAR Procedure – Voronoi Field Analysis. *Quant. Struct. -Act. Relat.* **17,** 313–326 (1998).

64. Cruciani, G.; Pastor, M.; Guba, W. VolSurf: A New Tool for the Pharmaceutic Optimization of Lead Compounds. *Eur. J. Pharm. Sci.* **11 (Suppl.,** S29–S39 (2000).

65. Pastor, M.; Cruciani, G.; McLay, I. M.; Pickett, S. D.; Clementi, S. GRid-INdependent Descriptors (GRIND): A Novel Class of Alignment-Independent Three-Dimensional Molecular Descriptors. *J. Med. Chem.* **43,** 3233–3243 (2000).

66. Gasteiger, J. Handbook of Chemoinformatics. From Data to Knowledge in 4 Volumes. *Wiley-VCH Weinheim, Ger.* 1870 pp. (2003).

67. Oprea, T. 3D QSAR Modeling in Drug Design. in *In Computational Medicinal Chemistry for Drug Discovery* (eds. I. Bultinck, P., D. W. & H., Langenaeker, W., Tollenaere, J. P.) pp 571–616 (2004).

68. Kubinyi, H. QSAR in Drug Design. in *In Handbook of Chemoinformatics* (ed.

Gasteiger, J.) pp 1532–1554. (Ed.; Wiley-VCH: Weinheim, 2003).

69. Todeschini, R.; Consonni, V. Handbook of Molecular Descriptors. *Wiley-VCH Weinheim, Ger.* 668 pp. (2000).

70. Mezey, P. G. *Shape in chemistry: an introduction to molecular shape and topology*. (Wiley-VCH, 1993).

71. Pearlman, R. S. Rapid generation of high quality approximate 3D molecular structures. *Chem. Des. Auto. News* **2,** 5–6 (1987).

72. Burden, F. R. Molecular identification number for substructure searches. *J. Chem. Inf. Comput. Sci.* **29,** 225–227 (1989).

73. Cramer, R. D., Patterson, D. E. & Bunce, J. D. Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. *J. Am. Chem. Soc.* **110,** 5959–5967 (1988).

74. Testa, B.; Kier, L. B. The Concept of Molecular Structure in Structure–Activity Relationship Studies and Drug Design. *Med. Res. Rev.* **11,** 35–48 (1991).

75. Jurs, P. C.; Dixon, J. S.; Egolf, L. M. van de & Waterbeemd, H. Representations of Molecules. In Chemometrics Methods in Molecular Design. *Ed.; VCH Publ. New York, NY* **Vol. 2,** pp 15–38 (1995).

76. Smith, E. G.; Baker, P. A. The Wiswesser Line-Formula Chemical Notation (WLN). *Chem. Inf. Manag. Cherry Hill, NJ* (1975).

77. Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **28,** 31–36 (1988).

78. Mekenyan, O.; Ivanov, J.; Veith, G. D.; Bradbury, S. P. Dynamic QSAR: A New Search for Active Conformations and Significant Stereoelectronic Indices. *Quant. Struct. -Act. Relat.* **13,** 302–307 (1994).

79. Mekenyan, O.; Nikolova, N.; Schmieder, P. Dynamic 3D QSAR Techniques: Applications in Toxicology. *J. Mol. Struct.* **622,** 147–165 (2003).

80. Basak, S. C.; Gute, B. D.; Grunwald, G. D. Use of Topostructural, Topochemical, and Geometric Parameters in the Prediction of Vapor Pressure: A Hierarchical QSAR Approach. *J. Chem. Inf. Comput. Sci.* **37,** 651–655 (1997).

81. Hosoya, H. Topological Index. A Newly Proposed Quantity Characterizing the Topological Nature of Structural Isomers of Saturated Hydrocarbons. *Bull. Chem. Soc. Jpn.* **44,** 2332–2339 (1971).

82. Randic, M.; Wilkins, C. L. Graph Theoretical Ordering of Structures as a Basis for Systematic Searches for Regularities in Molecular Data. *J. Phys. Chem.* **83,** 1525–1540 (1979).

83. Harary, F. Graph Theory. *Addison-Wesley Reading, MA* (1969).

84. Kier, L. B. A Shape Index from Molecular Graphs. *Quant. Struct. -Act. Relat.* **4,** 109–116 (1985).

85. Randic, M. Novel Shape Descriptors for Molecular Graphs. *J. Chem. Inf. Comput. Sci.* **41,** 607–613 (2001).

86. Wiener, H. Structural Determination of Paraffin Boiling Points. *J. Am. Chem. Soc.* **69,** 17–20 (1947).

87. Ivanciuc, O.; Balaban, A. T. The Graph Description of Chemical Structures. in *In Topological Indices and Related Descriptors in QSAR and QSPR* (ed. Devillers,

J., Balaban, A. T., E. .) 59–167 pp (Gordon and Breach Science Publishers: Amsterdam, 1999).

88. Ivanciuc, O.; Balaban, T.-S.; Balaban, A. T. Design of Topological Indices. Part 4. Reciprocal Distance Matrix, Related Local Vertex Invariants and Topological Indices. *J. Math. Chem.* **12,** 309–318 (1993).

89. Janezic, D.; Milicevic, A.; Nikolic, S.; Trinajstic, N. Graph Theoretical Matrices in Chemistry. *Univ. Kragujev. Kragujevac, Serbia* ; 205 pp. (2007).

90. Randic, M. Graph Theoretical Approach to Local and Overall Aromaticity of Benzenoid Hydrocarbons. *Tetrahedron* **31,** 1477–1481 (1975).

91. Kier, L. B.; Hall, L. H. The Nature of Structure–Activity Relationships and Their Relation to Molecular Connectivity. *Eur. J. Med. Chem.* **12,** 307–312 (1977).

92. Balaban, A. T. Highly Discriminating Distance-Based Topological Index. *Chem. Phys. Lett.* **89,** 399–404 (1982).

93. Burden, F. R. A Chemically Intuitive Molecular Index Based on the Eigenvalues of a Modified Adjacency Matrix. *Quant. Struct. Act. Relat.* **16,** 309–314 (1997).

94. Raevsky, O. A.; Trepalin, S. V.; Razdol'skii, A. N. New QSAR Descriptors Calculated from Interatomic Interaction Spectra. *Pharm. Chem. J.* **34,** 646–649 (2000).

95. Robinson, D. D.; Winn, P. J.; Lyne, P. D.; Richards, W. G. Self-Organizing Molecular Field Analysis: A Tool for Structure–Activity Studies. *J. Med. Chem.* **42,** 573–583 (1999).

96. Buolamwini, J. K.; Assefa, H. CoMFA and CoMSIA 3D QSAR and Docking Studies on Conformationally-Restrained Cinnamoyl HIV-1 Integrase Inhibitors: Exploration of a Binding Mode at the Active Site. *J. Med. Chem.* **45,** 841–852 (2002).

97. Xu, M.; Zhang, A.; Han, S.; Wang, L.-S. Studies of 3D-Quantitative Structure–Activity Relationships on a Set of Nitroaromatic Compounds: CoMFA, Advanced CoMFA and CoMSIA. *Chemosphere* **48,** 707–715 (2002).

98. Harary, F. Graph Theory, 2nd printing. (1971).

99. Balaban, A. T. *Chemical applications of graph theory*. (Academic Press, 1976).

100. Wiener, H. Structural determination of paraffin boiling points. *J. Am. Chem. Soc.* **69,** 17–20 (1947).

101. Wiener, H. Correlation of heats of isomerization, and differences in heats of vaporization of isomers, among the paraffin hydrocarbons. *J. Am. Chem. Soc.* **69,** 2636–2638 (1947).

102. Zhu, H.-Y., Klein, D. J. & Lukovits, I. Extensions of the Wiener number. *J. Chem. Inf. Comput. Sci.* **36,** 420–428 (1996).

103. Dobrynin, A. A. Formula for calculating the Wiener index of catacondensed benzenoid graphs. *J. Chem. Inf. Comput. Sci.* **38,** 811–814 (1998).

104. Nikolić, S. & Trinajstić, N. The Wiener index: Development and applications. *Croat. Chem. Acta* **68,** 105–129 (1995).

105. Hosoya, H. Topological index. A newly proposed quantity characterizing the topological nature of structural isomers of saturated hydrocarbons. *Bull. Chem. Soc. Jpn.* **44,** 2332–2339 (1971).

106. Gutman, I., Ruščić, B., Trinajstić, N. & C. F. Wilcox, J. Graph theory and

molecular orbitals. XII. Acyclic polyenes. *J. Chem. Phys.* **62,** 3399–3405 (1975).

107. Balaban, A. T. Chemical graphs. *Theor. Chim. Acta* **53,** 355–375 (1979).

108. Schultz, H. P. Topological organic chemistry. 1. Graph theory and topological indices of alkanes. *J. Chem. Inf. Comput. Sci.* **29,** 227–228 (1989).

109. Schultz, H. P., Schultz, E. B. & Schultz, T. P. Topological organic chemistry. 2. Graph theory, matrix determinants and eigenvalues, and topological indexes of alkanes. *J. Chem. Inf. Comput. Sci.* **30,** 27–29 (1990).

110. Klavžar, S. & Gutman, I. A comparison of the Schultz molecular topological index with the Wiener index. *J. Chem. Inf. Comput. Sci.* **36,** 1001–1003 (1996).

111. Cash, G. G. Heats of Formation of Polyhex Polycyclic Aromatic Hydrocarbons from Their Adjacency Matrixes. *J. Chem. Inf. Comput. Sci.* **35,** 815–818 (1995).

112. Randic, M. Characterization of molecular branching. *J. Am. Chem. Soc.* **97,** 6609–6615 (1975).

113. Bonchev, D. & Trinajstić, N. Information theory, distance matrix, and molecular branching. *J. Chem. Phys.* **67,** 4517–4533 (1977).

114. Balaban, A. T. Highly discriminating distance-based topological index. *Chem. Phys. Lett.* **89,** 399–404 (1982).

115. Balaban, A. T., Ionescu-Pallas, N. & Balaban, T. S. Asymptotic Values of Topological Indices J and J′(Average Distance Sum Connectivities) for Infinite Acyclic and Cyclic Graphs. *MATCH Commun. Math. Comput. Chem* **17,** 121–146 (1985).

116. Ruecker, G. & Ruecker, C. Counts of all walks as atomic and molecular descriptors. *J. Chem. Inf. Comput. Sci.* **33,** 683–695 (1993).

117. Kier, L. B. & Hall, L. H. An electrotopological-state index for atoms in molecules. *Pharm. Res.* **7,** 801–807 (1990).

118. Hall, L. H. & Kier, L. B. Electrotopological state indices for atom types: a novel combination of electronic, topological, and valence state information. *J. Chem. Inf. Comput. Sci.* **35,** 1039–1045 (1995).

119. Hall, L. H. & Story, C. T. Boiling point and critical temperature of a heterogeneous data set: QSAR with atom type electrotopological state indices using artificial neural networks. *J. Chem. Inf. Comput. Sci.* **36,** 1004–1014 (1996).

120. Filip, P. A., Balaban, T.-S. & Balaban, A. T. A new approach for devising local graph invariants: derived topological indices with low degeneracy and good correlation ability. *J. Math. Chem.* **1,** 61–83 (1987).

121. Ivanciuc, O., Balaban, T.-S. & Balaban, A. T. Chemical graphs with degenerate topological indices based on information on distances. *J. Math. Chem.* **14,** 21–33 (1993).

122. Balaban, A. T. & Catana, C. Search for nondegenerate real vertex invariants and derived topological indexes. *J. Comput. Chem.* **14,** 155–160 (1993).

123. Diudea, M. V. Cluj matrix invariants. *J. Chem. Inf. Comput. Sci.* **37,** 300–305 (1997).

124. Randic, M., Guo, X., Oxley, T. & Krishnapriyan, H. Wiener matrix: Source of novel graph invariants. *J. Chem. Inf. Comput. Sci.* **33,** 709–716 (1993).

125. Diudea, M. V. Wiener and hyper-Wiener numbers in a single matrix. *J. Chem.*

Inf. Comput. Sci. **36,** 833–836 (1996).

126. Diudea, M. V. Molecular topology. 21. Wiener index of dendrimers. *MATCH-COMMUNICATIONS Math. Comput. Chem.* 71–83 (1995).

127. Diudea, M. V & Parv, B. Molecular topology. 25. Hyper-Wiener index of dendrimers. *J. Chem. Inf. Comput. Sci.* **35,** 1015–1018 (1995).

128. Klein, D. J., Lukovits, I. & Gutman, I. On the definition of the hyper-Wiener index for cycle-containing structures. *J. Chem. Inf. Comput. Sci.* **35,** 50–52 (1995).

129. Lukovits, I. Formulas for the hyper-Wiener index of trees. *J. Chem. Inf. Comput. Sci.* **34,** 1079–1081 (1994).

130. Lukovits, I. & Linert, W. A novel definition of the hyper-Wiener index for cycles. *J. Chem. Inf. Comput. Sci.* **34,** 899–902 (1994).

131. Ivanciuc, O. & Balaban, A. T. Design of topological indices. Part 8. Path matrices and derived molecular graph invariants. *MATCH Commun. Math. Comput. Chem* **30,** 141–152 (1994).

132. Randić, M. Molecular topographic descriptors. *Stud. Phys. Theor. Chem* **54,** 101–108 (1988).

133. Bogdanov, B., Nikolić, S. & Trinajstić, N. On the three-dimensional Wiener number. *J. Math. Chem.* **3,** 299–309 (1989).

134. Randić, M., Jerman-Blaẑić, B. & Trinajstić, N. Development of 3-dimensional molecular descriptors. *Comput. Chem.* **14,** 237–246 (1990).

135. Kier, L. B. An index of flexibility from molecular shape descriptors. *Prog. Clin. Biol. Res.* **291,** 105–109 (1989).

136. Schultz, H. P., Schultz, E. B. & Schultz, T. P. Topological organic chemistry. 9. Graph theory and molecular topological indices of stereoisomeric organic compounds. *J. Chem. Inf. Comput. Sci.* **35,** 864–870 (1995).

137. Schultz, H. P., Schultz, E. B. & Schultz, T. P. Topological organic chemistry. 10. Graph theory and topological indices of conformational isomers. *J. Chem. Inf. Comput. Sci.* **36,** 996–1000 (1996).

138. Hansen, P. J. & Jurs, P. C. Chemical applications of graph theory. Part I. Fundamentals and topological indices. *J. Chem. Educ.* **65,** 574 (1988).

139. CORAL, http://www.insilico.eu/coral

140. Milan Randić & Pompe, M. The Variable Connectivity Index 1xf versus the Traditional Molecular Descriptors: A Comparative Study of 1xf Against Descriptors of CODESSA. *J. Chem. Inf. Comput. Sci.* **41 (3),** 631–638 (2001).

141 Randić, M. & Basak, S. C. Optimal Molecular Descriptors Based on Weighted Path Numbers. *J. Chem. Inf. Comput. Sci.* **39,** 261–266 (1999).

142 Toropov, A. A. & Toropova, A. P. QSAR modeling of toxicity on optimization of correlation weights of Morgan extended connectivity. *J. Mol. Struct. THEOCHEM* **578,** 129–134 (2002).

143. Toropov, A. A., Toropova, A. P., Raska Jr, I., Leszczynska, D. & Leszczynski, J. Comprehension of drug toxicity: Software and databases. *Comput. Biol. Med.* **45,** 20–25 (2014).

144. Toropova, A. P. *et al.* CORAL: quantitative structure–activity relationship models for estimating toxicity of organic compounds in rats. *J. Comput. Chem.*

32, 2727–2733 (2011).

145. Toropova, A. P. *et al.* CORAL: QSPR models for solubility of [C60] and [C70] fullerene derivatives. *Mol. Divers.* **15,** 249–256 (2011).

Capítulo 3. Métodos estadísticos en QSPR-QSAR

3.1. Introducción

Los estudios QSPR-QSAR tienen como objetivo principal desarrollar modelos de correlación utilizando la variable respuesta de un compuesto químico (propiedad/actividad) y los datos de la información química del mismo. Las estrategias basadas en regresión y clasificación se emplean para desarrollar modelos, para datos con respuestas cuantitativas y cualitativas, respectivamente.

Además de los métodos convencionales, las herramientas de aprendizaje automático son útiles en el modelado QSPR-QSAR, especialmente para estudios que involucran datos con información química compleja, de alta dimensión, y que tienen una relación no lineal con la respuesta considerada.

Los enfoques basados en regresión se emplean cuando los datos de respuesta de los compuestos químicos son completamente numéricos, es decir, cuantitativos, mientras que las respuestas químicas cualitativas o semi-cuantitativas se modelan utilizando técnicas de clasificación. Los métodos basados en regresión permiten la predicción cuantitativa de la respuesta (propiedad/actividad), mientras que los métodos de clasificación permiten la categorización de los puntos de datos en varios grupos o clases, es decir, como altamente activos y menos activos.

Los métodos basados en aprendizaje automático también son útiles en el desarrollo de modelos QSPR-QSAR. Es importante mostrar que las herramientas de aprendizaje automático que emplean inteligencia artificial se pueden utilizar para resolver problemas basados en regresión y clasificación.

Numerosas herramientas estadísticas son útiles para la selección de características a partir de una gran matriz de datos. Las herramientas de selección de variables permiten el uso de descriptores adecuados y relevantes para una respuesta particular, eliminando los datos espurios en el análisis.

La matriz de datos de descriptores puede explorarse a través de diversos métodos de corte para reducir la información química inter-correlacionada y redundante. Los modelos QSPR-QSAR desarrollados también se someten a varias pruebas de validación para verificar su confiabilidad estadística. Después de su construcción, un modelo QSPR-QSAR se verifica mediante el empleo de estrategias de validación múltiples, dando una estimación de su predictibilidad y estabilidad.

De acuerdo con los lineamientos de la OCDE, el desarrollo de un modelo QSPR-QSAR debe cumplir con la estrategia de un algoritmo inequívoco y debe superar los criterios de aptitud, robustez y predictividad.

El presente capítulo proporciona una descripción de algunas herramientas estadísticas útiles y muy utilizadas para el pretratamiento de los datos, la selección de variables, el desarrollo de modelos y la validación de modelos QSPR-QSAR.

3.2. Varias herramientas de quimiometría utilizadas en QSPR-QSAR

Los estudios QSPR-QSAR son un enfoque que correlacionan los datos experimentales de la propiedad/actividad, con descriptores que codifican la información química. Dicha correlación puede derivarse de un enfoque basado en regresión (en los casos en que la propiedad o respuesta es cuantitativa y está disponible en una escala continua) o en un enfoque basado en clasificación (en los casos en que la propiedad o respuesta es gradual o semi-cuantitativa).

Los enfoques basados en regresión más comúnmente utilizados son los siguientes:

- Regresión Lineal Multivariable (MLR)

• Mínimos Cuadrados Parciales (PLS)

Algunos de los enfoques más comunes basados en clasificación son los siguientes:

• Análisis Discriminante Lineal (LDA)

• Análisis de Agrupamiento

Las herramientas de aprendizaje automático como la Red Neuronal Artificial y la Máquina de Soporte Vectorial, son muy efectivas en el desarrollo de modelos predictivos, particularmente en el manejo de los datos con información química compleja y de alta dimensión que muestran una relación no lineal con la(s) respuesta(s) de los compuestos químicos.

La quimiometría es la disciplina química que utiliza métodos estadísticos para diseñar procedimientos, experimentos y objetos óptimos, y para proporcionar la máxima información química mediante el análisis de los datos químicos.

Algunas de las herramientas quimiométricas más populares y comúnmente usadas en QSPR-QSAR se discutirán brevemente en este capítulo. Sin embargo, antes de aplicar cualquier método de desarrollo de modelos estadísticos, puede ser necesario pretratar la matriz de los datos, y seguidamente se realizará la aplicación de un método de selección de variables adecuado.

3.3. Pretratamiento de la matriz de datos

Al preparar una matriz de datos en los estudios QSPR-QSAR, se debe tener cuidado asegurando que las estructuras moleculares se hayan extraído o importado correctamente, que los datos de la actividad biológica (u otra respuesta) se hayan tomado de una fuente auténtica (y que tengan errores experimentales permisibles) y que los valores de los descriptores se hayan calculado utilizando un programa validado.

Los datos de la variable respuesta para un conjunto molecular específico empleados para construir un modelo QSPR-QSAR deberían presentar un

patrón de distribución normal y los experimentos realizados para determinar su valor haber sido obtenidos mediante el mismo protocolo.

Se debe tener cuidado en la inclusión de compuestos duplicados en el conjunto de datos. Del mismo modo, debe considerarse la forma tautomérica correcta de la estructura de los compuestos.

Para el cálculo de los descriptores tridimensionales, se debería haber llevado a cabo una adecuada optimización de la estructura molecular mediante métodos de estructura electrónica apropiados.

Cuando se han calculado una gran cantidad de descriptores moleculares, se debe aplicar un método apropiado para eliminar los descriptores menos importantes o redundantes. Se pueden omitir los descriptores con un valor constante para todas las observaciones y los descriptores que muestran una varianza muy baja. Solamente se debe conservar un descriptor entre aquellos pares de descriptores que muestran una alta intercorrelación mutua. En algunos casos, también se puede requerir un escalado adecuado de los descriptores[1].

3.4. Selección de variables

La selección de descriptores para el desarrollo del modelo es un paso vital en los estudios QSPR-QSAR. Dicha selección se puede realizar de diferentes formas, incluida la selección gradual (utilizando un criterio de escalonamiento adecuado, por ejemplo, '*F* para inclusión' y '*F* para exclusión', basado en el estadístico parcial-*F*), la mejor selección de subconjuntos, los métodos genéticos y el análisis factorial[1].

3.5. Análisis Discriminante Lineal

LDA[2] puede separar dos o más clases de objetos y puede usarse para problemas de clasificación. LDA realiza la misma tarea que MLR al predecir un resultado cuando la propiedad de respuesta tiene valores cualitativos y los

descriptores moleculares son variables continuas. LDA explícitamente intenta modelar la diferencia entre las clases de datos.

En una situación de dos grupos, los miembros previstos se obtienen calculando una puntuación de la función discriminante (FD) para cada caso. Entonces, los casos con valores de FD menores que el valor de corte se clasifican como pertenecientes a un grupo, mientras que aquellos con valores mayores se clasifican en el otro grupo. La FD puede tomar la siguiente forma:

$$FD = c_1.X_1 + c_2.X_2 + \ldots + c_d.X_d + c_0 \qquad (3.1)$$

donde FD es la función discriminante, que es una combinación lineal (suma) de las variables discriminantes, $X_1 \ldots X_d$ son las variables predictoras, c es el coeficiente discriminante o el peso para cada variable, c_0 es el término constante del modelo, y d es el número de variables predictoras.

Los buenos predictores tienden a tener grandes valores de coeficientes estandarizados. Después de analizar un conjunto existente de datos para calcular la FD y clasificar los casos, todos los casos nuevos (en el conjunto de predicción) se pueden clasificar.

En un análisis paso a paso de FD, el modelo se construye paso a paso. Específicamente, en cada paso se revisan y evalúan todas las variables para determinar cuál contribuirá más a la discriminación entre grupos. Esa variable se incluirá en el modelo y el proceso se inicia nuevamente.

3.6. Análisis de Agrupamiento

A diferencia de LDA, el Análisis de Agrupamiento[3] no requiere del conocimiento previo sobre cuáles elementos pertenecen a cuáles grupos. Los agrupamientos se definen mediante el análisis de los datos. El análisis de agrupamiento maximiza la similitud de los casos dentro de cada grupo y maximiza la diferencia entre los grupos que inicialmente son desconocidos.

El análisis de agrupamiento jerárquico encuentra grupos de casos relativamente homogéneos basados en disimilitudes o distancias entre objetos. La forma más sencilla y generalmente aceptada de calcular distancias

entre objetos en un espacio multidimensional es mediante el cálculo de las distancias euclidianas.

Comienza con cada caso como un agrupamiento separado y luego combina los agrupamientos secuencialmente, reduciendo la cantidad de agrupamientos en cada paso hasta que solo queda un agrupamiento. Se puede generar un diagrama de árbol jerárquico o dendrograma (Figura 3.1) para mostrar los puntos de vinculación: los agrupamientos están vinculados a niveles crecientes de disimilaridad.

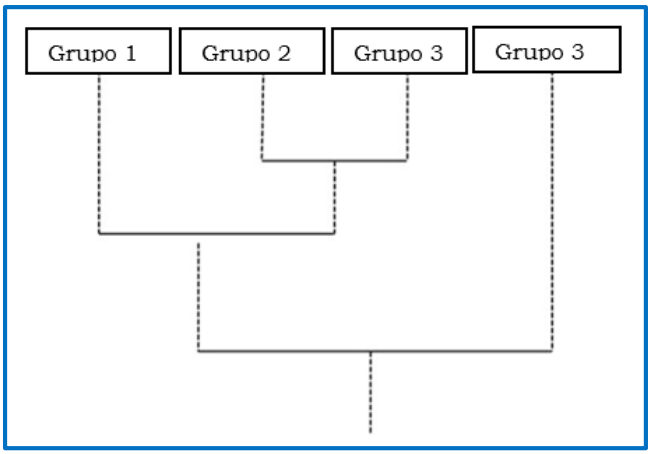


Figura 3.1. Ejemplo de un dendrograma.

El agrupamiento de *k-medias* es un método no-jerárquico de agrupamiento que se puede usar cuando se conoce el número de grupos presentes en los objetos o casos. Es un método no-supervisado de agrupamiento basado en el centroide.

En general, el método *k-medias* producirá exactamente *k* grupos diferentes. El método define *k* centroides, uno para cada grupo, ubicados lo más lejos posible el uno del otro. El siguiente paso es tomar cada punto perteneciente a un conjunto de datos dado y asociarlo al centroide más cercano. Cuando no hay ningún punto pendiente, las posiciones de los *k* centroides se vuelven a calcular. Este procedimiento se repite hasta que los centroides ya no se mueven[1].

3.7. Análisis de Regresión Lineal Multivariable

La Regresión Lineal Multivariable (MLR)[4] es un método comúnmente utilizado en QSPR-QSAR debido a su simplicidad, transparencia, reproducción y fácil interpretación. La expresión generalizada de una ecuación de MLR es la siguiente:

$$Y = a_0 + a_1.X_1 + a_2.X_2 + a_3.X_3 + \ldots + a_d.X_d \qquad (3.2)$$

En la expresión anterior, Y es la respuesta o variable dependiente, X_1, X_2, ..., X_d son los descriptores (características o variables independientes) presentes en el modelo con los coeficientes de regresión correspondientes a_1, a_2, ..., a_d, respectivamente, y a_0 es el término constante del modelo.

La interpretación de la contribución de los descriptores individuales X_1, X_2, ..., X_d es directa dependiendo del valor del coeficiente correspondiente y su signo algebraico. Cada coeficiente de regresión debería ser significativo en $p<0.05$, que puede verificarse a partir de una prueba *'t'*. Los descriptores presentes en un modelo MLR no deberían estar muy inter-correlacionados.

Para que un modelo sea estadísticamente confiable, el número de observaciones y el número de descriptores deben tener una relación de al menos 5:1, respectivamente. Un modelo MLR que se ajuste bien a los datos dará lugar a una gráfica de dispersión (actividad observada frente a actividad calculada) que muestra una desviación mínima de los puntos de la línea de ajuste.

La calidad de un modelo de MLR se determina a partir de una serie de parámetros como se describe a continuación.

1. Coeficiente de determinación *(R^2)*

Se puede definir el coeficiente de determinación *(R^2)* de la siguiente manera:

$$R^2 = 1 - \frac{\sum (Y_{obs} - Y_{cal})^2}{\sum (Y_{obs} - \overline{Y_{obs}})^2} \qquad (3.3)$$

En la ecuación anterior, Y_{obs} representa el valor de respuesta observada, mientras que Y_{cal} es la respuesta calculada con el modelo y $\overline{Y_{obs}}$ es el promedio de los valores de la respuesta observada. Para un modelo ideal, si la suma de residuos al cuadrado es 0, el valor de R^2 es 1.

A medida que el valor de R^2 se desvía de 1, la calidad de ajuste del modelo se deteriora. La raíz cuadrada de R^2 es el coeficiente de correlación *(R)*[1].

2. R^2 ajustado (R_a^2)

Si se aumenta el número de descriptores en un modelo para un número fijo de observaciones, los valores de R^2 siempre aumentarán, pero esto conducirá a una disminución en los grados de libertad y, por ende, tendrá una baja confiabilidad estadística.

Por lo tanto, un valor alto de R^2 no es necesariamente una indicación de que el modelo estadístico se ajusta bien a los datos disponibles. Para reflejar la varianza explicada (es decir, la fracción de la varianza de los datos explicada por el modelo) de una mejor manera, el valor de R^2 ajustado se ha definido de la siguiente manera:

$$R_a^2 = \frac{(N-1).R^2 - d}{N-1-d} \tag{3.4}$$

En la expresión anterior, d es la cantidad de variables utilizadas en la construcción del modelo[1].

3. Relación de Varianza *(F)*

Para juzgar la importancia general de los coeficientes de regresión, la relación de varianza (es decir, la relación entre el significado de la regresión y el valor medio de las desviaciones) puede definirse de la siguiente manera:

$$F = \frac{\frac{\sum\left(Y_{cal} - \overline{Y}\right)^2}{d}}{\frac{\sum\left(Y_{obs} - Y_{cal}\right)^2}{N-d-1}} \tag{3.5}$$

El valor de *F* tiene dos grados de libertad: *d* y *N-d-1*. El valor de *F* calculado de un modelo debe ser significativo en *p<0.05*. Para la significación global de los coeficientes de regresión, el valor de *F* debe ser alto.

4. Estimación de la desviación estándar *(S)*

Para tener un buen modelo, la estimación del error estándar de *Y* debe ser bajo y esto se puede definir de la siguiente manera:

$$S = \sqrt{\frac{\left(Y_{obs} - Y_{cal}\right)^2}{N - d - 1}} \tag{3.6}$$

Aquí, los grados de libertad son *N-d-1*.

3.8. Mínimos Cuadrados Parciales (PLS)

Al manejar una gran cantidad de descriptores inter-correlacionados y ruidosos para un número limitado de datos, PLS es una mejor opción en comparación a MLR. PLS, al ser una generalización de MLR[5], trata de extraer las variables latentes (LV) que son funciones de las variables originales, y representan la mayor variación posible del factor subyacente al modelar las respuestas.

Antes del análisis, las variables *X* e *Y* a menudo se transforman para hacer sus distribuciones bastante simétricas. Las variables de respuesta generalmente se transforman logarítmicamente y las variables *X* se deben escalar apropiadamente.

El PLS lineal encuentra algunas variables nuevas (LV), que son combinaciones lineales de las variables originales. Cuando el número de LV es igual al número de variables, el modelo PLS se vuelve igual que el modelo MLR. Es necesario realizar una prueba estricta de la importancia predictiva de cada componente PLS, y luego detener la adición de nuevos componentes cuando los componentes se vuelven no-significativos.

La validación cruzada (CV) es una forma práctica y confiable de probar esta capacidad predictiva. Una ecuación PLS puede expresarse de la misma

forma que en MLR; por lo tanto, las contribuciones de los descriptores individuales a la respuesta se pueden descubrir fácilmente[1].

3.9. Métodos de búsqueda basados en regresiones

En la literatura especializada existe una enorme cantidad de algoritmos matemáticos aproximados de MLRA para resolver el problema de encontrar *d* variables óptimas $\mathbf{d} = \{\mathbf{X_1}, \mathbf{X_2}, \mathbf{X_3}, \ldots, \mathbf{X_d}\}$ que dan origen al modelo a partir de un conjunto de descriptores *D* mucho mayor a *d*. Por conjunto óptimo debe entenderse el que produce los mejores parámetros estadísticos como el coeficiente de correlación (*R*) y la desviación estándar (*S*).

En esta sección, se mencionarán aquellos algoritmos que son las más famosos y reconocidos en el campo, y aquellos que se usaron a lo largo del transcurso del presente trabajo de tesis. Se empezará describiendo la búsqueda exacta y posteriormente los métodos aproximados.

3.9.1. Búsqueda Exhaustiva (FS)

Para encontrar el mejor modelo de *d* descriptores a partir de un número mucho mayor *D* es necesario realizar $D![d!(D - d)!]$ regresiones lineales; es decir, probar todos los casos posibles. Si se elige el mínimo valor de *S* como criterio de búsqueda, entonces aquel conjunto de *d* variables con menor *S* será la mejor solución posible dentro del conjunto inicial analizado.

En términos matemáticos, debemos encontrar el mínimo global $S(d)$ en un conjunto de $\mathrm{D!/(D-d)!\,D!}$ "puntos" $\mathbf{d}_i = \{X_{i1}, X_{i2}, \ldots, X_{id}\}$. Como cada descriptor $\mathbf{X_i}$ es un vector de *N* componentes, entonces cada punto $\mathbf{d}_i$ es una matriz de $d \times N$.

Debido a que habitualmente se trabaja con un conjunto inicial **D** grande, la búsqueda exacta es impracticable salvo que se cuente con una supercomputadora, y aun así continuará resultando un procedimiento sumamente costoso.

Los métodos estándares de optimización numérica conocidos hasta el momento no resultan apropiados para este tipo de problemas, por lo cual se han propuesto una multitud de métodos aproximados.

3.9.2. Método de regresión "de a pasos"

El método de Regresión de a pasos se usa desde hace ya mucho tiempo y es clásico en el trabajo QSPR-QSAR. Su popularidad deriva de tratarse de un procedimiento rápido, sencillo, nada costoso, y que se halla disponible en cualquier paquete computacional comercial. Esta aproximación presenta básicamente tres variantes, aunque el fundamento del método es el mismo: métodos inclusión de a pasos (SI), exclusión de a pasos (SE), y de a pasos (SW).

La inclusión de a pasos consiste en calcular en una primera etapa el mejor modelo de una variable, luego en cada etapa subsiguiente adicionar una nueva variable óptima que mejore la calidad del modelo, y el modelo óptimo se encuentra cuando no es posible mejorar más la relación.

La exclusión de a pasos, por su parte, emplea un procedimiento opuesto: comienza con un modelo que contiene varias variables y en cada etapa remueve aquella que no contribuya a mejorar la correlación. El proceso finalmente termina cuando se alcanza un conjunto óptimo, donde la remoción de cualquiera de las variables presentes empeora la calidad del modelo.

El método de a pasos es un método que se usa bastante en la actualidad, y se trata de una combinación de SI y SE. En cada paso existen cuatro alternativas posibles: agregar una variable, eliminar una variable, intercambiar dos variables, o finalizar el proceso de búsqueda. Usando estas cuatro opciones se pueden establecer múltiples variantes de SW, dependiendo de cómo se lo ejecute[6].

El método de regresión de a pasos no garantiza alcanzar una solución óptima que coincida con la solución exacta, simplemente porque cada vez que se introduce/remueve una variable en el modelo, ya hay otras presentes en el mismo que restringen y determinan la calidad de la ecuación.

Aquí aparece lo que se conoce con el nombre de "efecto de mezcla de variables", resultante de que las variables se combinan entre sí para producir un determinado efecto final sobre los parámetros estadísticos. El uso de una regresión de a pasos no tiene en cuenta dicho efecto, y esto explica que el método constituya una aproximación. Una manera de disminuir el efecto de mezcla es usar variables que se encuentren menos inter-correlacionadas entre sí y que resulten en cierto grado ortogonales.

3.9.3. Método del Reemplazo (RM)

El Método del Reemplazo[7] surgió de la carencia de contar con un programa computacional comercial que permita analizar cientos de descriptores moleculares para encontrar los mejores subconjuntos de variables.

La principal ventaja de RM es que requiere realizar un número de regresiones lineales que es mucho menor al caso FS y produce resultados finales que coinciden o son cercanos a los exactos. Se trata de una aproximación con una "receta" fácil de entender y aplicar que tiene en cuenta el efecto de mezcla de variables comentado antes.

La idea principal de RM es que se puede obtener un mínimo en *S* tomando en cuenta el error relativo (*der*) de los coeficientes de regresión del ajuste lineal con *d* descriptores, $\mathbf{d} = \{\mathbf{X_1}, \mathbf{X_2}, ..., \mathbf{X_d}\}$. El fundamento de emplear este camino para efectuar la búsqueda de variables radica en la simple observación de que los modelos obtenidos haciendo FS producen *der* bajas en los coeficientes, y ello nos hizo pensar que sería un criterio apropiado para ir probando variables en el modelo.

En la actualidad no sabemos de otro método reportado en la literatura que también este basado en los *der*. La razón tal vez sea que experimentos numéricos conducidos tiempo atrás por otros investigadores sugirieran que los coeficientes de regresión y sus errores asociados tenían un comportamiento aleatorio, dado por la inestabilidad de las ecuaciones de regresión.

La esencia del procedimiento RM es la siguiente: primero se elige un conjunto de *d* descriptores de manera aleatoria y se hace una regresión lineal. Luego se escoge uno de los descriptores del conjunto, digamos X_i, se lo reemplaza con cada uno de los *D* descriptores del conjunto total (excepto por sí mismo), y se conserva el conjunto resultante con menor S.

Debido a que en este primer paso se puede escoger cualesquiera de los *d* descriptores en el modelo inicial, se tendrán *d* caminos distintos que conducen a soluciones finales. Luego se elige la variable en el modelo resultante que posea el mayor valor *der* en su coeficiente (omitiendo la que fue reemplazada en el paso previo) y se reemplaza por todos los descriptores (excepto ella misma), reteniendo nuevamente el mejor conjunto resultante. De ese modo se reemplazan las variables remanentes omitiendo las reemplazadas en pasos previos.

Una vez finalizado este proceso se comienza nuevamente con la variable que tiene el mayor valor *der* en su coeficiente y se repite el ciclo completo. Este proceso se repite tantas veces como sea necesario hasta que el conjunto de descriptores resulte invariante. Al final, tendremos el mejor modelo para el camino *i*. Se procede exactamente de la misma manera para todos los caminos posibles $i = 1, 2, \ldots, d$, y se elige el conjunto de variables con menor *S*.

Los experimentos numéricos muestran que en esta forma se consigue un conjunto óptimo de *d* descriptores con las características señaladas previamente.

3.10. Importancia de los parámetros de calidad en QSPR-QSAR

El avance en recursos computacionales rápidos y económicos hace que sea factible calcular una gran cantidad de descriptores usando varias herramientas de programación. Como consecuencia, no se puede negar el riesgo de correlaciones fortuitas con el número creciente de variables incluidas en el modelo QSPR-QSAR en comparación con el número limitado de compuestos generalmente empleados para el desarrollo del modelo[8].

Por otro lado, empleando diversas herramientas de optimización, es factible obtener modelos que puedan ajustar bien los datos experimentales, pero siempre existe la posibilidad de sobre-ajustarlos.

Los parámetros de ajuste de los datos no corroboran una buena capacidad de predicción del modelo, ya que aquellos son parámetros para la calidad estadística del modelo. Esta es la razón principal por la que las herramientas de validación se deben aplicar en el modelo QSPR-QSAR desarrollado, para verificar su predictividad para las nuevas moléculas no ensayadas.

En la Figura 3.2 se muestra un diagrama de flujo para el desarrollo de un modelo QSPR-QSAR confiable, junto con los diversos métodos de validación con los parámetros comúnmente utilizados.

3.11. Los Principios de la OCDE

Los principios de la OCDE representan el mejor esquema posible de todos los puntos esenciales que deben abordarse al desarrollar modelos QSPR-QSAR fiables y reproducibles[9]. Los principios fueron formulados por expertos de QSAR en una reunión celebrada en Setúbal, Portugal, en 2002 así como las directrices para la validación de los modelos QSPR-QSAR, en particular para fines regulatorios.

Estos principios fueron luego aprobados por los países miembros de la OCDE, QSAR y las comunidades reguladoras en la 37ª Reunión Conjunta del Comité de Químicos y el Grupo de Químicos, Plaguicidas y Biotecnología en noviembre de 2004.

Las cinco directrices adoptadas por la OCDE que denotan la validez del modelo QSAR son las siguientes:

• Principio 1- Una respuesta definida (actividad/propiedad/toxicidad)

• Principio 2- Un algoritmo inequívoco

• Principio 3- Un dominio de aplicabilidad definido

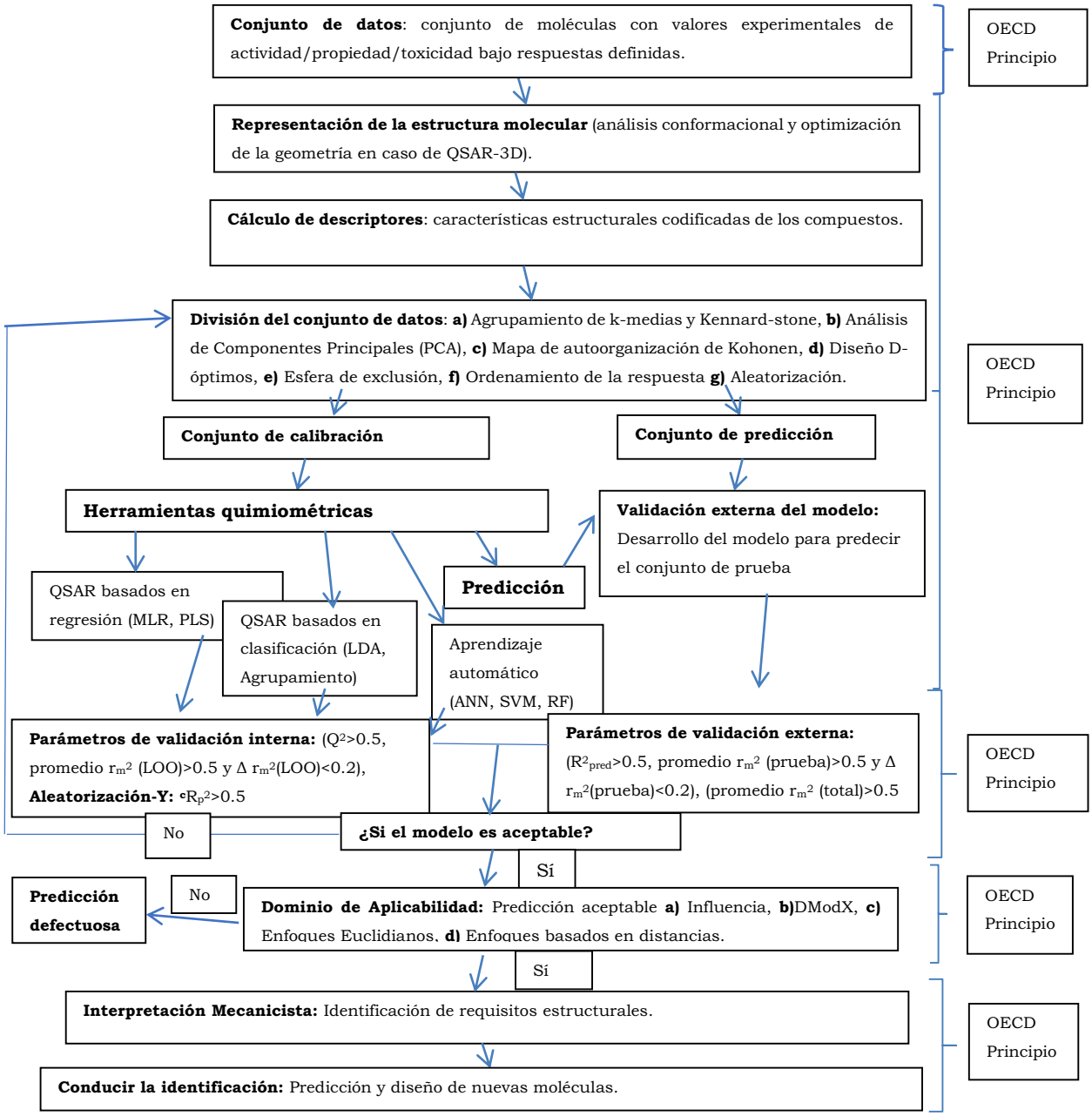


Figura 3.2. Pasos fundamentales para la generación de un modelo QSAR y métodos de validación empleados.

• Principio 4- Medidas apropiadas de bondad de ajuste, robustez y predictividad

• Principio 5- Una interpretación mecanicista, si es posible.

El desafío actual en el proceso de desarrollo de un modelo QSPR-QSAR ya no es desarrollar un modelo que sea estadísticamente sólido para predecir la actividad dentro del conjunto de calibración, sino desarrollar un modelo con

la capacidad de predecir con precisión la actividad de nuevos productos químicos[1].

3.12. Validación interna

La validación interna de un modelo QSPR-QSAR se realiza en base a las moléculas utilizadas en el desarrollo del modelo. Implica la predicción de la actividad de las moléculas estudiadas, seguida de la estimación de parámetros para detectar la precisión de las predicciones. Para juzgar la calidad y la bondad de ajuste del diseño del modelo, la validación interna es una técnica ideal. Sin embargo, la principal desventaja de este enfoque es la falta de predicción del modelo cuando se aplica a un nuevo conjunto de datos[10].

3.13. Validación externa

No se puede juzgar la predictibilidad del modelo desarrollado a partir de la validación interna de un conjunto completamente nuevo de compuestos, ya que la validación interna considera los compuestos químicos que pertenecen al mismo conjunto de compuestos utilizados para el desarrollo del modelo.

Así, para la validación externa, el conjunto de datos disponible generalmente se divide en conjuntos de calibración y de predicción, luego se desarrolla un modelo con el conjunto de calibración, y luego se utiliza el modelo construido para verificar la validación externa empleando las moléculas del conjunto de predicción que no se utilizan en el proceso de desarrollo del modelo. La validación externa asegura la capacidad predictiva y la aplicabilidad del modelo QSPR-QSAR desarrollado para la predicción de moléculas que no han sido probadas antes[11].

3.13.1. Selección de los conjuntos de calibración y predicción

En general, la división del conjunto molecular en conjuntos de calibración y predicción debe ejecutarse de tal manera que los puntos que representan ambos conjuntos, de calibración y predicción estén dispersos

dentro del espacio descriptor ocupado por el conjunto de datos completo y cada punto del conjunto de predicción esté cerca de al menos un compuesto del conjunto de calibración.

Los siguientes enfoques son empleados principalmente por quienes desarrollan los modelos QSPR-QSAR para la selección de los conjuntos de calibración y predicción[11]:

1. Selección aleatoria: El conjunto de datos puede dividirse por un mero proceso de selección aleatorio.

2. Basado en la respuesta *Y*: este enfoque se basa en el muestreo de la actividad (respuesta *Y*). El rango completo de la respuesta se divide en grupos y los compuestos que pertenecen a cada grupo se asignan a los conjuntos de calibración o predicción al azar o de forma personalizada.

3. Basado en la respuesta *X*: las propiedades y la similitud estructural de las moléculas se consideran para la agrupación de compuestos similares. Después de eso, una fracción de compuestos estudiados se asigna al conjunto de calibración o al conjunto de predicción manualmente o de una manera regular. Las herramientas más comúnmente empleadas para la división racional de los conjuntos de datos son:

• Agrupación *k-medias*,

• Selección de mapa de autoorganización de Kohonen,

• Diseño molecular estadístico,

• Selección de Kennard-Stone,

• Esfera de Exclusión, y

• Selección de extrapolación orientada al conjunto de predicción.

3.13.2. Dominio de aplicabilidad (DA)

1. Concepto de DA

El DA se define como una región teórica en el espacio químico construido tanto por los descriptores del modelo como por la respuesta

modelada. El dominio de aplicabilidad juega un papel crucial para estimar la incertidumbre en la predicción de un compuesto particular en función de cuán similar es a los compuestos empleados para construir el modelo QSPR-QSAR.

Por lo tanto, la predicción de una respuesta estudiada a través de un modelo es aplicable solo si el compuesto que se predice cae dentro del DA del modelo, ya que no es factible predecir todo el universo de compuestos usando un único modelo QSPR-QSAR[12].

2. Tipos de aproximaciones DA

Las técnicas más comúnmente empleadas para estimar regiones de interpolación en un espacio multivariable son las siguientes:

(a) intervalos en el espacio descriptor

(b) métodos geométricos

(c) métodos basados en la distancia

(d) distribución de densidad de probabilidad

(e) rango de la variable de respuesta

Los primeros cuatro enfoques se basan en la metodología utilizada para la caracterización del espacio de interpolación en el espacio descriptor del modelo. Por el contrario, el último depende únicamente del espacio de respuesta de las moléculas del conjunto de calibración.

Se puede identificar un compuesto fuera del DA, si:

(a) al menos un descriptor está fuera del rango, para los enfoques de rangos y (b) la distancia entre la sustancia química y el centro del conjunto de datos de calibración excede el umbral para los enfoques de distancia.

El umbral para todos los tipos de métodos de distancia, es la distancia más grande entre los puntos de datos del conjunto de calibración y el centro del conjunto de datos del conjunto de calibración[1].

3.14. Algunos parámetros de validación interna y externa

1. Validación cruzada dejar-uno-afuera (loo)[13]

Para determinar la validación cruzada loo, el conjunto de calibración se modifica principalmente mediante la eliminación de un compuesto del conjunto. Luego, el modelo se reconstruye en base a las moléculas restantes del conjunto de calibración utilizando la combinación de descriptores originalmente seleccionada, y la actividad del compuesto eliminado se calcula en base a la ecuación QSAR resultante.

Este ciclo se repite hasta que todas las moléculas del conjunto de calibración se han eliminado una vez, y los datos de actividad predichos obtenidos para todos los compuestos del conjunto de calibración se usan para el cálculo de varios parámetros de validación interna.

Finalmente, la predicción del modelo se juzga usando la sumatoria de los residuos al cuadrado predichos *(PRESS)* y R^2 (Q^2) con la validación cruzada para el modelo, mientras que el valor de la desviación estándar del error de predicción *(SDEP)* se calcula a partir de *PRESS*.

$$PRESS = \sum \left(Y_{obs} - Y_{pred} \right)^2 \tag{3.7}$$

$$SDEP = \sqrt{\frac{PRESS}{n}} \tag{3.8}$$

$$Q^2 = 1 - \frac{\sum \left(Y_{obs(cal)} - Y_{pred(cal)} \right)^2}{\sum \left(Y_{obs(cal)} - \overline{Y}_{calibración} \right)^2} = \frac{PRESS}{\sum \left(Y_{obs(cal)} - \overline{Y}_{calibración} \right)^2} \tag{3.9}$$

En las ecuaciones (3.7) -(3.9), Y_{obs} e Y_{pred} corresponden a los valores de actividad observados y calculados de loo, n se refiere al número de observaciones, $Y_{obs(cal)}$ es la actividad observada del conjunto de calibración, $Y_{pred(cal)}$ es la actividad calculada de las moléculas del conjunto de calibración basado en la técnica de loo. El valor de umbral de Q^2 es 0.5[1].

2. Validación cruzada dejar-n%-afuera (lmo)

El principio básico de la técnica lmo o dejar-n%-afuera es que una porción definida del conjunto de calibración se mantiene y se elimina en cada ciclo (siendo n% el porcentaje de moléculas que se remueven del conjunto por etapas y corresponde a *nm* moléculas).

Para cada ciclo, el modelo se construye en base a las moléculas restantes (y utilizando los descriptores originalmente seleccionados) y luego se predice la actividad de los compuestos eliminados utilizando el modelo desarrollado. Después de que se hayan completado todos los ciclos, los valores de actividad predichos de los compuestos se usan para el cálculo del lmo-Q^2.

3. Verdadero Q^2

Hawkins *et al.*[14] propusieron el concepto del parámetro "verdadero Q^2", calculado a partir de la aplicación de la estrategia de selección de variables en cada ciclo de validación. El parámetro es una mejor herramienta para evaluar la predicción del modelo, principalmente en el caso de conjuntos de datos pequeños, en comparación con el enfoque tradicional de la división del conjunto de datos en conjuntos de calibración y predicción.

4. El parámetro de R_m^2 para la validación interna

Un valor aceptable de Q^2 no indica necesariamente que los datos de actividad predichos se encuentran en proximidad cercana a los observados, aunque pueda existir una buena correlación general entre dichos valores. Por lo tanto, para obviar este problema y para indicar mejor la predictibilidad del modelo, el parámetro R_m^2 introducido por Roy *et al.*[15] se puede calcular mediante las siguientes ecuaciones:

$$\overline{R_m^2} = \frac{\left(R_m^2 + R_m'^2\right)}{2} \tag{3.10}$$

$$\Delta R_m^2 = \left|R_m^2 - R_m'^2\right| \tag{3.11}$$

Aquí, $R_m^2 = R^2\left(1 - \sqrt{\left(R^2 - R_0^2\right)}\right)$ y $R_m'^2 = R^2\left(1 - \sqrt{\left(R^2 - R_0'^2\right)}\right)$. Los parámetros R^2 y R_0^2 son los coeficientes de correlación al cuadrado entre los valores predichos y los valores calculados (dejar-uno-afuera) de los compuestos con y sin

intercepto, respectivamente. El parámetro $R_0^{'2}$ tiene el mismo significado, pero usa los ejes invertidos.

El $\overline{R_m^2}$ es el valor promedio de R_m^2 y $R_m^{'2}$, y ΔR_m^2 es la diferencia absoluta entre R_m^2 y $R_m^{'2}$. En el caso de validación interna del conjunto de calibración, los parámetros $\overline{R_{mloo}^2}$ y ΔR_{mloo}^2 pueden utilizarse y demostrar que el valor de ΔR_{mloo}^2 debe ser preferiblemente inferior a 0.2, lo que indica que el valor de $\overline{R_{mloo}^2}$ es superior a 0.5. Roy *et al.*[16] propusieron que el cálculo del parámetro R_m^2 debería basarse en los valores escalados de los datos de respuesta observados y predichos. El escalado se puede hacer en base a la siguiente ecuación:

$$escalado Y_i = \frac{Y_i - Y_{\min(obs)}}{Y_{\max(obs)} - Y_{\min(obs)}} \tag{3.12}$$

Aquí, Y_i se refiere a la respuesta observada/predicha para el compuesto *i-ésimo* (1, 2, 3, ..., n) en el conjunto de calibración/prueba. Además de estos, $Y_{\max(obs)}$ e $Y_{\min(obs)}$ indican los valores máximos y mínimos, respectivamente, para la respuesta observada en los compuestos del conjunto de calibración.

5. Parámetros para la correlación casual: aleatorización-Y

La aleatorización-Y se realiza para garantizar la solidez del modelo QSPR-QSAR desarrollado. En la prueba de aleatorización-Y, la validación se realiza permutando los valores de respuesta (Y) con respecto a la matriz X que se ha mantenido inalterada[17,18].

6. Validación externa R_{pred}^2

El R_{pred}^2 refleja el grado de correlación entre los datos de propiedad observados y predichos del conjunto de predicción.

$$R_{pred}^2 = 1 - \frac{\sum \left(Y_{obs(pred)} - Y_{pred(pred)} \right)^2}{\sum \left(Y_{obs(pred)} - \overline{Y}_{calibración} \right)^2} \tag{3.13}$$

Aquí, $Y_{obs(pred)}$ y $Y_{pred(pred)}$ son los datos de propiedad observados y predichos para los compuestos del conjunto de predicción, mientras que $\overline{Y}_{calibración}$ indica el

valor medio de propiedad observada de las moléculas del conjunto de calibración. Por lo tanto, los modelos con valores de R^2_{pred} por encima del valor estipulado de 0.5 se consideran que presentan una buena capacidad predictiva.

7. Criterios de Golbraikh y Tropsha

Golbraikh y Tropsha[19] propusieron un conjunto de parámetros para determinar la predictibilidad externa del modelo QSPR-QSAR. Según estos autores, los modelos se consideran satisfactorios si todas las condiciones siguientes se satisfacen:

a) $Q^2_{calibración} > 0.5$

b) $R^2_{pred} > 0.6$

c) $\frac{R^2 - R_0^2}{R^2} < 0.1$ y $0.85 \le k \le 1.15$ o $\frac{R^2 - R_0'^2}{R^2} < 0.1$ y $0.85 \le k' \le 1.15$

d) $\left|R^2 - R_0'^2\right| < 0.3$

El significado de los términos R^2 y R_0^2 ya se discutió en la sección "parámetro R^2_m de validación interna".

8. El parámetro de validación externa $R^2_{m(pred)}$

Con el fin de verificar la proximidad entre los datos observados y predichos, el parámetro $R^2_{m(pred)}$ ha sido desarrollado por Roy *et al.*[15]. El valor de $R^2_{m(pred)}$ se calcula utilizando los coeficientes de correlación al cuadrado entre la actividad observada y la predicha de los compuestos del conjunto de predicción. Para que la predicción sea aceptable, el valor de $\Delta R^2_{m(pred)}$ debe ser preferiblemente inferior a 0.2, siempre que el valor de $R^2_{m(pred)}$ sea superior a 0.5.

3.15. Conclusiones

Los estudios QSPR-QSAR implican el uso de un número significativo de herramientas estadísticas y, por lo tanto, requiere de un buen conocimiento de la quimiometría. El modelo matemático establecido puede proporcionar

una relación tanto lineal como no lineal entre la respuesta y los atributos químicos a través de análisis basados en regresión y en clasificación.

Debido a que se establecen relaciones matemáticas cuantitativas, la validación utilizando un algoritmo estadístico adecuado se vuelve esencial para confirmar la estabilidad y la predictibilidad de los modelos. El juicio para la elección del método de validación depende del número de datos analizados y de descriptores utilizados, e inclusive del objetivo del análisis.

Bibliografía

1. Roy, K., Kar, S. & Das, R. N. Statistical methods in QSAR/QSPR. in *A primer on QSAR/QSPR modeling* 37–59 (Springer, 2015).
2. Agresti, A. *An introduction to categorical data analysis*. **135,** (Wiley New York, 1996).
3. Everitt, B., Landau, S. & Leese, M. Cluster analysis 4th ed. *London: Arnold* (2001).
4. Snedecor, G. W. & Cochran, W. G. Statistical methods 6th edition Oxford and IBH Publishing Co. *New delhi* (1967).
5. Wold, S., Sjöström, M. & Eriksson, L. PLS-regression: a basic tool of chemometrics. *Chemom. Intell. Lab. Syst.* **58,** 109–130 (2001).
6. Dixon, W. J. *Biomedical computer programs P-series*. (University of California Press, 1979).
7. Duchowicz, P. R., Castro, E. A. & Fernández, F. M. Alternative algorithm for the search of an optimal set of descriptors in QSAR-QSPR studies. *MATCH Commun. Math. Comput. Chem* **55,** 179–192 (2006).
8. Topliss, J. G. & Costello, R. J. Chance correlations in structure-activity studies using multiple regression analysis. *J. Med. Chem.* **15,** 1066–1068 (1972).
9. Jaworska, J. S., Comber, M., Auer, C. & Van Leeuwen, C. J. Summary of a workshop on regulatory acceptance of (Q) SARs for human health and environmental endpoints. *Environ. Health Perspect.* **111,** 1358 (2003).
10. Wold, S. Cross-validatory estimation of the number of components in factor and principal components models. *Technometrics* **20,** 397–405 (1978).
11. Roy, K. On some aspects of validation of predictive quantitative structure–activity relationship models. *Expert Opin. Drug Discov.* **2,** 1567–1577 (2007).
12. Gramatica, P. Principles of QSAR models validation: internal and external. *Mol. Inform.* **26,** 694–701 (2007).
13. Roy, K. & Mitra, I. On various metrics used for validation of predictive QSAR models with applications in virtual screening and focused library design. *Comb. Chem. High Throughput Screen.* **14,** 450–474 (2011).
14. Hawkins, D. M., Basak, S. C. & Mills, D. Assessing model fit by cross-validation. *J. Chem. Inf. Comput. Sci.* **43,** 579–586 (2003).

15. Roy, K. *et al.* Comparative studies on some metrics for external validation of QSPR models. *J. Chem. Inf. Model.* **52,** 396–408 (2012).

16. Roy, K. *et al.* Some case studies on application of "rm2" metrics for judging quality of quantitative structure–activity relationship predictions: emphasis on scaling of response data. *J. Comput. Chem.* **34,** 1071–1082 (2013).

17. Mitra, I., Roy, P. P., Kar, S., Ojha, P. K. & Roy, K. On further application of r m2 as a metric for validation of QSAR models. *J. Chemom.* **24,** 22–33 (2010).

18. Mitra, I., Saha, A. & Roy, K. Exploring quantitative structure–activity relationship studies of antioxidant phenolic compounds obtained from traditional Chinese medicinal plants. *Mol. Simul.* **36,** 1067–1079 (2010).

19. Golbraikh, A. & Tropsha, A. Beware of q2! *J. Mol. Graph. Model.* **20,** 269–276 (2002).

Capítulo 4. Propiedades fisicoquímicas estudiadas y sus aplicaciones en el desarrollo de modelos QSPR en pesticidas

4.1. Introducción

Las propiedades de interés agronómico y de importancia en el área de la ciencia de los pesticidas van a ser desarrolladas en este capítulo de acuerdo a como se han presentado a lo largo del presente trabajo de tesis. Las mismas se analizan a partir del conocimiento que se ha ido adquiriendo en el desarrollo de los modelos QSPR para compuestos plaguicidas.

Las principales diferencias entre los perfiles de comportamiento de los productos químicos orgánicos en el medio ambiente son atribuibles a sus propiedades fisicoquímicas. Algunas de las propiedades claves se reconocen como solubilidad acuosa, presión de vapor, los tres coeficientes de partición entre aire, agua y octanol, la constante de disociación en agua (cuando sea relevante) y la susceptibilidad a la degradación o el factor de bioconcentración.

La identidad química puede parecer un problema trivial, pero la mayoría de los compuestos tienen varios nombres, y son muy sutiles las diferencias entre isómeros (por ejemplo, *cis* y *trans*) y estás pueden ignorarse. Los identificadores más comúnmente aceptados son el nombre IUPAC y el número del Sistema de Resúmenes Químicos (CAS).

Más recientemente, se han buscado métodos para expresar la estructura en forma de notación lineal para que la entrada en la computadora de una serie de símbolos se pueda utilizar para definir una estructura tridimensional para fines ambientales. En este sentido el formato molecular SMILES (sistema de entrada molecular lineal simplificado), permite poder obtener la masa molar o el peso molecular a partir de la estructura.

4.2. Enfoque QSPR para el coeficiente de sorción en suelo

El coeficiente de sorción en suelo (*Koc*) se puede estimar para compuestos orgánicos. Puede definirse como "la relación de la cantidad de un compuesto químico absorbido al suelo por unidad de peso de carbono orgánico (oc) en el suelo o sedimento a la concentración del producto químico en solución a equilibrio"[1]. Este es representado por la siguiente ecuación:

$$Koc = \frac{(\mu g\ absorbidos / g\ carbono\ orgánico)}{(\mu g / mL\ solución)} \qquad (4.1)$$

Las unidades de *Koc* son típicamente expresadas en l/kg o ml/g.

El coeficiente de sorción en suelo provee una indicación de la medida a la cual se particiona un compuesto químico entre las fases sólidas y la solución del suelo, o entre el agua y los sedimentos en los ecosistemas acuáticos. Describe la biodegradación y el impacto de la contaminación de los pesticidas orgánicos[2] cuando estos compuestos interactúan con la materia orgánica de los suelos y los sedimentos, ya sea en la superficie, el suelo o el agua potable[3].

La estimación confiable del parámetro *Koc* es muy importante en agricultura, y a menudo se realizan para evaluaciones de destino ambiental, ya que su medición experimental es difícil, costosa y lleva mucho tiempo. Por lo tanto, la predicción del coeficiente de sorción en suelo para un gran número de estructuras químicas es muy conveniente en la evaluación de riesgos[4].

En el ámbito de la teoría de las relaciones cuantitativas de estructura y propiedad (QSPR)[5–7], se puede predecir una propiedad experimental de un compuesto químico, es decir, *Koc*, mediante el conocimiento de su estructura química.

La estructura se cuantifica por medio de un conjunto de descriptores moleculares adecuados, en otras palabras, cantidades numéricas que llevan información específica sobre los aspectos constitucionales, topológicos, geométricos, hidrofóbicos y/o electrónicos[8–10]. Por consiguiente, los descriptores se correlacionan estadísticamente con la propiedad experimental,

lo que da como resultado un modelo matemático que se puede utilizar para descubrir paralelismos útiles.

Los métodos de estimación tradicional son fiables y se sabe que muchos modelos QSPR publicados que predicen el coeficiente de sorción del suelo implican el coeficiente de partición octanol/agua experimental (*Kow*) o la solubilidad en agua (*Sw*)[11], pero el índice de conectividad molecular de primer orden *(MCI)* se ha empleado satisfactoriamente para predecir los valores de *Koc* para compuestos orgánicos hidrofóbicos[12,13], mientras que otros QSPR se basan en descriptores moleculares teóricos[14–16].

Sin embargo, por lo general, se han realizado pocos trabajos para examinar la predictividad (validación) del modelo y el dominio de aplicabilidad químico de dichos modelos sobre una amplia gama de compuestos, especialmente para los nuevos compuestos químicos[17–19].

El módulo original de KOCWIN de EPI Suite (PCKOC) empleaba a *MCI* y una serie de grupo de factores de contribución para predecir *Koc*[20]. Este método de contribución de grupos demostró mejorar los métodos de estimación tradicionales basados en el coeficiente de partición octanol/agua y de solubilidad en agua.

Un estudio QSPR anterior de Gramática *et al.*[17] en un conjunto altamente heterogéneo de 643 compuestos orgánicos no-iónicos predice el coeficiente de sorción en suelo expresado en unidades logarítmicas (*logKoc*).

El conjunto de calibración con 93 compuestos utilizados en dicho trabajo es peculiar, porque es mucho más pequeño que el conjunto de predicción de 550 compuestos (proporción 1:6).

Los mejores descriptores moleculares de Dragon se seleccionan a través de la técnica de Algoritmos Genéticos (GA) basada en MLRA, que conduce a un modelo QSPR de 4 descriptores con una capacidad predictiva del 78% en el conjunto de predicción. Los mejores resultados se obtienen mediante el modelado de consenso de 10 modelos diferentes en la población del modelo GA.

En este trabajo, informamos nuevos modelos QSPR alternativos para el coeficiente de sorción en suelo en el mismo conjunto molecular estudiado por Gramática *et al.*[17], utilizando un enfoque que no considera la representación conformacional de la estructura química al basarse únicamente en los aspectos constitucionales y topológicos de las moléculas[18].

Como es sabido, cada modelo que incluye descriptores tridimensionales generalmente implica altos costos computacionales y largos tiempos durante el cálculo de la optimización de la geometría molecular. Por lo tanto, el enfoque QSPR independiente de la conformación se puede considerar como una metodología muy útil.

Además, también exploramos el rendimiento de los modelos QSPR basados en descriptores óptimos[19]. Dentro de esta técnica, el descriptor óptimo calculado depende tanto de la estructura molecular como de la propiedad bajo análisis (*Koc*), pero no depende explícitamente de la geometría molecular tridimensional. Hemos demostrado la importancia de los descriptores óptimos en estudios anteriores QSPR[21–25].

4.2.1. Resultados y Discusión

Comenzamos nuestro estudio QSPR explorando el rendimiento de los descriptores moleculares calculados con el programa de libre acceso PaDEL. Las características estructurales más representativas del conjunto de calibración de 93 compuestos heterogéneos se buscan a través de la técnica RM.

De esta manera, los mejores modelos de MLR basados en 1-6 descriptores moleculares se encuentran en un grupo que tiene 17536 variables. Para eliminar los descriptores “colineales” (idénticos), los pares linealmente dependientes se identifican dentro de RM, y solo se conserva una variable de cada par para su posterior análisis. Este proceso conduce a un conjunto que contiene 3491 descriptores linealmente independientes. Seguimos la práctica común de mantener la dimensión (*d*) del modelo lo más pequeña posible.

Los mejores modelos de MLR se enumeran en la Tabla 4.1, mientras que en el apartado de Tablas anexas en (CD) se presenta la Tabla 4.3.A que proporciona una breve descripción del significado de los descriptores. Se aprecia en la Tabla 4.1 que el parámetro RMS_{cal} continúa mejorando más allá de cuatro descriptores, pero RMS_{pred} no mejora significativamente.

De acuerdo con ello, elegimos una relación estructura-propiedad que tiene cuatro descriptores con un poder predictivo aceptable en el conjunto de predicción:

$$\log Koc = 0.18\,SP3 + 0.30\,CrippenLogP - 0.090\,gmax + 0.16\,XLogP + 1.18 \qquad (4.2)$$

$N_{cal} = 93$, $R^2_{cal} = 0.87$, $RMS_{cal} = 0.45$, $R^2_{ij\max} = 0.58$, $o2.5 = 0$

$R^2_{loo} = 0.85$, $RMS_{loo} = 0.47$, $RMS^{aleat} = 1.02$

$N_{pred} = 550$, $R^2_{pred} = 0.81$, $RMS_{pred} = 0.53$

En esta ecuación, N es el número de compuestos; $R^2_{ij\max}$ denota el coeficiente de correlación máximo entre pares de descriptores; $o2.5$ indica el número de compuestos salientes en el conjunto de calibración que tiene un residuo (diferencia entre la actividad experimental y la calculada) mayor que 2.5 veces el valor de S_{cal}

Tabla 4.1. _Los mejores modelos QSPR lineales obtenidos a partir de descriptores independientes de la conformación. El modelo seleccionado aparece en negrita.

D	Descriptores	R^2_{cal}	R^2_{pred}	RMS_{cal}	RMS_{pred}
1	*CrippenLogP*	0.72	0.68	0.65	0.67
2	*CrippenLogP XLogP*	0.80	0.76	0.55	0.59
3	*CrippenLogP gmax TpiPC*	0.84	0.79	0.49	0.56
4	***SP3 CrippenLogP gmax XLogP***	**0.87**	**0.81**	**0.45**	**0.53**
5	*ALogp2 CrippenLogP maxHBint2 TpiPC XLogP*	0.87	0.81	0.44	0.52
6	*BCUTw-1l CrippenLogP gmax ETA_Epsilon_3 WPOL XLogP*	0.89	0.81	0.41	0.53

Los descriptores independientes de la conformación que aparecen en la ecuación (4.2) pertenecen a cuatro clases diferentes[26]: (i) un descriptor de

camino de Chi de PaDEL: *SP3*, camino de acceso simple 3; (ii) un descriptor de Crippen: *CrippenLogP*, *logP* de Crippen; (iii) un descriptor de tipo de átomo de estado electrotopológico: *gmax*, el máximo estado-E; y (iv) el descriptor *XLogP*.

En la Figura 4.1 se proporciona un gráfico para el logaritmo de *Koc* predicho como una función de los valores experimentales para los conjuntos de calibración y de prueba. El gráfico de dispersión de residuos en la Figura 4.1.A del modelo tiende a seguir un patrón aleatorio alrededor de la línea cero, sugiriendo que la suposición de la técnica de MLR se cumple.

La matriz de correlación para la ecuación (4.2) aparece en la Tabla 4.4.A, que muestra la ausencia de altas correlaciones entre los pares de descriptores, mientras que sus valores numéricos se incluyen en la Tabla 4.5.A. La ecuación (4.2), tiene una potencia predictiva aceptable en el conjunto de predicción externo de 550 compuestos, de acuerdo con el valor estadístico de R^2 y los parámetros de RMS_{pred} .

Tal modelo aprueba el proceso de validación interna de validación cruzada mediante la exclusión de una molécula a la vez. La técnica de aleatorización-Y demuestra que la ecuación (4.2) tiene un valor de RMS_{cal} menor a RMS^{aleat} y, por lo tanto, se encuentra una relación válida estructura-*logKoc*.

Los criterios de validación externos recomendados en Ref. 27 para asegurar la capacidad predictiva también se logran y se resumen en la Tabla 4.6.A.

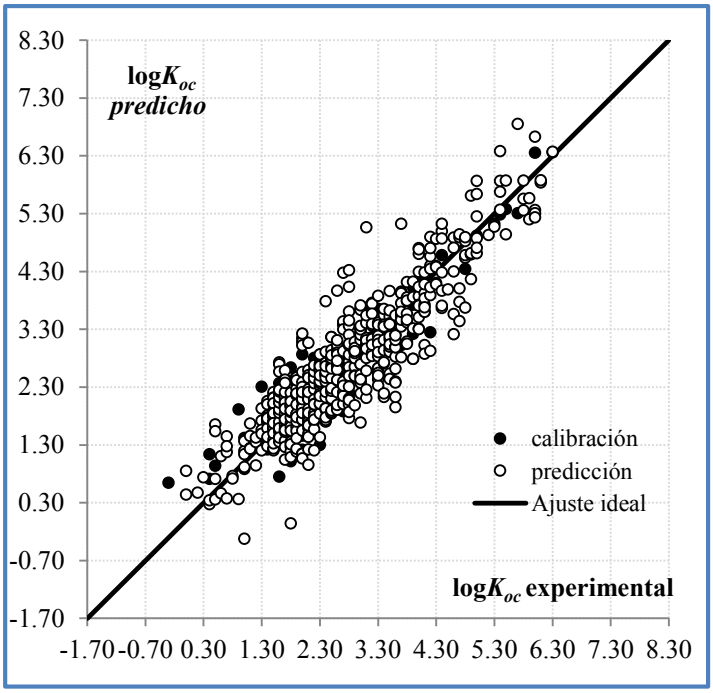


Figura 4.1. Se proporciona un gráfico para el logaritmo de Koc predicho como una función de los valores experimentales para los conjuntos de calibración y predicción.

La calidad estadística de la ecuación (4.2) es bastante similar a la de varios modelos QSPR informados anteriormente por Gramática *et al.*[17]. Por ejemplo, nuestro QSPR con RMS_{cal} = 0.45 y RMS_{pred} = 0.53 es mejor que el modelo publicado de cuatro descriptores topológicos con RMS_{cal} = 0.52 y RMS_{pred} = 0.56.

Además, la ecuación (4.2) también es comparable al modelo de consenso de tres descriptores propuesto en ese documento[17] (RMS_{cal} = 0.52 y RMS_{pred} = 0.53), aunque dicho modelo tiene la desventaja de incluir descriptores geométricos.

Como una siguiente etapa en este estudio QSPR, incluimos definiciones óptimas de descriptores moleculares para analizar el rendimiento de tales variables estructurales específicas de sorción en suelo.

El descriptor flexible *DCW* se optimiza aumentando el valor R^2 del conjunto de calibración, hasta que el modelo comience a perder capacidad predictiva en el conjunto de predicción (medido por su valor de RMS_{pred}).

La mejor representación estructural para los 93 compuestos de calibración es el grafo con hidrógenos, donde la calidad estadística para la evolución gradual del modelo lineal se presenta en la Tabla 4.1. El primer descriptor local seleccionado es *NNC* (representa el código del vecino más cercano a un determinado vértice), luego los siguientes atributos son *ec0* (conectividad extendida de Morgan de orden cero) y *NOSP* (la presencia de nitrógeno, oxígeno, azufre o fósforo) en ese orden.

En la Tabla 4.2 se observa que el descriptor óptimo de mejor calidad incluye tales tipos de tres variables, y 64 atributos activos se basan en ellos (se muestra en la Tabla 4.7.A). Los detalles más completos para el modelo QSPR son los siguientes:

$$\log Koc = 0.073\,DCW + 0.31 \qquad (4.3)$$

$N_{cal}=93$, $R^2_{cal}=0.87$, $RMS_{cal}=0.45$, $o2.5=1$, $R^2_{loo}=0.86$, $RMS_{loo}=0.45$

$RMS^{aleat}=1.11$, $N_{pred}=550$, $R^2_{pred}=0.76$, $RMS_{pred}=0.61$

Los parámetros utilizados para el cálculo de *DCW* son T = 1 y N^{iter} = 7. Las Figuras 4.2.A y 4.3.A demuestran que la técnica de MLR también se cumple para la ecuación (4.3). Un ejemplo para el cálculo de *DCW* para formaldehído se proporciona en la Tabla 4.2.

Tabla 4.2. La búsqueda por pasos para encontrar los mejores atributos estructurales construyendo el descriptor óptimo; el resultado seleccionado aparece en negrita.

Atributo estructural	R^2_{cal}	R^2_{pred}	RMS_{cal}	RMS_{pred}	N_{act}
NNC	0.84	0.73	0.49	0.64	50
NNC ec0	0.86	0.75	0.46	0.62	70
NNC ec0 NOSP	**0.87**	**0.76**	**0.45**	**0.61**	**64**

Tabla 4.3. Un ejemplo del cálculo del descriptor óptimo para formaldehído por la sumatoria de los valores de *CW*: *DCW* = -0.64892

atributo estructural	*CW*
ec0-O...1...	0.12508
ec0-C...3...	1.00094
ec0-H...1...	-0.18254
ec0-H...1...	-0.18254
NNC-O...101	0.24867
NNC-C...303	-0.75284
NNC-H...101	-0.07978
NNC-H...101	-0.07978
NOSP01000000	-0.74613

Nuestros resultados revelan que la ecuación (4.2) tiene un mejor rendimiento en el conjunto de predicción que la ecuación (4.3). Ambos modelos QSPR se obtienen a través de diferentes enfoques, es decir, permitiendo que el descriptor molecular que representa la estructura química dependa o no de la propiedad *logKoc* estudiada.

Como siguiente paso, investigamos lo que sucede cuando el conjunto anterior de descriptores 3491 0D-2D de PaDEL se combina con el descriptor *DCW* óptimo. Los mejores modelos del método MLR con 1-6 variables encontrados en dicho grupo de descriptores 3492 (Tabla 4.8.A) no mejoran el poder predictivo de nuestro primer modelo, ya que las estadísticas del conjunto de calibración son mejores, pero no ocurre lo mismo para el conjunto de predicción.

En un nuevo intento de mejorar la ecuación (4.2), consideramos la inclusión de las predicciones del programa EPI Suite como descriptores moleculares semiempíricos, calculados a través de los valores pronosticados *logKowEpi* y *logSwEpi*.

Después de buscar los mejores modelos de MLR en el conjunto compuesto por 3493 descriptores independientes de PaDEL y EPI Suite (Tabla 4.4), se logra la siguiente relación estructura-*Koc*:

$$\log Koc = 0.60\,MLFER.E - 0.36\,SubFP302 + 0.48\log KowEpi + 0.72 \qquad (4.4)$$

$N_{cal} = 93$, R^2_{cal}=0.87, RMS_{cal}=0.44, $R^2_{ij\max} = 0.21$, $o2.5 = 0$, R^2_{loo}=0.86, RMS_{loo}=0.46

RMS^{aleat}=1.02, $N_{pred} = 550$, $R^2_{pred} = 0.84$, $RMS_{pred} = 0.48$

El rendimiento de la ecuación (4.4) es mejor que la ecuación (4.2) y, por lo tanto, consideramos que este nuevo modelo QSPR es la relación estructura-coeficiente de sorción en suelo más adecuada para los 643 compuestos orgánicos no iónicos. La Figura 4.2 y la Figura 4.4.A grafican las predicciones, mientras que las Tablas 4.4.A y 4.6.A proporcionan la matriz de correlación y los criterios de validación externa para la ecuación (4.4).

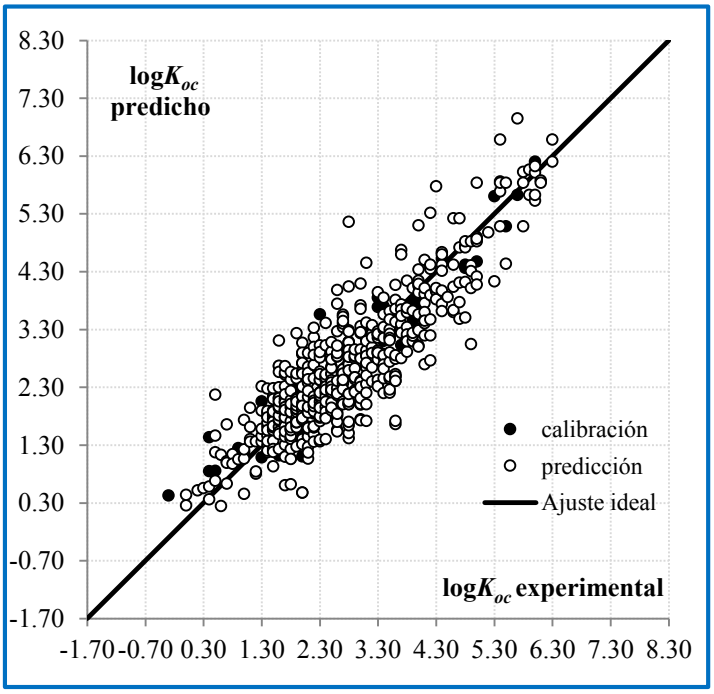


Figura 4.2. Valores experimentales y predichos para logKoc de acuerdo al modelo QSPR basado en la ecuación (4.4).

Tabla 4.4. Los mejores modelos QSPR lineales obtenidos a partir de un conjunto de 3493 descriptores independientes de la conformación obtenidos con PaDEL y EPI Suite; el modelo seleccionado aparece en negrita

d	Descriptores	R^2_{cal}	R^2_{pred}	RMS_{cal}	RMS_{pred}
1	*logKowEpi*	0.77	0.76	0.59	0.59

2	*LFER_E logKowEp*	0.86	0.83	0.46	0.50
3	***MLFER_E SubFP302 logKowEpi***	**0.87**	**0.84**	**0.44**	**0.48**
4	*mindO MLFER_E KRFP1105 logKowEpi*	0.87	0.84	0.42	0.48
5	*MAXDP2 ZMIC1 TpiPC KRFP3788 logKowEpi*	0.90	0.84	0.40	0.49
6	*ATSC3c AATSC3c MATS4p MLFER_E AD2D393 logKowEpi*	0.91	0.84	0.37	0.49

Los descriptores moleculares 2D que aparecen en esta última ecuación pertenecen a tres clases diferentes: (i) un descriptor de relación de energía libre lineal molecular (MLFER): *MLFER.E*, que mide la refracción molar excesiva; (ii) un descriptor indicador de subestructura: *SubFP302*, nos informa de la presencia de enlaces giratorios; y (iii) un descriptor del programa EPI Suite: *logKowEpi*, logaritmo del coeficiente de partición octanol/agua.

Como los descriptores toman valores numéricos positivos, la ecuación (4.4) indica que un compuesto que tenga valores más altos para los descriptores *MLFER.E* y *logKowEpi* con un valor menor para *SubFP302*, tenderá a tener un coeficiente de sorción en suelo más alto.

MLFER.E mide la refracción molar del soluto menos la refracción molar de un alcano de volumen equivalente. Este descriptor se puede estimar fácilmente a partir del conocimiento del índice de refracción de un compuesto, y sugiere la preferencia de la fase del suelo a interactuar con compuestos de soluto que tienen pares de electrones π y σ.

El descriptor *SubFP302* tiene una interpretación clara ya que cuantifica la presencia (igual a uno) o ausencia (igual a cero) de enlaces giratorios en la estructura química. Este descriptor indicador identifica enlaces giratorios que permiten la libre rotación alrededor de ellos mismos, es decir, cualquier enlace simple, que no está en un anillo, vinculado a un átomo pesado no terminal.

Finalmente, el logaritmo del coeficiente de partición octanol/agua del descriptor *logKowEpi* es una propiedad fisicoquímica bien conocida que ha sido ampliamente utilizada en estudios anteriores de modelos QSPR para correlacionar los valores del *logKoc*. Por lo tanto, los compuestos hidrofóbicos con altos valores de *logKowEpi* tienden a exhibir una mayor retención por parte de la materia orgánica de suelos y sedimentos.

El análisis del dominio de aplicabilidad (DA) del nuevo modelo QSPR propuesto revela que 16 compuestos están fuera de los 550 compuestos incluidos en el conjunto de predicción, por ello no pertenecen al dominio de aplicabilidad del modelo, por el valor obtenido de $h_i > h^* = 0.13$. Los valores de influencia (h_i) también se proporcionan en la Tabla 4.1.A. Asumimos que este comportamiento particular se debe a la complejidad del conjunto de datos, es decir, a la gran heterogeneidad estructural de las moléculas consideradas en este estudio. Por lo tanto, los valores de *logKoc* predichos para todos, con la excepción de tales 16 compuestos del conjunto de predicción, pueden considerarse igualmente confiables ya que caen dentro del DA.

Como comparación final, nuestro mejor modelo QSPR con RMS_{cal} =0.44 y RMS_{pred} =0.48 tiene un mejor rendimiento en los compuestos heterogéneos que el proporcionado por EPI Suite: RMS_{cal} =0.47 y RMS_{pred} =0.56 (método de conectividad) y RMS_{cal} =0.48 y RMS_{pred} =0.56 (método basado en el coeficiente de partición). Esto significa que nuestro modelo QSPR desarrollado a partir de la ecuación (4.4) representa una herramienta alternativa/complementaria al programa EPI Suite para predecir la propiedad estudiada en el presente conjunto de datos de 643 compuestos orgánicos no-iónicos.

4.2.2. Datos experimentales (643 moléculas)

Los valores experimentales del coeficiente de sorción en suelo fueron colectados de la literatura[17], dicho coeficiente es cuantificado como la tasa entre la concentración del compuesto químico en el suelo y el carbono orgánico normalizado en agua. En el presente conjunto de datos, los valores de *logKoc* se encuentran dentro del rango de intervalos de (-0.31, 6.02) en el conjunto de calibración y (0, 6.33) en el conjunto de predicción; la lista completa de los 643 compuestos estudiados aquí está incluida en la Tabla 1.A en el apartado de Tablas anexas en (CD).

El conjunto de datos es altamente heterogéneo, e incluye prácticamente todos los principales grupos funcionales presentes en pesticidas y varios compuestos orgánicos contaminantes.

Por otro lado, a fin de poder comparar los resultados, el logaritmo del coeficiente de partición de suelo es obtenido a través de la Interface del Programa de Estimación (EPI Suite) desde el módulo KOCWIN (*logKocEpi*)[28].

EPI Suite calcula *logKocEpi* a partir de dos técnicas diferentes: (a) basada en el índice de primer orden de conectividad molecular (MCI); y (b) basado en *logKow* (más que en MCI). En ambos casos, el programa emplea una serie de factores de contribución de grupos.

4.2.3. Descriptores moleculares

Las moléculas son primero dibujadas con el programa de acceso libre ACDLabs ChemSketch[29], con el formato molecular MDL mol (V2000).

El conjunto de descriptores moleculares independientes de la conformación es calculado con el programa Laboratorio de Exploración de Datos Farmacéuticos (PaDEL) 2.20[30], porque este tiene la ventaja de que está disponible de forma gratuita y es un programa de código abierto.

PaDEL actualmente calcula 1444 descriptores 0D–2D y 12 tipos de descriptores indicadores (16092 variables)[31]. Los descriptores indicadores involucran la presencia o el conteo de subestructuras químicas específicas: aquí tratamos a los descriptores indicadores como descriptores de tipo constitucional que describen la composición molecular.

Además, son adicionados los descriptores semiempíricos del programa EPI Suite, tales como, el logaritmo calculado del coeficiente de partición octanol/agua KOWWIN (*logKowEpi*) y el logaritmo calculado de la solubilidad en agua a partir del módulo WATERNT (*logSwEpi*)[28].

Por lo tanto, el número total de descriptores no-conformacionales explorados en este trabajo es 17538. Es nuestra intención poder capturar, con este gran número de descriptores, las características estructurales más relevantes que pueden afectar la propiedad estudiada.

4.2.4. Selección de los mejores descriptores moleculares

Nosotros empleamos la técnica del Método del Reemplazo (RM)[32–38] a fin de que los modelos MLR puedan ser generados en base al conjunto de calibración, en el análisis simultáneo de un gran conjunto que tiene D = 17538 descriptores con el fin de identificar un subconjunto óptimo de d descriptores (d es mucho menor que D), con valores más pequeños de desviación estándar (S_{cal}) o del valor de raíz cuadrada media de la desviación (RMS_{cal}). La Tabla 4.2.A incluye una lista de las ecuaciones matemáticas empleadas en el presente estudio. Todos los algoritmos usados en nuestros cálculos son programados bajo el lenguaje Matlab[39] y están disponibles.

4.2.5. Cálculo de descriptores flexibles

Por medio del programa de acceso libre CORAL[40] es fácil poder definir los diferentes tipos de descriptores óptimos. La representación estructural puede basarse sea en un grafo o un SMILES, lo que determina el tipo de atributo estructural (SA) o descriptor local disponible para poder construir el modelo QSPR.

Por otra parte, es necesario decidir la combinación de SAs que sean los más apropiados y esto se hace a partir del método de inclusión de a pasos, es decir, primero se busca el mejor SA simple, entonces, luego se busca el segundo mejor SA y se combina con el primer mejor SA, luego se combinan el primer y el segundo SA previamente seleccionados, con un nuevo mejor SA, y así sucesivamente.

El descriptor flexible u óptimo (*DCW*) es una combinación lineal de los pesos de correlación (*CW*); nos referimos a ellos en la Tabla 4.8.A. El *CW* es calculado para cada SA en el conjunto de calibración a través del método de simulación de Monte Carlo (MC). El *DCW* depende de dos parámetros para poder ser calculado: el número de umbral (T), y el número de iteraciones óptimas (N^{iter}), la selección apropiada del valor de T y N^{iter} evita el sobreajuste del modelo. Los atributos raros son los que ocurren en menos que T

compuestos, y en este trabajo T es un entero positivo analizado en el rango de 0-5.

4.2.6. Validación del modelo

Los modelos de regresión lineal son teóricamente validados a través del método de validación cruzada dejar-uno-fuera (loo)[27]. Una validación más confiable se basa en un conjunto de predicción externo de estructuras. La misma partición en conjuntos de calibración y prueba de la literatura[17] es usada en el presente análisis; esto quiere decir que 93 compuestos están en el conjunto de calibración y 550 compuestos están en el conjunto de predicción.

Además, nosotros mezclamos los valores de la propiedad experimental con la aleatorización-Y[41] y 10000 casos: es la forma de comprobar que el modelo no resulta de una correlación fortuita cuando el valor de RMS^{aleat} es más grande que RMS_{cal}.

4.2.7. Dominio de aplicación

Un modelo QSPR predictivo es aquel capaz de predecir las moléculas que caen dentro de su dominio de aplicación (DA)[42], por lo tanto, la capacidad predictiva de la propiedad no es resultado de la extrapolación sustancial (predicción no confiable). La definición del DA es dependiente de los descriptores empleados para construir el modelo y de la propiedad experimental.

Dentro del enfoque de influencia[43], los compuestos del conjunto de predicción deben tener un valor de influencia calculado (h_i) más pequeño que el valor de influencia de control (h^*).

4.2.8. Conclusiones

Hemos establecido satisfactoriamente un modelo de relación estructura-propiedad para el coeficiente de sorción en suelo, un parámetro

útil que relaciona el proceso de sorción en suelo y, por lo tanto, determina la distribución, el destino y la persistencia de los compuestos químicos en el medio ambiente. El domino de aplicación de los compuestos químicos explorados incluye un conjunto de 643 compuestos heterogéneos de moléculas orgánicas no-iónicas, que tienen un rango de más de seis unidades logarítmicas de la propiedad.

El modelo QSPR encontrado para el conjunto de calibración de 93 compuestos tiene una capacidad predictiva aceptable en el conjunto de predicción, que incluye a 550 compuestos, y es capaz de cumplir con otras condiciones matemáticas, tales como, la validación cruzada, aleatorización-Y, y el análisis de dominio de aplicabilidad.

En el estudio actual, la formulación del descriptor flexible no mejora la calidad de las predicciones del coeficiente de sorción en suelo, pero sí se logra encontrar un modelo aceptable que incluye descriptores semiempíricos calculados con el programa EPI Suite.

Nuestros resultados se comparan favorablemente con otros trabajos previamente reportados en la literatura. Los modelos que hemos propuesto involucran descriptores calculados a través de programas computacionales de libre acceso como PaDEL, CORAL y EPI Suite.

El trabajo de investigación realizado está enfocado en el uso de nuevos métodos basados en aproximaciones constitucionales y topológicas de los estudios QSPR. Hemos desarrollado un enfoque QSPR independiente de la conformación, de manera de evitar la representación conformacional de las estructuras químicas, y por lo tanto, no es necesaria la información experimental de la estructura cristalina u otra información de la geometría experimental.

4.3. Predicción del Factor de Bioconcentración con QSPR

El factor de bioconcentración (BCF) representa la capacidad de bioconcentración de un compuesto químico, definido como la tasa entre su concentración en el organismo y la concentración en agua en estado de

equilibrio estable bajo condiciones de laboratorio. El entorno acuático es a menudo el sumidero final de muchos contaminantes, debido a la inmisión directa o los procesos hidrológicos/atmosféricos.

Un paso clave para comprender el efecto de los compuestos químicos en la biota es describir la relación entre las concentraciones en los organismos, en el medio ambiente y la toxicidad potencial. Los procesos capaces de influir en estas relaciones son la bioconcentración, bioacumulación y biomagnificación.

La bioconcentración y la bioacumulación son procesos por los cuales la concentración de un químico en un organismo excede la concentración en el ambiente circundante. El primero solo tiene en cuenta la exposición a través de rutas no dietéticas (por ejemplo, superficies respiratorias y dérmicas), mientras que el segundo se refiere a todas las rutas de exposición posibles[44]. La combinación de bioconcentración y bioacumulación dentro de la cadena alimentaria da como resultado la denominada biomagnificación, es decir, el aumento de la concentración química en la cadena alimentaria al aumentar el nivel trófico[45].

Cuando ocurren estos procesos, los organismos (especialmente en la parte superior de la cadena trófica) pueden llegar a estar contaminados, con efectos a largo plazo difíciles de predecir. Esta es la razón por lo que las capacidades de bioacumulación y bioconcentración constituyen un riesgo ecológico, que deben tenerse en cuenta en la evaluación de riesgos ambientales de los productos químicos.

El Reglamento Europeo REACH (EC No 1907/2006) identifica la evaluación de riesgos de sustancias bioacumulativas como una prioridad. Para este propósito, se pueden usar datos medidos de BCF.

El parámetro BCF es un punto final de gran importancia, debido a su impacto en ecotoxicología: las sustancias son identificadas como bioacumulativas cuando su valor del *logBCF* es mayor a 3.3, y no serían acumulativas por debajo de este límite. Como alternativa, se puede utilizar el coeficiente de partición octanol-agua (*Kow*) como criterio de selección (*logKow*>4.5)[46]. Los peces generalmente se utilizan para evaluar la

contaminación, debido a su papel en la cadena trófica y la disponibilidad de protocolos de prueba estandarizados[47].

La determinación experimental de BCF es, sin embargo, costosa (aproximadamente 35000 euros por cada compuesto químico) y requiere el uso de más de 100 animales para cada estudio estándar[48]. Esta es la razón por la cual la regulación REACH identifica a la metodología QSAR[49] como una herramienta valiosa para la reducción de pruebas innecesarias en animales y para complementar la falta de datos experimentales.

Desde finales de la década de 1970, se han realizado grandes esfuerzos para predecir el valor de BCF a partir de la estructura molecular o propiedades medidas[50]. El método más común es establecer una relación empírica con *Kow*, como modelo lineal[51,52], bilineal[53] y como un polinomio[54].

Durante las últimas décadas, se introdujeron modelos más complejos, como el modelo basado en fragmentos de Meylan *et al.*[55] o el modelo híbrido de Zhao *et al.*[56].

Hoy en día, los principales aspectos críticos de la predicción del valor de BCF están relacionados con:

(1) la disponibilidad de varios modelos para BCF, que plantea el problema sobre cual modelo debe ser utilizado con productos químicos nuevos no probados.

(2) el amplio uso de *Kow* experimental o calculado como descriptor principal. La bioconcentración es, de hecho, el resultado neto del reparto de lípidos y agua (*Kow*) y otros procesos, como el metabolismo, la dilución del crecimiento, la eliminación fecal y la excreción e interacciones específicas con tejidos[44]. Cuando se producen estos procesos, el valor de BCF estimado a partir de *Kow* puede diferir del valor de BCF real.

(3) la compensación óptima entre sesgo, complejidad y significado mecanicista. De hecho, al aumentar la complejidad del modelo (es decir, el número de variables a parámetros incluidos), aumenta el ajuste a los datos de calibración. Sin embargo, a menudo causa un ajuste excesivo (disminuyendo la capacidad de predicción) y no garantiza una mayor cantidad de procesos

contabilizados o una mejor comprensión. Por otro lado, cuando los modelos son demasiado simples, no son capaces de capturar mecanismos importantes que influyen en la respuesta y, por lo tanto, la precisión de la predicción disminuye.

El presente trabajo recurre a la misma base de datos con valores conocidos y verificados de BCF experimental empleada por Gissi *et al.*[57], con el fin de reportar un modelo QSPR como una nueva alternativa que involucra información de pesticidas.

El enfoque de modelos QSPR independientes de la conformación[58–61] empleados aquí, no consideran la representación conformacional de la estructura química, porque solo relacionan sus representaciones constitucionales y topológicas. Es interesante notar que este enfoque es independiente de la conformación aunque no es independiente de la geometría, ya que un grafo depende de la geometría molecular.

Al igual que en el estudio anterior sobre el coeficiente de sorción en suelo, la exclusión de los aspectos estructurales 3D se realiza con el fin de evitar ambigüedades debido a la existencia de los compuestos químicos en varios estados conformacionales, lo cual puede llevar a una menor capacidad predictiva del modelo QSPR debido a la incorrecta representación de tales estados.

4.3.1. Datos experimentales de ANTARES (851 moléculas)

La base de datos ANTARES[57,62] incluye valores experimentales de BCF recolectados de entre varias base de datos fiables y que se encuentran disponibles públicamente. Esta base de datos contiene compuestos con diferentes clases químicas, de los cuales 159 compuestos son pesticidas.

Compuestos caracterizados con datos ambiguos, compuestos inorgánicos, o mezclas isoméricas fueron descartados[57], esto lleva a un conjunto que contiene 851 valores experimentales de BCF en un intervalo de -1.70-5.69. La lista completa de compuestos estudiados aquí está presentada en la Tabla 9.A.

4.3.2. Obtención de los descriptores moleculares

Las 851 estructuras moleculares fueron primero dibujadas con el programa gratuito ACDLabs ChemSketch[29] con el formato molecular en MDL mol (V2000). La conversión de las moléculas se realizó con el programa para Windows Open Babel[63]. Los descriptores moleculares independientes de la conformación fueron calculados como se comenta a continuación.

Utilizamos el programa PaDEL versión 2.20 para el cálculo de descriptores moleculares, incluyendo a descriptores indicadores.

Cinco descriptores semiempíricos, fueron calculados a partir de los módulos del programa EPI Suite[28], que emplea el formato molecular SMILES. EPI Suite utiliza una serie de factores de grupos de contribución para calcular (en unidades logarítmicas decimales): (i) el coeficiente de partición octanol/agua *logKowEpi*; (ii) la solubilidad en agua *logSw1Epi* y *logSw2Epi*: el segundo parámetro está basado en el *logKowEpi*; (iii) el coeficiente de sorción en suelo *logKoc1* y el *logKoc2*: el primer parámetro está basado en el índice de conectividad de primer orden (MCI), mientras que el segundo está basado en el *logKowEpi*.

También hemos calculado el factor de bioconcentración *logBCFEpi* con el fin de comparar las predicciones del programa EPI Suite con aquellas obtenidas en nuestro trabajo.

Se calculó el mejor descriptores flexible (*DCW*) mediante CORAL, ya que este tipo de descriptor demostró su importancia en estudios QSPR-QSAR previos[24,25,61,64].

Otros descriptores moleculares fueron calculados con el programa de libre acceso Descriptores Moleculares de Estructuras 2D (Mold[2])[65], el cual genera 777 variables estructurales 1D–2D a partir del formato molecular MDL sdf.

Los descriptores basados en densidades de carga topológica fueron calculados mediante el programa gratuito RECON 5.5[66] el cual codifica información electrónica y estructural importante para las interacciones moleculares. La robustez de RECON ha sido previamente demostrada en otros

trabajos[67]. RECON es un algoritmo para la reconstrucción rápida de densidades de carga molecular y de las propiedades electrónicas de carga basadas en la densidad molecular, para lo cual dispone pre-calculados fragmentos de densidades de carga atómica de funciones de onda *ab initio*. El método está basado en la Teoría de Bader de átomos en moléculas (AIM)[68]. Una librería de fragmentos de densidad de carga atómica ha sido construida con el objetivo de permitir la recuperación rápida de los fragmentos y del montaje molecular. En el presente caso, las moléculas fueron calculadas en formato SMILES para poder obtener 248 descriptores "Átomos Equivalentes Transferibles (TAE)"[69].

Finalmente, otro tipo de descriptores moleculares 2D fueron calculados con el programa Mapas Cuadráticos, Bilineales y N-Lineales (QuBiLs)[70], basados en el uso de matrices de densidad electrónica grafo-teóricas y pesos atómicos, que se encuentra en el módulo del programa gratuito multi-plataforma ToMoCoMD-CARDD. Para ello utiliza el formato molecular MDL sdf.

El módulo QuBiLs-MAS calcula 8448 descriptores algebraicos cuando se seleccionan las múltiples opciones siguientes: a) formas algebraicas: 'bilineal', 'lineal', 'cuadrática'; b) restricciones: 'basada en el átomo', 'no-quiral', y 'duplex'; c) formas matriciales (máximo orden 15): 'no-estocástica', 'estocástica simple', 'estocástica doble', y 'probabilidad mutua'; corte: 'mantener todo'; grupos: 'totales'; propiedades: 'LogP Ghose-Crippen', 'polarizabilidad', 'carga', 'área superficial polar', 'electronegatividad', 'refractividad', 'masa', y 'volumen de Van der Waals'; invariantes: 'distancia euclideana', 'media aritmética', y 'desviación estándar', junto con opción no-estandarizado.

Luego de obtener todos los descriptores mediante los diferentes programas de cálculo, el número total de variables no-conformacionales alcanza un total de 27017.

4.3.3. Partición molecular con el Método de Subconjuntos Balanceados

Varias estrategias de validación han sido propuestas durante los últimos años en el desarrollo de los modelos QSPR[71-73], las cuales consisten en la capacidad para predecir la propiedad de los compuestos no considerados durante el desarrollo del modelo.

Con esta idea, el conjunto molecular completo es particionado en tres conjuntos: calibración (cal), validación (val) y predicción (pred). El conjunto de calibración es usado para ajustar el modelo y sus parámetros a través de la técnica de RM, mientras que el conjunto de validación ayuda a aceptar el modelo parcialmente. Finalmente, el conjunto de predicción incluye compuestos "nunca vistos" durante la etapa de calibración, y demuestran la verdadera capacidad predictiva del modelo QSPR ensayado.

Una partición molecular debe lograr establecer relaciones estructura-propiedad que resulten similares en los tres conjuntos; en otras palabras, las moléculas del conjunto de calibración deben ser representativas de las moléculas incluidas en los conjuntos de validación y de predicción.

Para poder cumplir con este objetivo, la división del conjunto de datos fue llevada a cabo por medio del Método de Subconjuntos Balanceados (BSM)[74,75], un procedimiento propuesto por nuestro grupo que permite obtener conjuntos balanceados.

El BSM está basado en el método de Análisis de Agrupamiento de k-medias (k-MCA)[76]: la esencia de k-MCA es crear k-agrupamientos o grupos de compuestos, de manera que, los compuestos en el mismo grupo son muy similares en términos de distancias métricas (por ejemplo, en base a la distancia euclídea), y compuestos en diferentes grupos son muy diferentes.

Esto genera conjuntos de calibración, validación y predicción que pueden cumplir con el dominio de aplicación del modelo en el rango de la propiedad experimental bajo estudio, y cuyas estructuras químicas son similares en todos los conjuntos, en línea con el principio de BSM.

4.3.4. Búsqueda del modelo QSPR

Los 27017 descriptores moleculares no-conformacionales calculados con PaDEL, EPI Suite, CORAL, Mold[2], RECON, y QuBiLs-MAS fueron analizados con el objetivo de remover los descriptores "colineales". En este sentido, se identificaron los pares de descriptores linealmente dependientes, y sólo una variable de cada par se mantuvo.

Por lo tanto, los descriptores carentes de información relevante son removidos, al igual que aquellas variables con valores constantes o casi constantes, y también aquellas variables que tienen al menos un valor faltante. Este procedimiento conduce a 8122 descriptores estructurales linealmente independientes para el análisis.

Empleamos la técnica del Método del Reemplazo (RM) aplicada al conjunto de calibración para seleccionar los descriptores moleculares más representativos de la propiedad.

La Tabla 4.2.A incluye una lista de las ecuaciones matemáticas utilizadas en el presente estudio. Todos los algoritmos usados en nuestros cálculos son programados bajo el lenguaje MATLAB[39] y están disponibles para posteriores consultas.

Los modelos de regresión lineal son validados teóricamente a través del procedimiento de validación cruzada dejar-uno-fuera (loo)[27], y además, a través de una rigurosa validación cruzada dejar-30%-fuera (lmo), con 200000 casos. De acuerdo a Golbraikh y Tropsha[27], la varianza explicada en validación cruzada (R^2_{loo} y $R^2_{l30\%o}$) debe ser mayor que 0.5, aunque esto es necesario, no es suficiente para demostrar el poder predictivo real.

Los modelos establecidos son a su vez, validados con un nuevo criterio basado en el error absoluto medio (*MAE*)[73]. La calidad de las estimaciones del conjunto de predicción se determina a través del parámetro *MAE* y su desviación estándar σ, ambos parámetros son calculados después de omitir el 5% de las moléculas del conjunto de predicción que poseen valores altos del residuo. Este procedimiento evita la influencia de errores de predicción muy

altos, que pueden afectar demasiado la calidad de las predicciones en el conjunto de predicción externo.

Con el fin de obtener buenas predicciones en el conjunto de predicción, es necesario considerar que un error del 10% en el rango del conjunto de calibración debería ser aceptable, mientras que un valor de error mayor del 20% en el rango del conjunto de calibración debería ser un error muy grande.

Finalmente, aplicamos la técnica de aleatorización-Y[41] y 10000 casos, de manera de comprobar que el modelo no resulta de una correlación fortuita cuando el valor de RMS^{aleat} es más grande que RMS_{cal} .

En este trabajo, el DA es determinado a través de dos metodologías alternativas. La primera está basada en un enfoque bien conocido de influencia[43], donde los compuestos del conjunto de predicción deben tener un valor de influencia calculado menor al valor de influencia límite h^*. El segundo está basado en un enfoque de estandarización[77]: un compuesto dado i del conjunto de predicción tiene d valores de descriptores estandarizados s_{ik}, k = 1, ..., d con un valor máximo de $s_{ik}^{\max} \leq 3$. En el caso de que $s_{ik}^{\max} > 3$ y su valor mínimo $s_{ik}^{\min} < 3$, entonces el parámetro s_i^{nuevo} tiene que ser calculado y debe cumplir completamente la condición:

$s_i^{nuevo} = \langle s_i \rangle + 1.28.\sigma_{si} \leq 3$, donde $\langle s_i \rangle$ es la media de los valores s_{ik} para i y σ_{si} es la desviación estándar para cada valor.

Por último, para poder encontrar la importancia relativa del *j*-ésimo descriptor del modelo lineal, los coeficientes de regresión fueron estandarizados (b_j^s). Cuanto mayor sea el valor absoluto del coeficiente estandarizado de un descriptor dado, mayor será la importancia en el modelo de tal descriptor[78].

4.3.5. Resultados y Discusión

La técnica BSM fue aplicada a la base de datos de ANTARES de 851 compuestos heterogéneos, a partir de allí, se generaron subconjuntos

balanceados de tamaños similares. La Tabla 4.9.A indica los miembros de cada conjunto como validación (^) y predicción (*). De esta manera, los compuestos de calibración y validación constituyen el 66.75 % de toda la base de datos.

Los descriptores moleculares más representativos son buscados en el conjunto de calibración a través del método de selección de variables RM. El mejor modelo se basa en 1-7 descriptores teóricos que están indicados en la Tabla 4.5, mientras que una breve descripción del significado de los descriptores es añadida en la Tabla 4.10.A.

Tabla 4.5. El mejor modelo QSPR multidimensional de BCF. El modelo seleccionado se encuentra en negrita

d	descriptores	R^2_{cal}	RMS_{cal}	R^2_{val}	RMS_{val}	R^2_{pred}	RMS_{pred}
1	*DCW*	0.64	0.83	0.50	0.91	0.54	0.85
2	*PC406; DCW*	0.66	0.80	0.54	0.88	0.57	0.82
3	*GATS3c; Sub295; DCW*	0.69	0.77	0.56	0.86	0.62	0.77
4	*GATS3c; Sub295; AP402; DCW*	0.70	0.76	0.58	0.84	0.62	0.77
5	*AATS5e; GATS3c; Sub295; K1406; DCW*	0.71	0.74	0.58	0.83	0.65	0.75
6	***ATS8m; AATS5e; Sub295; K1406; AP391; DCW***	**0.73**	**0.71**	**0.60**	**0.81**	**0.67**	**0.71**
7	*ATS8m; AP391; DCW; D708; FPIP9; SD_B_AB_nCi_2_NS0_T_KA_ psa-e_MAS; N2_B_AB_nCi_2_ NS3_T_KA_a-psa_MAS*	0.75	0.69	0.60	0.82	0.67	0.71

De la Tabla 4.5, se observa que los parámetros de calibración continúan mejorando con la adición de cada nuevo descriptor molecular en la ecuación lineal, un comportamiento característico en la selección de variables, pero RMS_{val} no mejora significativamente después de la adición del sexto descriptor.

Con el fin de mantener la dimensión del modelo lo más pequeña como sea posible, seleccionamos dicho modelo como la mejor regresión lineal QSPR:

$$\log BCF = -3.06.10^{-5}\, ATS8m + 0.071\, AATS5e - 0.70 Sub295 - 0.87\, K1406 + \\ + 0.48\, AP391 + 0.069\, DCW + 0.51 \qquad (4.5)$$

$N_{cal} = 284$, $R^2_{cal} = 0.73$, $RMS_{cal} = 0.71$, $R^2_{ij\max} = 0.13$, $o3 = 0$

$R^2_{aleat} = 0.10$, $RMS^{aleat} = 1.31$, $R^2_{loo} = 0.72$, $RMS_{loo} = 0.73$, $R^2_{l30\%o} = 0.67$, $RMS_{l30\%o} = 0.79$

$N_{val} = 284$, $R^2_{val} = 0.60$, $RMS_{val} = 0.81$

$N_{pred} = 283$, $R^2_{pred} = 0.67$, $RMS_{pred} = 0.71$

De estos resultados $R^2_{ij\max}$ indica la ausencia de correlaciones serias entre los seis descriptores seleccionados. La ecuación (4.5) no involucra compuestos del conjunto de calibración con valores altos de residuos.

El gráfico que relaciona las predicciones con los valores experimentales se presenta en la Figura 4.3. El gráfico de dispersión de residuos en la Figura 4.5.A tiende a obedecer un patrón aleatorio alrededor de la línea de cero, lo que sugiere que la ecuación (4.5) predice todo el conjunto de datos sin errores sistemáticos.

El modelo QSPR de la ecuación (4.5) tiene una calidad aceptable en el conjunto de predicción externo de 283 valores de *BCF* según los parámetros R^2_{pred} y RMS_{pred}. Tal modelo aprueba el proceso de validación interna de dejar-uno-afuera y dejar-30%-afuera como proceso de validación cruzada, a través de la predicción de una o más moléculas excluidas a la vez del conjunto de calibración.

La técnica de aleatorización-Y demuestra que el modelo tiene valores de $RMS_{cal} < RMS^{aleat}$ y $R^2_{aleat} < R^2_{cal}$, y que hay una relación válida entre la estructura y la propiedad, sin correlaciones aleatorias. Además, también se logra superar los criterios de validación externos recomendados[27] para asegurar la capacidad predictiva:

$$1 - R^2_0 / R^2_{pred}\,(6.04.10^{-4}) < 0.1 \;\text{ o }\; 1 - R^2_0 / R^2_{pred}\,(0.21) < 0.1;$$
$$0.85 \le k\,(1.0035) \le 1.15 \text{ o } 0.85 \le k\,(0.9095) \le 1.15;$$
$$R^2_m\,(0.66) > 0.5$$

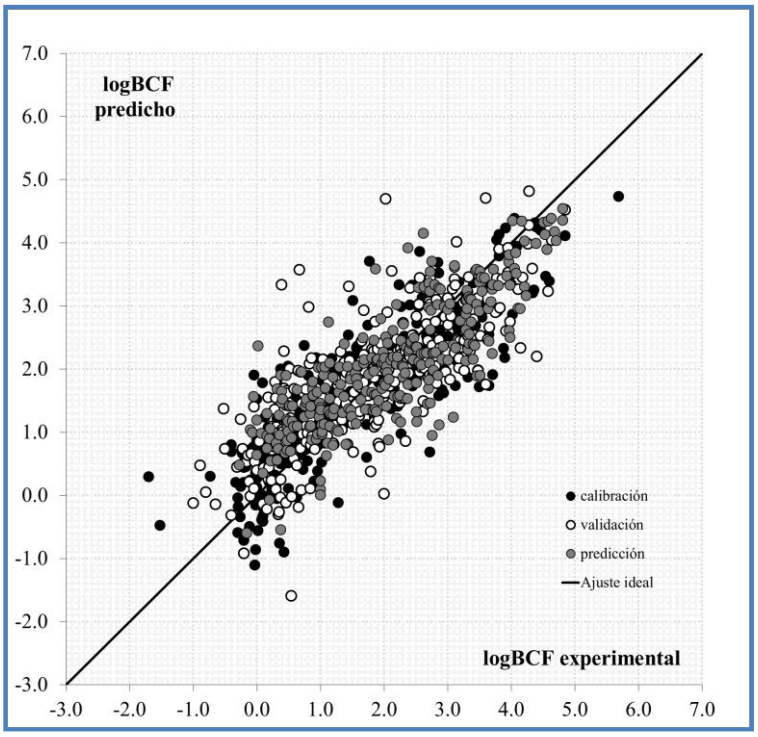


Figura 4.3. Valores predichos y experimentales de logBCF según la ecuación (4.5).

El rendimiento de nuestro modelo QSPR en los 283 compuestos del conjunto de predicción se encuentra clasificado como "intermedio" según el criterio MAE, lo que significa que es un modelo aceptable[73]. Para el conjunto de predicción, *MAE*(100%) = 0.57 y σ(100%) = 0.43, mientras que si se omite el 5% de los compuestos con altos valores de residuos conduce a *MAE*(95%) = 0.51 y σ(95%) = 0.34.

Los seis descriptores moleculares independientes de la conformación que forman parte de la relación cuantitativa estructura-propiedad propuesta para *logBCF* pertenecen a diferentes clases[49,79,80]:

i. dos descriptores de autocorrelación de la estructura topológica: *ATS8m*, la autocorrelación de Broto Moreau-distancia 8/ponderado por masa, y *AATS5e*, la autocorrelación media de Broto-Moreau-distancia 5/ponderado por las electronegatividades de Sanderson.

Las variables estructurales introducidas por Broto-Moreau son autocorrelaciones bidimensionales entre pares de átomos (*i*, *j*) en una molécula, con el objetivo principal de capturar el grado de interacción entre

ellos. La naturaleza de los átomos se considera a través de una propiedad dada (w), es decir, peso atómico, masa atómica, polarizabilidad, electronegatividad o volumen. Estos índices se calculan a partir del grafo al sumar los productos de los términos $w_i.w_j$, incluyendo las contribuciones atómicas terminales, en todos los caminos de una longitud establecida (lag).

ii. un descriptor de CORAL: *DCW*, descriptor óptimo basado en los atributos de HSG ec2 y SMILES. En el enfoque de grafo, ec2 es el índice de conectividad extendida de Morgan de segundo orden. Cabe señalar que el índice de orden cero ec0 para el vértice (átomo) *j* representa el grado de vértice para *j* (número de átomos vecinos), mientras que los índices de orden superior eck se obtienen a través de una fórmula recursiva basada en ec0[24,25]. En el enfoque basado en SMILES, *s* representa un atributo de un elemento: es decir, si un SMILES es una secuencia de elementos como 'ABCDE', entonces el atributo estructural *s* puede representarse con 'A', 'B', 'C', 'D',' E'.

Los siguientes descriptores tienen una interpretación estructural directa:

iii. un descriptor indicador 2D de pares de átomos: *AP391*, la presencia de C-C a la distancia topológica 6;

iv. un descriptor indicador de Klekota Roth: *K1406*, que indica la presencia de un patrón SMARTS $[!\#1]C(=O)[OH]$; y

v. un descriptor indicador de subestructura: *Sub295*, la presencia de un enlace C-ONS.

Todos los descriptores moleculares de la ecuación (4.5) tienen valores numéricos positivos con excepción de *DCW,* que puede tener valores positivos o negativos. El signo del coeficiente de regresión en el modelo lineal indica si la contribución del descriptor aumenta o disminuye el valor predicho de *logBCF*. Los valores numéricos positivos más altos de *DCW*, *AATS5e* y *AP391* y valores más bajos para *ATS8m*, *Sub295* y *K1406* tienden a predecir valores de *logBCF* más altos.

Después de la estandarización, el descriptor más importante de la ecuación (4.5) es *DCW* ($b_j^s = 0.66$) , teniendo así valores numéricos que

cambian más de acuerdo con las variaciones numéricas de la propiedad experimental. Los descriptores restantes *ATS8m* ($b_j^s = 0.14$), *AATS5e* ($b_j^s = 0.12$), *Sub295* ($b_j^s = 0.21$), *K1406* ($b_j^s = 0.16$) y *AP391* ($b_j^s = 0.17$) se complementan dentro de la ecuación lineal y tienen una relevancia comparable.

La matriz de correlación al cuadrado del modelo se proporciona en la Tabla 4.11.A, que muestra la ausencia de altas correlaciones entre pares de descriptores, como se mencionó anteriormente. También calculamos el factor de inflación de la varianza (*VIF*), un parámetro que mide la multicolinealidad entre los descriptores. Un valor de *VIF* de 1 para un descriptor específico significa que no existe una correlación entre este descriptor y todos los descriptores restantes del modelo, y un valor de *VIF* superior a 10 indica que la multicolinealidad es un problema en el conjunto de datos[81]. De la Tabla 4.11.A, se observa que el parámetro *VIF* para cada descriptor de la ecuación (4.5) está cerca de 1. Los valores numéricos de los descriptores aparecen en la Tabla 4.12.A.

Ahora demostramos que el modelo QSPR propuesto mediante la ecuación (4.5) es generalizable y útil para su aplicación, es decir, nuestro modelo no está determinado solo por la composición del conjunto de calibración debido a la partición de conjuntos de datos específicos con la técnica BSM. Para esto, realizamos 1000 operaciones diferentes de partición molecular aleatoria y recalculamos las estadísticas del modelo propuesto por nosotros en el presente trabajo. Descubrimos que, para 1000 conjuntos de predicción aleatorios externos, la ecuación (4.5) conduce a R^2_{pred} que varían entre 0.56-0.76 y valores de RMS_{pred} que van desde 0.66-0.85.

Estas conclusiones sugieren que el modelo final de la ecuación (4.5) tiene una estabilidad aceptable en su capacidad predictiva. La buena predicción de nuestro modelo QSPR en el conjunto de predicción no es casual, y los descriptores moleculares implicados en la ecuación (4.5) funcionan satisfactoriamente en las diferentes particiones de conjuntos de calibración y de predicción.

Siguiendo con la exploración del dominio de aplicación del modelo QSPR desarrollado, un compuesto con alta influencia reforzaría el modelo si el compuesto está en el conjunto de calibración, pero tal compuesto en el conjunto de predicción podría tener datos predichos poco confiables, como resultado de una extrapolación sustancial del modelo[42].

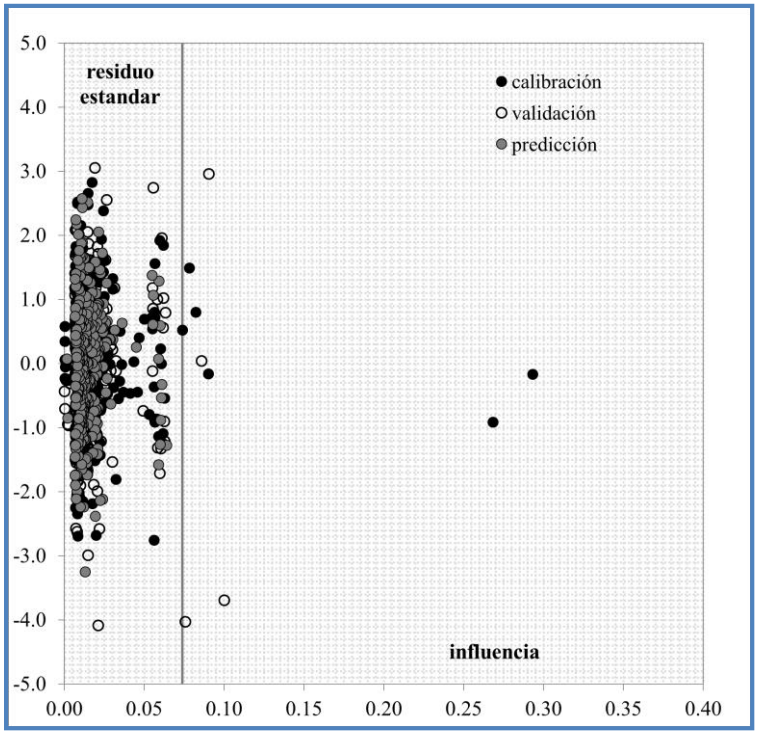


Figura 4.4. Diagrama de Williams para la ecuación (4.5). La línea indica el valor de influencia de control de 0.0739.

En nuestro caso, se encontró que los 283 compuestos del conjunto de predicción pertenecían al DA, ya que sus valores de influencia caen por debajo del límite h^* (0.0739). El diagrama de Williams para la ecuación (4.5) (residuos estandarizados en función de los valores de influencia) se proporciona en la Figura 4.4. Algunos compuestos que pertenecen a los conjuntos de calibración y validación tienen altos niveles de influencia que refuerzan el modelo, como los compuestos químicos **41**, **59**, **265**, **403**, **427**, **468**, **504**, **505**, **522** y **659**.

Este resultado obtenido con el enfoque de influencia para el conjunto de predicción coincide con el obtenido al utilizar el enfoque de estandarización, ya que las dos condiciones $s_{ik}^{\max} \leq 3$ o $s_i^{nuevo} \leq 3$ son seguidas por todos los 283 compuestos del conjunto de predicción. Por lo tanto, los valores de log*BCF*

predicho para los compuestos del conjunto de predicción se pueden considerar como confiables.

Algunos compuestos tienen residuos estandarizados superiores a tres unidades: esto puede ser puramente atribuido al conjunto de datos estructuralmente heterogéneo de 851 compuestos, que no se puede esperar que se modelen utilizando solo un modelo de 6 descriptores (ecuación 4.5).

Se puede hacer una comparación entre la calidad estadística de nuestro modelo QSPR propuesto en la ecuación (4.5) y el reportado por Gissi *et al.*[57]. Por medio de 836 compuestos en una división 608:152:76 ($N_{cal} : N_{val} : N_{pred}$), la calidad estadística lograda por el modelo ANN reportado de 9 descriptores aparece resumida en la Tabla 5.6 (modelo-1). Consideramos que nuestro modelo mejora dicho resultado reportado, principalmente por las siguientes cuatro razones que exponemos:

a) número de moléculas consideradas: contemplamos las 851 moléculas en el estudio QSPR propuesto sin excluir ninguna, contrariamente al modelo QSPR reportado, que emplea 836 compuestos y excluye 15 compuestos debido a limitaciones en el software de cálculo de los descriptores.

b) tamaño del modelo: la ecuación (4.5) involucra seis descriptores en lugar de nueve.

c) idoneidad de la partición del conjunto de datos: utilizamos una división 284:284:283, mientras que el modelo reportado usa 608:152:76. Por lo tanto, se consideran más compuestos en el conjunto de predicción durante el presente estudio QSPR que para determinar la capacidad predictiva en el modelo-1.

d) simplicidad: nuestro modelo lineal es más simple que el modelo ANN no lineal informado, y no depende de las conformaciones moleculares de los compuestos heterogéneos.

Mediante la definición del dominio de aplicación del modelo-1 reportado a través de cuatro métodos de filtrado independientes[57], 27 compuestos se excluyen adicionalmente de los conjuntos de validación y predicción (en total 42 compuestos excluidos del conjunto de datos iniciales).

Aunque el modelo-2 obtiene un mejor resultado estadístico en comparación con el modelo-1 (Tabla 4.6), dicho modelo considera solo el 8.53% de los compuestos en el conjunto de predicción. En cambio, el 33.25% de los compuestos de predicción es considerado por la ecuación (4.5). De hecho, nuestro modelo propuesto conduce a un mejor resultado en los 283 compuestos del conjunto de predicción con RMS_{pred} = 0.71, en comparación con RMS_{pred} = 0.82 para el modelo-2 en 69 compuestos.

Finalmente, comparamos los valores predichos de log*BCF* obtenidos por la ecuación (4.5) con las predicciones calculadas mediante el uso del módulo BCFBAF del programa EPI Suite, en la misma partición de BSM utilizada en este trabajo. De la Tabla 4.6 se encontraron estadísticas similares para los conjuntos de calibración, validación y predicción, aunque se lograron mediante metodologías diferentes en ambos casos. Sin embargo, cuando se trazan las predicciones en función de los valores experimentales para EPI Suite en la Figura 4.5, junto con el gráfico de dispersión de residuos en la Figura 4.6.A, se observa que muchos compuestos se predicen con el mismo valor: se han predicho 131 compuestos con un valor de log*BCF* = 0.50. En este sentido, consideramos que la ecuación (4.5) se comporta como un mejor modelo QSPR.

Tabla 4.6. Comparación de la calidad estadística de diferentes modelos QSPR para BCF en el conjunto de datos ANTARES.

Modelo	N	detalles de la partición	R^2_{cal}	RMS_{cal}	R^2_{val}	RMS_{val}	R^2_{pred}	RMS_{pred}
Presente trabajo ecuación (4.5)	851	284:284:283	0.73	0.71	0.60	0.81	0.67	0.71
Modelo-1 9 descriptores ANN[57]	836	608:152:76	0.73	0.67	0.63	0.79	0.62	0.84
Modelo-2 9 descriptores ANN[57]	809	608:132:69	0.73	0.67	0.77	0.62	0.66	0.82
Módulo EPI Suite BCFBAF	851	284:284:283	0.70	0.77	0.64	0.77	0.69	0.70

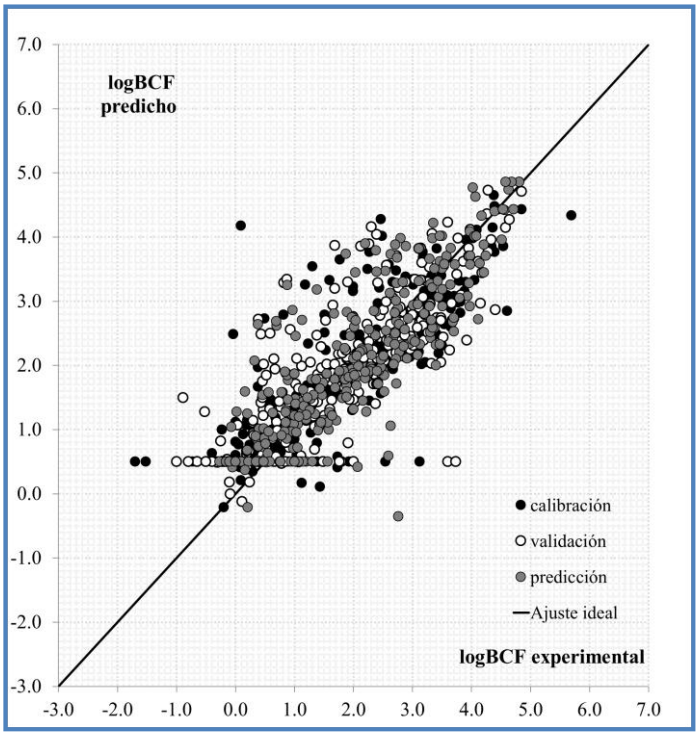


Figura 4.5. Valores de logBCF predichos y experimentales según EPI Suite.

4.3.6. Conclusiones

Se propone un modelo QSPR alternativo para poder estimar el factor de bioconcentración de los compuestos químicos. Para ello, se analizaron simultáneamente un gran número de descriptores moleculares no-conformacionales para encontrar la mejor capacidad predictiva del modelo.

El conjunto de datos ANTARES incluye estructuras moleculares altamente heterogéneas junto con 159 plaguicidas, por lo que el dominio de aplicabilidad de nuestro mejor modelo QSPR considera en su definición diferentes clases químicas para la predicción de BCF, y por lo tanto, podría aplicarse a la predicción de plaguicidas heterogéneos de distintos grupos químicos.

El presente trabajo analiza una gran cantidad de descriptores moleculares (27017 descriptores), para seleccionar los mejores en el modelo de regresión lineal final. De esta manera, enfocamos nuestro trabajo en

describir mejor la estructura química y emplear diferentes tipos de programas de descriptores para mejorar la calidad estadística del modelo establecido.

La consideración de los aspectos constitucionales y topológicos de las estructuras moleculares, en el enfoque QSPR independiente de la conformación utilizado, logra una vez más alcanzar resultados satisfactorios en la predicción de las propiedades/actividades de las sustancias químicas.

4.4. Estudio QSPR de la solubilidad acuosa de pesticidas

Hay un incremento cada vez mayor de detecciones de plaguicidas en el agua, incluidas las fuentes de agua potable. En la actualidad, existen importantes incumplimientos normativos sobre los niveles que pueden exceder la concentración máxima permitida en aguas superficiales, subterráneas y aguas potables tratadas para evitar su contaminación[82,83]. Los pesticidas y sus productos de degradación se distinguen por su fuerte toxicidad y persistencia en el medio ambiente.

La predisposición de un pesticida a ser eliminado del suelo por el agua de escorrentía o de irrigación y para alcanzar una corriente de agua superficial, está directamente relacionada con su solubilidad en agua (S_w). Este parámetro se define como la concentración de un producto químico disuelto en agua cuando el agua está en contacto y en equilibrio con el compuesto químico puro.

La solubilidad en agua establece condiciones de prueba para estudios de destino ambiental (por ejemplo, biodegradación, bioacumulación) y efectos (en humanos y otros organismos vivos), y es usado como un indicador para otros parámetros relevantes ambientalmente, como el coeficiente de partición octanol/agua y el coeficiente de partición carbono orgánico/agua, entre otros.

El error experimental en las mediciones de solubilidad puede ser bastante grande, especialmente para compuestos con un valor de solubilidad muy bajo. La evaluación precisa de la solubilidad en agua se complica por una serie de factores, que incluyen la ionización, la formación de sales y el

polimorfismo. Estos efectos pueden alterar significativamente los valores de solubilidad en agua[84].

La aplicación de la técnica de las relaciones cuantitativas estructura-propiedad y del modelado asistido por computadora son herramientas valiosas e intensamente utilizadas para predecir con exactitud las propiedades físicas y químicas de los compuestos[84]. Los modelos fiables pueden proporcionar información sobre las características moleculares que pueden influir en la solubilidad en agua, mejorando drásticamente la determinación de esta propiedad.

Recientemente, Das y Roy[85] han desarrollado modelos predictivos sobre la solubilidad acuosa de un gran conjunto de fármacos, compuestos similares a fármacos y agroquímicos, con descriptores bidimensionales llamados índices de átomos topoquímicos ampliados (ETA), otros descriptores topológicos, estructurales, espaciales y electrónicos no-ETA y el parámetro de lipofilicidad $C\log P$. Emplearon la función de aproximación genética (AFG), los mínimos cuadrados genéticos parciales (GA/PLS) y el método de regresión lineal múltiple paso a paso (MLR) para generar los modelos. Los descriptores conformacionales independientes son utilizados por el grupo de investigación de Toropov[86] a través del programa CORAL para crear modelos QSPR para la solubilidad en agua.

Además, Zeng *et al.*[87] han aplicado aproximaciones de la teoría del funcional de la densidad (DFT) y métodos QSPR en metil-feniléteres halogenados para modelar su solubilidad en agua. Encuentran que la solubilidad se ve muy afectada por tres variables: la energía del orbital molecular desocupado más bajo, la carga parcial atómica más positiva en la molécula y el momento cuadrupolar.

La solubilidad acuosa de 209 congéneres de cloro-trans-azobenceno ha sido modelada por Wilczyńska-Piliszek *et al.*[88], utilizando una combinación de Algoritmo Genético-Red Neuronal Artificial (GA-ANN). Bhhatarai y Gramática[89] han estudiado los productos químicos perfluorados no-iónicos utilizando descriptores bidimensionales. Además, Benfenati *et al.*[90] han aplicado diferentes modelos computacionales predictivos para analizar la

solubilidad en agua en compuestos orgánicos. Concluyen que, para todos los modelos obtenidos, los valores de los compuestos altamente solubles pueden predecirse mejor que los poco solubles.

El propósito de este estudio es establecer un análisis QSPR para la solubilidad en agua de plaguicidas, utilizando un amplio conjunto de descriptores y de datos experimentales de plaguicidas heterogéneos informados en la literatura. Nuestro objetivo es proponer modelos basados en un conjunto extenso y variado de compuestos, y obtener modelos simples pero confiables, en donde solo se consideran descriptores moleculares no-conformacionales.

Es sabido que los modelos QSPR basados únicamente en características moleculares constitucionales y topológicas, evitan las ambigüedades que pueden resultar de la existencia de compuestos químicos en varios estados conformacionales[58,91].

Por lo tanto, se exploran tres enfoques diferentes de modelos QSPR utilizando: i) descriptores convencionales 0D, 1D y 2D y descriptores indicadores generados por los programas de descriptores de acceso libre PaDEL versión 2.20[30,31], EPI Suite[28] y Mold[2] [65]; ii) descriptores flexibles obtenidos a través del programa CORAL[25,40]; y iii) ambos conjuntos de descriptores combinados.

Se eligen modelos simples que incluyen de 1 a 8 descriptores como las mejores combinaciones de variables predictivas seleccionadas independientemente.

4.4.1. Datos experimentales de PPDB (1211 moléculas)

El análisis QSPR se realiza sobre 1211 plaguicidas aprobados y están disponibles en la Tabla 4.13.A, sus nombres, estructuras y los valores experimentales de solubilidad en agua medida a 20 °C se obtuvieron de la Base de Datos de Propiedades de Pesticidas en línea (PPDB)[92]. La PPDB ha sido desarrollada por la Unidad de Investigación de Agricultura y Medio

Ambiente de la Universidad de Hertfordshire. La solubilidad expresada como g/L se convierte en unidades logarítmicas ($\log S_w$).

4.4.2. Descriptores moleculares

Las estructuras moleculares de los pesticidas se generan en notación SMILES y las estructuras bidimensionales se extraen con el programa Discovery Studio versión 3.5[93], y se guardan en formato MDL mol (V2000) sin realizar ninguna optimización de la geometría.

Se aplican dos enfoques diferentes para calcular los descriptores:

a) El programa de acceso libre PaDEL versión 2.20[30,31], EPI Suite[28] y Mold[2] [65] se utilizan para obtener 17974 descriptores moleculares independientes de la conformación y descriptores indicadores. De este total, 1444 descripciones 0D-2D y 12 tipos de descriptores indicadores (16092) se calculan con PaDEL, 184 descriptores con EPI Suite y 254 descriptores con Mold[2]. Los descriptores que se encuentran dependientes linealmente y los valores constantes se excluyen del grupo de variables.

b) Los descriptores moleculares flexibles se obtienen del programa gratuito CORAL[25,40] utilizando las notaciones SMILES de los compuestos como entrada junto con los valores $\log S_w$ experimentales. En la Tabla 4.20.A se encuentra el valor de *DCW* calculado para los conjuntos de calibración, validación y predicción. En este estudio, *T* varía de 0 a 5 y el número máximo de iteraciones utilizadas es 50.

4.4.3. Validación del modelo

Para estar seguros de que el conjunto de calibración es representativo de los conjuntos de validación y predicción, el conjunto de datos se particiona con BSM[94]. El Método del Reemplazo (RM)[32] programado en el lenguaje Matlab[39] se aplica para generar modelos de Regresión Lineal Múltiple (MLR) en el conjunto de calibración.

Con el fin de medir la estabilidad del modelo QSPR tras la inclusión/exclusión de moléculas, los modelos MLR se validan teóricamente a través del método de validación cruzada dejar-uno-afuera (loo). Este es un criterio de validación general para probar el modelo, siempre que la varianza de loo (R^2_{loo}) sea mayor que 0.5. Sin embargo, esta es una condición necesaria pero no suficiente para determinar el poder predictivo[27]. Un criterio de validación más robusto es aplicar los mismos principios ($R^2_{pred} > 0.5$) al conjunto de predicción externo.

Para descartar correlaciones fortuitas, los valores experimentales se mezclan mediante el método de aleatorización-Y[41] para que no se correspondan con los compuestos respectivos.

El DA para los modelos propuestos en este estudio se determinan mediante el enfoque de influencia[43,95].

4.4.4. Resultados y discusión

Realizamos un estudio QSPR en 1211 compuestos diversos y bien conocidos por su acción pesticida. Se exploran tres enfoques QSPR para modelar la solubilidad con diferente tipo de descriptores: 1) descriptores convencionales; 2) descriptores flexibles; y 3) descriptores híbridos.

La metodología general aplicada en los tres enfoques es primero verificar la capacidad predictiva de los descriptores moleculares y luego evaluar los modelos para los datos experimentales de $\log S_w$ en el conjunto de predicción. Esto permite aprovechar al máximo la información estructural, la respuesta disponible, y así ampliar el DA del modelo diseñado.

4.4.4.1. Descriptores convencionales

Los resultados para los mejores 8 modelos encontrados usando este primer enfoque se muestran en la Tabla 4.7. Se exploran modelos que involucran 1-8 descriptores moleculares junto con descriptores indicadores; los mejores resultados predictivos se observan para 6 y 7 descriptores.

Ambos modelos presentan RMS_{val} similar, pero el modelo con siete descriptores tiene una diferencia menor entre RMS_{val} y RMS_{cal}. Por lo tanto, el modelo de siete descriptores se selecciona como el mejor resultado para el enfoque de descriptores convencionales, y los valores calculados de $\log S_w$ de la ecuación (4.6) frente a los valores experimentales para este modelo, se muestran en la Figura 4.8.

$$\log S_w = 1.46\ GATS2m + 1.81\ GATS1p - 0.35\ CrippenLogP + 4.25\ SIC3 + \\ + 9.90.10^{-11}\ SpDiamD - 1.34\ VAdjMat + 2.25\ MACCSFP35 + 0.16 \quad (4.6)$$

$N_{cal} = 404$, $R^2_{cal} = 0.56$, $RMS_{cal} = 1.57$, $R^2_{loo} = 0.53$, $RMS_{loo} = 1.62$

$N_{val} = 404$, $R^2_{val} = 0.54$, $RMS_{val} = 1.49$, $o3 = 7$, $RMS^{aleat} = 2.30$, $h^* = 0.059$

$N_{pred} = 403$, $R^2_{pred} = 0.56$, $RMS_{pred} = 1.38$

Según las predicciones de la ecuación (4.6), 10 compuestos en este modelo no están dentro del dominio de aplicabilidad y siete compuestos son salientes[96] (**53**, **213**, **233**, **637**, **853**, **1120**, **1155**). Los datos de solubilidad en agua para el compuesto **53** no son confiables, hay al menos dos valores muy diferentes informados en la literatura sobre su solubilidad[92,97].

El resto de los salientes son compuestos con una solubilidad muy baja o muy alta. Por ejemplo, el compuesto **853** muestra la solubilidad más baja en el conjunto de datos, mientras que el compuesto **637** tiene el valor más alto. La solubilidad de los compuestos salientes **213** y **1155** es significativamente baja y los compuestos **233** y **1120** son muy solubles en agua ($S_w \approx 10^3$ g/l). Es comprensible que, en un conjunto molecular tan grande y diverso, los compuestos con valores extremos de solubilidad sean encontrados como salientes en el modelo propuesto. No obstante, la ecuación (4.6) cumple las siguientes condiciones de validación externa[98]:

$1 - R^2_0 / R^2_{pred} < 0.1$ (0.0005) o $1 - R'^2_0 / R^2_{pred} < 0.1$ (0.11); $0.85 \le k \le 1.15$ (0.99) o $0.85 \le k' \le 1.15$ (0.66); $R^2_m > 0.5$ (0.55)

Dos descriptores GATS muestran una correlación positiva con $\log S_w$ en este modelo. Estos son descriptores de autocorrelación 2D originados en la autocorrelación de la estructura topológica de Geary que codifica la estructura molecular como una propiedad fisicoquímica en un vector, relacionando la topología de una estructura con el atributo seleccionado.

El número que sigue al símbolo del descriptor representa la distancia topológica entre pares de átomos (lag) y el conteo de letras para la propiedad fisicoquímica considerada en el componente de ponderación. En el presente modelo, el descriptor *GATS2m* representa un descriptor de autocorrelación de distancia 2 ponderado por la masa, mientras que *GATS1p* describe las polarizabilidades atómicas a una distancia topológica de uno.

El descriptor indicador *MACCSFP35*, que representa la presencia de un átomo de metal alcalino del grupo IA, el descriptor 2D basado en la matriz SpDiamD (diámetro espectral de la matriz de distancia topológica) y el índice de contenido de información estructural *SIC3* (simetría de vecinos de orden 3), también presentan una correlación positiva con $\log S_w$.

Por otro lado, dos descriptores en este modelo presentan un efecto negativo con la solubilidad en agua: *CrippenLogP* y *VAdjMat*. El primero es un descriptor basado en átomos que mide el carácter lipófobo de una molécula, y el segundo es la información de vértices adyacentes de magnitud 1+log2.m, donde m es el número de enlaces metal-metal.

Por lo tanto, este modelo predice que la polarizabilidad, la presencia de átomos alcalinos y la asimetría de la estructura molecular tienen contribuciones positivas a la solubilidad en agua, mientras que el carácter lipófobo y la presencia de metales tienen una contribución negativa, como se esperaba. Los detalles de los descriptores se encuentran en la Tabla 4.14.A.

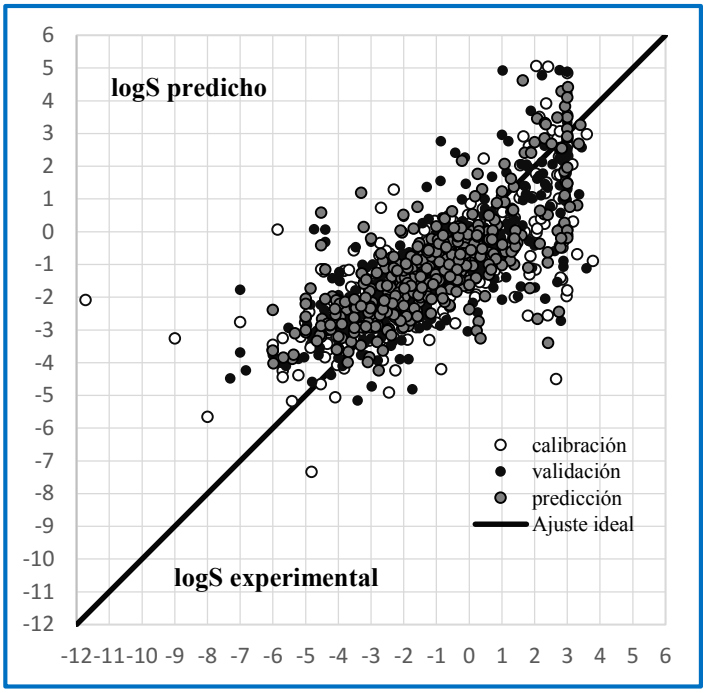


Figura 4.8. Valores experimentales y predichos para los conjuntos de calibración, validación y predicción para la ecuación (4.6).

Tabla 4.7. Descriptores identificados para modelar la solubilidad acuosa. El mejor modelo aparece en negrita.

d	descriptores	R^2_{cal}	RMS_{cal}	R^2_{val}	RMS_{val}	R^2_{pred}	RMS_{pred}
1	*CrippenLogP*	0.27	2.02	0.49	1.70	0.46	1.69
2	*CrippenLogP, piPC8*	0.31	1.96	0.50	1.60	0.43	1.62
3	*GATS1i, CrippenLogP, SpMAD_Dt*	0.38	1.86	0.50	1.57	0.50	1.51
4	*GATS1i, CrippenLogP, ZMIC0, SpDiam_D*	0.42	1.80	0.54	1.52	0.54	1.45
5	*ATSC1e, GATS1i, CrippenLogP, SpDiam_D, SubFPC297*	0.48	1.70	0.55	1.49	0.55	1.42
6	*ATSC1e, GATS2m, GATS1i, CrippenLogP, SpAD_D, SubFPC297*	0.52	1.64	0.53	1.50	0.55	1.39
7	***GATS2m, GATS1p, CrippenLogP, SIC3, SpDiam_D, VAdjMat, MACCSFP35***	**0.56**	**1.57**	**0.54**	**1.49**	**0.56**	**1.38**
8	*AATS3e, AATS2p, AATSC1e, GATS1p, CrippenLogP, SpDiam_D, VAdjMat, PubchemFP406*	0.59	1.52	0.51	1.55	0.55	1.40

4.4.4.2. Descriptor flexible

Para encontrar los atributos estructurales más eficientes para cada RS durante el diseño del descriptor flexible, el descriptor *DCW* se optimiza aumentando R^2 en el conjunto de calibración, hasta que el modelo pierde

capacidad predictiva en el conjunto de validación, sin involucrar el conjunto de predicción. Los parámetros estadísticos para los mejores modelos QSPR encontrados probando diferentes combinaciones basadas en CORAL se presentan en la Tabla 4.8.

El análisis de estos resultados revela que la mejor opción es un enfoque que incluye representaciones de HFG. El descriptor óptimo involucra tres tipos de variables, y 168 atributos activos se basan en ellas. La Figura 4.9 muestra que los valores predichos y experimentales para los conjuntos de calibración, validación y predicción siguen una línea recta. La ecuación resultante para este modelo con un descriptor *DCW* es:

$$\log S_w = 0.10\ DCW + 0.44 \qquad (4.7)$$

$N_{cal} = 404$, $R^2_{cal} = 0.70$, $RMS_{cal} = 1.30$, $R^2_{loo} = 0.69$, $RMS_{loo} = 1.31$

$N_{val} = 404$, $R^2_{val} = 0.60$, $RMS_{val} = 1.40$, $o3 = 5$

$N_{pred} = 403$, $R^2_{pred} = 0.54$, $RMS_{pred} = 1.41$, $RMS^{aleat} = 2.33$, $h^* = 0.01$

Tabla 4.8. Búsqueda del mejor modelo QSPR utilizando descriptores moleculares flexibles.

atributo estructural	R^2_{cal}	RMS_{cal}	R^2_{val}	RMS_{val}	R^2_{pred}	RMS_{pred}
^1S_k	0.45	1.76	0.44	1.65	0.46	1.53
^2S_k	0.75	1.19	0.54	1.55	0.45	1.56
^0EC_j	0.40	1.83	0.45	1.62	0.47	1.53
$Pt2_k$	0.46	1.74	0.50	1.56	0.46	1.54
NNC_j	0.52	1.64	0.50	1.56	0.49	1.49
VS2	0.59	1.52	0.45	1.7	0.48	1.51
^2S_k, NNC_j	0.69	1.31	0.54	1.52	0.50	1.49
^2S_k, $Pt2_k$	0.74	1.21	0.54	1.52	0.48	1.51
^2S_k, $Pt2_k$, NNC_j	**0.70**	**1.30**	**0.55**	**1.51**	**0.53**	**1.43**

Todos los compuestos están dentro del DA y el error sistemático está ausente. Cinco compuestos en el conjunto de calibración (**53**, **352**, **764**, **853**,

1120) muestran residuos absolutos mayores a 3 veces el valor S_{cal} y se consideran como valores salientes.

Aplicamos la aleatorización-Y para demostrar que $RMS_{cal} < RMS^{aleat}$ y también el criterio de validación externa[98] para asegurar que se logre una relación válida estructura-actividad:

$1-R_0^2/R_{pred}^2 < 0.1$ (0.000) o $1-R_0^{'2}/R_{pred}^2 < 0.1$ (0.12); $0.85 \le k \le 1.15$ (0.90)

o $0.85 \le k^{'} \le 1.15$ (0.75); $R_m^2 > 0.5$ (0.59)

La Tabla 4.21.A. incluye un ejemplo para el cálculo de *DCW* del compuesto **2**, al igual que los atributos estructurales que contribuyen al valor del descriptor flexible.

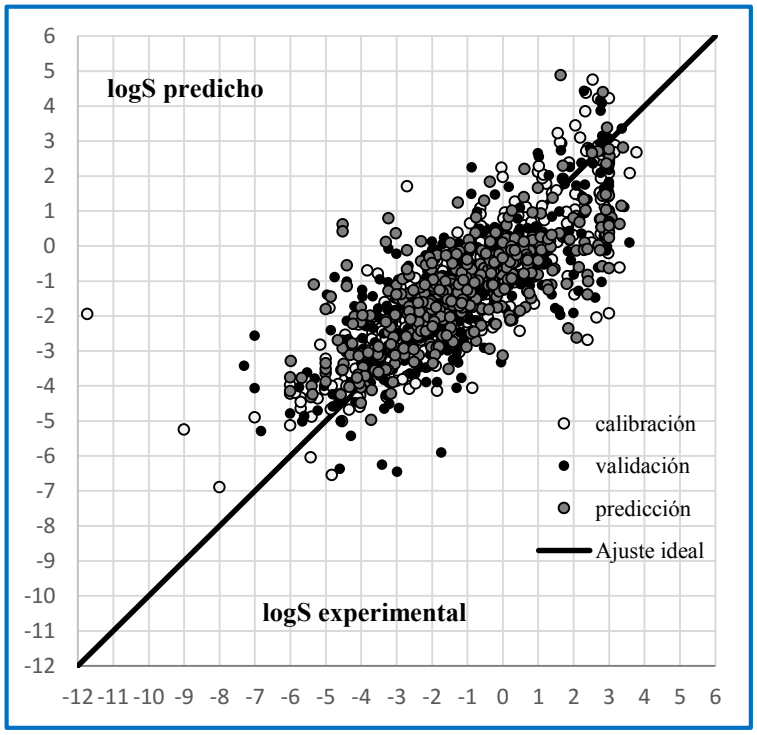


Figura 4.9. Valores experimentales y predichos para la ecuación (4.7).

4.4.4.3. Modelos híbridos

El tercer enfoque que se explora combina PaDEL, EPI Suite, Mold[2], descriptores indicadores y descripciones flexibles de CORAL. La combinación entre varios descriptores flexibles o entre descriptores flexibles y descriptores

moleculares convencionales produce modelos robustos con una mejor capacidad de predicción.

El mejor modelo híbrido implica 3 descriptores (Tabla 4.9, modelo 3), siendo el descriptor llamado *DCW* el mejor descriptor encontrado en el modelo anterior (modelo de descriptores moleculares flexibles, ecuación (4.7)).

El modelo que contiene cuatro variables en el enfoque híbrido resulta más complejo y no conlleva a una mejora significativa en el rendimiento. Los resultados obtenidos para todos los modelos propuestos con los diferentes enfoques se pueden observar en las Tablas 4.17.A-4.19.A.

Se puede observar a partir de los datos que el compuesto **853** es una molécula saliente en todos los modelos propuestos, presentando un valor extremadamente bajo (-11,73) que significa una solubilidad acuosa muy baja. Excluir este compuesto del conjunto de calibración produce un mejor modelo (Tabla 4.9, modelo 3a) representado por la ecuación (4.8). Se puede ver en la Tabla 4.9 y la Figura 4.10 que dicho modelo posee el mejor poder predictivo entre todos los modelos explorados:

$$\log S_w = 2.03\ SIC2 - 0.47\ MACCSFP106 + 0.11\ DCW - 0.67 \qquad (4.8)$$

$N_{cal} = 404$, $R^2_{cal} = 0.75$, $RMS_{cal} = 1.15$, $R^2_{loo} = 0.75$, $RMS_{loo} = 1.16$

$o3 = 4$, $RMS^{aleat} = 2.26$, $h^* = 0.030$

$N_{val} = 404$, $R^2_{val} = 0.62$, $RMS_{val} = 1.38$

$N_{pred} = 403$, $R^2_{pred} = 0.56$, $RMS_{pred} = 1.37$

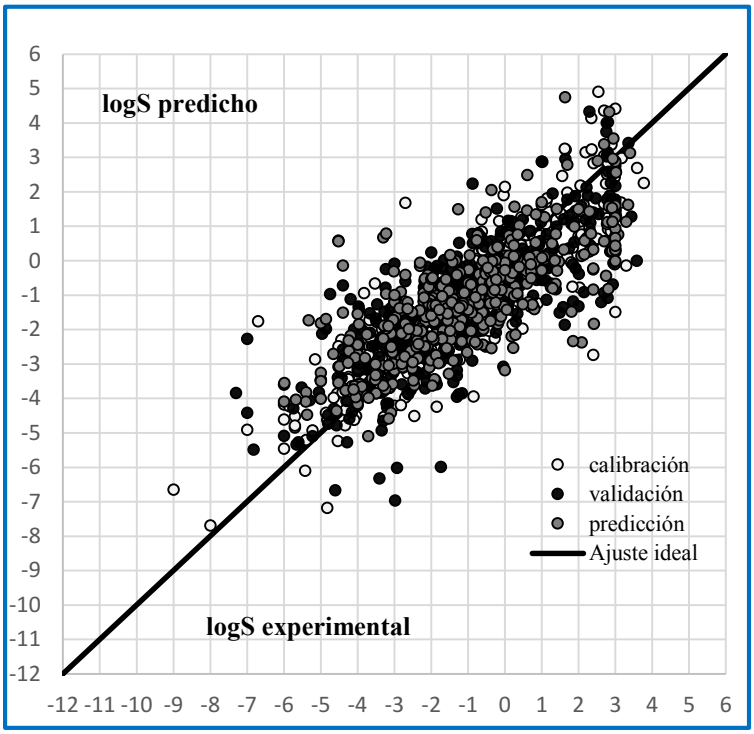


Figura 4.10. Valores experimentales y predichos por la ecuación (4.8) sin incluir el compuesto ***853****.*

Cuatro compuestos en este modelo son salientes (**53**, **352**, **764** y **1120**). Los descriptores *DCW* y *SIC2* (que denota la simetría del vecino de orden 2) presentan una correlación positiva con la propiedad. El descriptor indicador *MACCSFP106* tiene un coeficiente negativo en este modelo. La ecuación (4.8) también cumple las condiciones de validación externa[98]:

$1-R_0^2/R_{pred}^2<0.1$ (0.012) o $1-R_0'^2/R_{pred}^2<0.1$ (0.16); $0.85\le k\le 1.15$ (0.97) o $0.85\le k'\le 1.15$ (0.68); $R_m^2>0.5$ (0.55)

Tabla 4.9. Descriptores identificados para modelar la solubilidad acuosa

d	descriptores	R_{cal}^2	RMS_{cal}	R_{val}^2	RMS_{val}	R_{pred}^2	RMS_{pred}
1	*DCW*[a]	0.70	1.30	0.55	1.51	0.53	1.43
2	*DCW*[b], *MACCSFP106*	0.71	1.28	0.61	1.38	0.55	1.40

3	***DCW, SIC2, MACCSFP106***	**0.73**	**1.22**	**0.61**	**1.38**	**0.56**	**1.38**
4	*DCW, SIC2, MACCSFP106, SubFP236*	0.75	1.19	0.61	1.38	0.56	1.37
3a[b]	***DCW, SIC2, MACCSFP106***	**0.75**	**1.15**	**0.62**	**1.38**	**0.56**	**1.37**

DCW[a] se refiere al descriptor obtenido utilizando CORAL en la representación de HFG para los atributos $Pt2_k$, NNC_j y $2S^k$. El modelo 3a[b] corresponde al modelo 3, con el compuesto **853** eliminado del conjunto de calibración.

No se observan fuertes correlaciones entre los descriptores empleados para construir los modelos de las ecuaciones (4.6) y (4.8). En las Tablas 4.15.A y 4.16.A se describen las matrices de correlación para tales ecuaciones.

4.4.5. Conclusiones

Desarrollamos un modelo simple que predice con éxito la solubilidad en agua de un conjunto grande y diverso de pesticidas a través de una estrategia que no requiere el conocimiento de la conformación molecular como parte de la representación estructural.

El enfoque híbrido que involucra un descriptor molecular, un descriptor indicador y un descriptor flexible calculado con el programa CORAL muestra la mejor capacidad predictiva. Este modelo se valida mediante la aleatorización-Y, validación cruzada e incluye un dominio de aplicabilidad definido de manera adecuada.

4.5. Estudio QSPR de la solubilidad acuosa en compuestos heterogéneos incluidos pesticidas

La solubilidad en agua es una propiedad fundamental, específica de un compuesto químico, definida como la concentración de una sustancia química

disuelta en agua, cuando el agua está en contacto y en equilibrio con la sustancia química pura. Como regla general, un ingrediente muy soluble (solubilidad en agua por encima de varios g/l) no se puede extraer del agua con los procedimientos de extracción disponibles. Los muy insolubles (solubilidad en agua inferior a 0.5-1 mg/l) son difíciles de analizar a niveles traza porque tienen una tendencia a adsorberse en todas partes, especialmente en los materiales de vidrio; lo que conduce a bajas recuperaciones de extracción, a menos que se agregue algún solvente orgánico a las muestras antes de la extracción.

La solubilidad en agua indica la tendencia de un pesticida a ser eliminado del suelo por el agua de escorrentía o de irrigación y de poder alcanzar el agua superficial. También indica la tendencia a precipitar de la superficie del suelo. Sin embargo, este parámetro no puede usarse sólo, para predecir la lixiviación a través del suelo, aunque la distribución de los pesticidas en el ambiente está condicionada por una variedad de coeficientes de partición en agua, y varios autores han mostrado correlaciones entre estos coeficientes de partición y la solubilidad en agua[99].

La solubilidad en agua de los plaguicidas puede ser función de varios parámetros además de la estructura química de los compuestos, tales como: el pH, la temperatura, la concentración de sales, y la concentración de materia orgánica en el agua. La solubilidad también puede variar ampliamente dentro de una clase dada de pesticidas[100].

Además, la solubilidad en agua de un compuesto químico proporciona información considerable sobre el destino y transporte de una sustancia química en el medio ambiente. Los productos químicos altamente solubles en agua tienden a permanecer disueltos en la columna de agua, no se disuelven en el suelo o sedimentos, ni se bioconcentran en organismos acuáticos, generalmente tienen menos probabilidad de volatilizarse del agua, y es más probable que se biodegraden.

Los productos químicos poco solubles en agua son todo lo contrario: se disuelven en el suelo o los sedimentos y se bioconcentran en organismos acuáticos, se volatilizan más fácilmente del agua y es menos probable que se biodegraden.

Otros procesos de destino de los compuestos químicos que pueden ser afectados por la solubilidad en agua, incluyen la fotólisis, la hidrólisis, la oxidación y el lavado a la atmósfera por la lluvia o la niebla[101].

La solubilidad en agua es empleada como base para otros parámetros ambientales, como el coeficiente de partición octanol/agua (*Kow*), coeficiente de partición carbono orgánico-agua (*Koc*)[102] y la constante de la ley de Henry. Esto es un desencadenante regulatorio para ciertos puntos finales del tipo fisicoquímico y ecotoxicológico.

Para que un soluto orgánico se disuelva en agua, primero las moléculas de soluto deben estar separadas unas de otras. En segundo lugar, las moléculas de disolvente deben separarse lo suficiente para crear una cavidad lo suficientemente grande como para acomodar el soluto. Una vez que el soluto ocupa la cavidad, habrá nuevas fuerzas atractivas entre el soluto y el solvente. Finalmente, las moléculas de agua en la capa de solvatación formarán enlaces de hidrógeno adicionales a las moléculas de agua vecinas.

Por lo tanto, la solubilidad en agua depende no solo de la afinidad de un soluto por el agua, sino también de su afinidad por su propia estructura. Las moléculas fuertemente unidas requieren una energía considerable para separarlas. Tales compuestos tienen puntos de fusión altos (para sólidos) y, en general, los sólidos con una temperatura de fusión alta tienen una solubilidad pobre en cualquier disolvente[102].

La eliminación de una molécula de su red cristalina significa un aumento en la entropía, y esto puede ser difícil de modelar con precisión. Por esta razón, y el hecho de que el error experimental en las mediciones de solubilidad puede ser bastante alto para compuestos con muy baja solubilidad, la predicción de la solubilidad en agua no es tan precisa como para otras propiedades, como el coeficiente de partición octanol/agua[103].

Dedek[104] realizó un estudio sobre los factores de la solubilidad que afectan la penetración de pesticidas a través de la piel y la ropa de protección. La penetración de pesticidas sin solventes depende de la solubilidad en agua en lugar de la solubilidad del aceite. La penetración (y por lo tanto también la distribución en el cuerpo) se ve aumentada con una mejor solubilidad en agua por parte del pesticida. Con respecto a la concentración como un factor que

afecta la penetración, la solubilidad del plaguicida en ciertos solventes es de importancia. Las tasas máximas de penetración se encuentran en soluciones saturadas, independientemente de la concentración absoluta.

La concentración debe considerarse en relación con la solubilidad en el solvente, de acuerdo con la ley de partición de Nernst. En las capas de polímeros orgánicos puros, la solubilidad de los compuestos y, por lo tanto, la tasa de penetración, debe ser proporcional a la solubilidad en el material orgánico específico e inversamente proporcional a los valores de la solubilidad del compuesto. Se encontró que algunos compuestos organofosforados (por ejemplo, metilparation) presenta una penetración excepcional al ser muy polar. Los materiales probados con las mejores propiedades de protección fueron los vulcanizados de caucho de butilo[104].

La evaluación de la solubilidad en agua se complica por una serie de consideraciones, incluida la ionización y la formación de sales.

Estos efectos pueden alterar significativamente la solubilidad en agua[105]. Además, la solubilidad puede variar considerablemente con la temperatura, por lo que los datos de solubilidad deben informarse a una temperatura determinada. Las solubilidades en agua pueden informarse de varias maneras: en agua pura, a un pH específico, a una fuerza iónica específica, como especie no disociada (solubilidad intrínseca), y así sucesivamente.

Según la legislación europea sobre productos químicos, REACH (Reglamento CE Nº 1907/2006), existen disposiciones para el uso de datos generados por los métodos cuantitativos de relación estructura-actividad (QSAR). Los modelos QSAR buscan correlaciones matemáticas entre la estructura química y la actividad biológica[106–108].

David S. Palmer *et al.*[109] desarrollaron un modelo QSPR para estudiar la solubilidad en agua para un conjunto de 988 moléculas orgánicas. Para la selección del mejor modelo, se aplicó el método de regresión de Bosques Aleatorios (RF), con mejores resultados que los obtenidos por otros métodos como Mínimos Cuadrados Parciales (PLS), Máquinas de Soporte Vectorial y Red Neuronal Artificial (ANN).

En otro estudio, John S. Delaney[110] propuso un método de estimación de la solubilidad en agua a partir de la estructura química en diferente tipo de compuestos incluidos pesticidas. El modelo se derivó de un conjunto de 2874 solubilidades medidas usando el método de regresión lineal a partir de nueve descriptores moleculares. El parámetro más significativo fue el valor de log*P* calculado, seguido del peso molecular, la proporción de átomos pesados en los anillos aromáticos y el número de enlaces que pueden rotar.

El modelo se comportó consistentemente bien en tres conjuntos de validación, prediciendo solubilidades dentro de un factor de 5-8 con respecto a sus valores medidos, y fue comparado con la "ecuación de solubilidad general" bien establecida para moléculas de tamaño medicinal/agroquímico.

Un estudio presentado por Bhhatarai[111] se centró en productos químicos perfluorados no-iónicos. Su modelo utiliza descriptores bidimensionales: *T (F..F)*, que representa la suma de distancias entre el par de átomos de flúor (dicho valor aumenta con el número y la distancia entre dos átomos de flúor en una molécula) y *SIC1* (que representa el contenido de información estructural) proporcionan información principalmente sobre la simetría estructural en la molécula.

En el trabajo presentado por Raevsky *et al.*[112] se construyeron 32 modelos cuantitativos de relación estructura-propiedad (QSPR) para la predicción de la solubilidad acuosa intrínseca de productos químicos líquidos y cristalinos. Los conjuntos de datos contenían 1022 compuestos líquidos y 2615 compuestos cristalinos. Se usaron para construir modelos locales los métodos MLR, SVM, RF, el método del k-vecino más cercano (kNN), Propiedad Media Aritmética (AMP) y Propiedad de Regresión Local (LoReP).

Se obtuvieron los mejores modelos QSPR: para productos químicos líquidos con *RMS* de predicción en el rango de 0.50-0.60 unidades logarítmicas; y para productos químicos cristalinos con *RMS* entre 0.80-0.90 unidades logarítmicas. En el caso de los modelos globales, la gran cantidad de descriptores dificulta la interpretación mecanicista. Los modelos locales usan solo uno o dos descriptores, de modo que la solubilidad de un compuesto químico medicinal que trabaje con conjuntos de sustancias químicas relacionadas estructuralmente puede estimarse fácilmente.

Sin embargo, la construcción de modelos locales estables requiere la presencia de vecinos estrechamente relacionados para cada sustancia química considerada. Es probable que un consenso de los modelos QSPR globales y locales sea el enfoque óptimo para la construcción de modelos QSPR predictivos estables con interpretación mecanicista.

Un modelo QSPR se construyó para obtener la predicción de la solubilidad acuosa de plaguicidas pertenecientes a 4 clases químicas[113]: ácido, urea, triazina y carbamato. El conjunto completo de 77 plaguicidas se dividió en un conjunto de calibración de 58 plaguicidas y un conjunto de predicción de 19 plaguicidas. Se desarrolló un modelo de 6 descriptores, con un coeficiente de correlación cuadrático (R^2) de 0.89 y un error estándar de estimación de 0.52 unidades logarítmicas, aplicando MLRA y mediante la selección de subconjuntos por GA.

Un modelo QSPR propuesto por Kyrylo Klimenko *et al.*[114] estima el valor de la solubilidad acuosa a diferentes temperaturas a partir de dos pasos, en el primer paso estima el valor del parámetro k incluido en la ecuación lineal $\log S_w = kT + c$, donde S_w es la solubilidad y T la temperatura. El segundo paso usa la técnica RF para crear un modelo QSPR de alta eficiencia. El rendimiento del modelo se evalúa mediante validación cruzada y mediante conjuntos de predicción externos. La capacidad predictiva del modelo desarrollado se compara con la aproximación COSMO-RS, que tiene una base quimicocuántica y termodinámica. La comparación muestra una capacidad de predicción ligeramente mejor para el modelo QSPR presentado en esta publicación.

La solubilidad acuosa a pH = 7.4 es una propiedad muy importante para los compuestos químicos medicinales porque este es el valor de pH de los medios fisiológicos. En otro trabajo propuesto por Raevsky *et al.*[115], se describe la aplicación de tres métodos diferentes tales como: SVM, RF, MLR y tres modelos de relación estructura-propiedad cuantitativa local (regresión corregida por los vecinos más cercanos, RCNN), propiedad media aritmética (AMP) y propiedad de regresión local (LoReP) para construir QSPR estables con una clara interpretación mecanicista.

El conjunto de datos contenía valores experimentales de solubilidad acuosa a pH=7.4 de 387 sustancias químicas (349 en el conjunto de calibración y 38 en el conjunto de predicción, incluidas 16 mediciones propias). El conjunto de descriptores iniciales contenía 210 descriptores fisicoquímicos, calculados a partir de los programas HYBOT, Dragon, SYBYL y VolSurf+.

Se obtuvieron 6 modelos QSPR con buena estadística y basados en los fundamentos de la solubilidad acuosa y la optimización del espacio descriptivo. Esos modelos tienen un *RMS* cercano al error experimental (0.70) y son susceptibles de interpretación física. Los modelos QSPR desarrollados en este estudio pueden ser útiles para compuestos químicos medicinales. Los modelos globales de MLR, RF y SVM pueden ser valiosos para la consideración de los factores comunes que influyen en la solubilidad. Los modelos locales RCNN, AMP y LoReP pueden ser útiles para la optimización de la solubilidad acuosa en pequeños conjuntos de productos químicos relacionados.

Los modelos QSPR resultan apropiados para minimizar el tiempo, el costo y los recursos, como un sustituto de los datos experimentales y como un complemento de los datos experimentales[116].

La capacidad predictiva de los modelos depende en gran medida de los compuestos utilizados en el conjunto de calibración. En ese sentido, la capacidad de evaluar la confianza en el valor predicho es crucial para la correcta interpretación y aplicación de los modelos QSPR.

Se puede usar el concepto del dominio de aplicabilidad (DA)[117], definido como el espacio de la respuesta y la estructura química en los que un modelo realiza predicciones con una fiabilidad determinada[118].

4.5.1. Datos experimentales de WATERNT (5610 moléculas)

La base de datos fue tomada del módulo para estimación de la solubilidad WATERNT del programa EPI Suite[28]. Las solubilidades (S_w, mol/L) fueron tomadas de diferentes fuentes confiables a 25 °C. La base de datos consta de 5610 compuestos heterogéneos y pesticidas, en un intervalo logarítmico de solubilidades medidas de -13.17 a 1.70. En la Tabla 4.23.A se incluyen los detalles de los compuestos estudiados.

4.5.2. Diseño del modelo QSPR para la solubilidad acuosa

Los descriptores moleculares se calcularon mediante los programas PaDEL (14464), Mold2 (777), DataWarrior (34), QuBiLs-MAS (8448) y CORAL (1), obteniendo así 23724 descriptores.

Mediante la técnica BSM se realiza la partición molecular en los conjuntos de calibración, validación y predicción de igual dimensión cada uno (1870 moléculas, 33,33% del total). El análisis sobre el conjunto de calibración de la dependencia lineal, exclusión de descriptores de valores únicos, exclusión de descriptores sin datos, conduce a $D = 7339$ variables estructurales independientes.

La Tabla 4.23.A indica los miembros de cada conjunto como validación (^) y predicción (*). Como una etapa siguiente, los descriptores moleculares más representativos son buscados en el conjunto de calibración a través del método de selección de variables RM. Las mejores regresiones lineales QSPR de 1-7 descriptores más representativos para $\log S_w$ involucran a los descriptores de la Tabla 4.10. Dichas combinaciones entre descriptores de naturaleza rígida y flexible son elegidas entre $D = 7339$ variables independientes.

Tabla 4.10. Los mejores descriptores seleccionados para el estudio de la solubilidad acuosa.

d	descriptores	R^2_{cal}	RMS_{cal}	R^2_{val}	RMS_{val}	R^2_{pred}	RMS_{pred}
1	*DCW*	0.82	0.93	0.81	0.98	0.81	0.96
2	*Sub295; DCW*	0.83	0.90	0.82	0.95	0.83	0.93
3	*Sub295; D585; DCW*	0.84	0.88	0.83	0.93	0.84	0.90
4	***XLogP; Sub295; D585; DCW***	**0.84**	**0.87**	**0.83**	**0.91**	**0.84**	0.88
5	*PC408; Sub295; Qub1; D585; DCW*	0.84	0.86	0.83	0.91	0.84	0.88
6	*XLogP; Sub295; Qub2; D586; D733; DCW*	0.85	0.86	0.84	0.90	0.85	0.86
7	*XLogP; Sub295; KR4839; Qub2; D586; D733; DCW*	0.85	0.85	0.84	0.89	0.85	0.85

Con el fin de mantener la dimensión del modelo lo más baja posible, seleccionamos el modelo de 4 descriptores como la mejor regresión lineal QSPR, dado que RMS_{cal} y RMS_{val} no mejoran significativamente si aumenta el número de descriptores:

$$\log S_w = -0.08\,XLogP + 1.01\,Sub295 - 0.23\,D585 + 0.07\,DCW - 0.60 \tag{4.9}$$

$N_{cal} = 1870$, $R^2_{cal} = 0.84$, $RMS_{cal} = 0.87$

$R^2_{ij\,max} = 0.53$, $o3 = 15$, $R^2_{aleat} = 0.02$, $RMS^{aleat} = 2.15$ (100000 casos)

$R^2_{loo} = 0.84$, $RMS_{loo} = 0.88$, $R^2_{l30\%o} = 0.83$, $RMS_{l30\%o} = 0.89$ (100000 casos)

$N_{val} = 1870$, $R^2_{val} = 0.83$, $RMS_{val} = 0.91$

$N_{pred} = 1870$, $R^2_{pred} = 0.84$, $RMS_{pred} = 0.88$

La representación gráfica del modelo encontrado conduce a las Figuras 4.11 y 4.7.A, que demuestran que la regresión multivariable propuesta es válida.

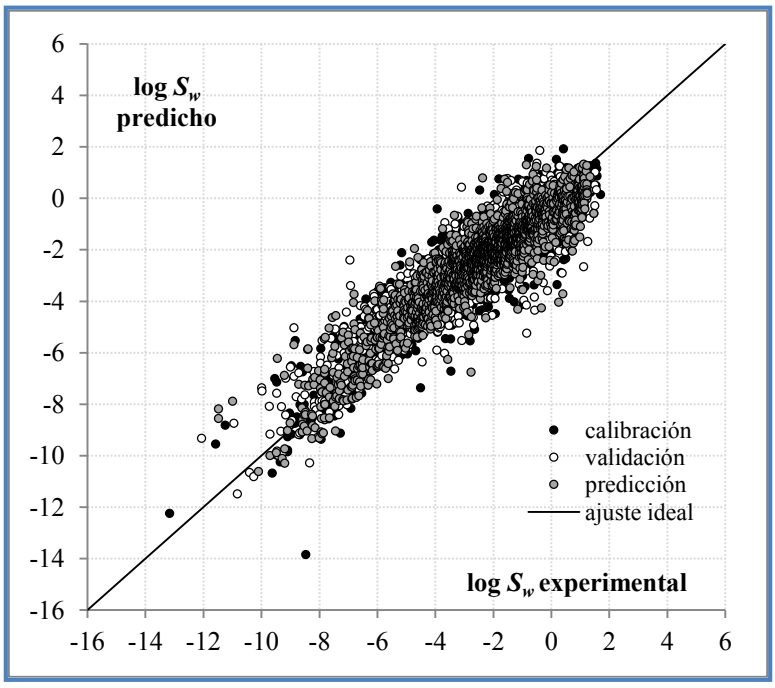


Figura. 4.11. Valores de logS_w experimentales y predichos mediante la ecuación (4.9).

La ecuación (4.9) supera la etapa de validación: se predice bien a las moléculas del conjunto de predicción no contempladas en el ajuste del modelo. La capacidad predictiva de nuestro modelo QSPR sobre las 1870 moléculas de predicción es 'buena' según el criterio MAE: para el conjunto de predicción $MAE(100\%) = 0.67$ y $\sigma(100\%) = 0.57$, mientras que si se omite el 5% de los compuestos con altos valores de residuos conduce a $MAE(95\%) = 0.58$ y $\sigma(95\%) = 0.42$.

Según demuestra la Figura 4.7A, 15 moléculas del conjunto de calibración poseen un residuo alto mayor a $3.S_{cal}$. Este comportamiento puede atribuirse a la gran cantidad de compuestos analizados y a la heterogeneidad estructural: dicloruro de paraquat; clorhexidina; ácido etilenodiaminotetracético; hexaclorofeno; leucopterina; hexatriacontano; óxido de tributilfosfina; glafenina; 4,5-dihidroxi-3,6-bis(fenilazo)-2,7-ácido disulfónico naftaleno; demeton; disperso azul 79; procion azul mx-r; base libre de guazatina; tecloftalam; nitenpiram.

La correlación máxima entre pares de descriptores del modelo de 4 descriptores no es significativa ($R^2_{ij\,max} = 0.53$). La relación cuantitativa estructura-propiedad representada por la ecuación (4.9) cumple con los parámetros loo, lmo(30%) y aleatorización-Y. Además, se cumplen las siguientes condiciones:

$R^2_m > 0.5\ (0.84)$; $1 - R^2_0 / R^2_{pred} < 0.1\ (1.72.10^{-4})$ o $1 - R'^2_0 / R^2_{pred} < 0.1\ (0.024)$;
$0.85 \le k \le 1.15\ (1.006)$ o $0.85 \le k' \le 1.15\ (0.928)$

Los descriptores no-conformacionales incluidos en el modelo son:

a) un descriptor propiedad: *XLogP*, el logaritmo del coeficiente de partición octanol/agua de Wang

b) un descriptor indicador de subestructura: *Sub295*, presencia de enlaces entre C y los heteroátomos O, N o S

c) un descriptor de Burden: *D585*, el autovalor más alto de la matriz de Burden pesado por electronegatividades de Sanderson de orden 6.

d) un descriptor flexible: *DCW*, basado en los atributos hfg-vs2 y SMILES-sss.

La presencia del coeficiente partición octanol/agua en la ecuación (4.9) y su contribución negativa a la propiedad resulta fácil de entender, pues es una medida de la hidrofobicidad de un compuesto. Los demás descriptores presentes en la ecuación lineal permiten establecer un complemento entre las variables de manera de alcanzar una calidad estadística aceptable.

Según el signo de los coeficientes de regresión de la ecuación (4.9), cuanto menores sean simultáneamente los valores numéricos de *XLogP* y *D585*, y mayores sean los valores de *Sub295* y *DCW* en una estructura química considerada, mayor tenderá a ser el valor predicho de su solubilidad acuosa.

En la Figura 4.12 se observa que los compuestos del conjunto de predicción (puntos en gris) que exceden el valor de influencia límite h^* están próximos a compuestos de los conjuntos de calibración y predicción (puntos en negro y blanco, respectivamente), por tanto, sus predicciones pueden considerarse confiables.

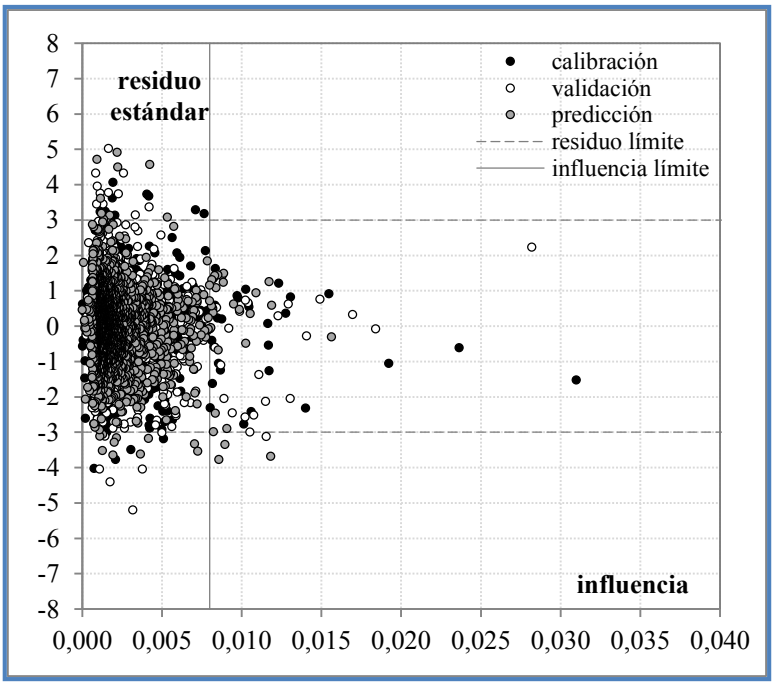


Figura 4.12. Gráfico de Williams para la ecuación (4.9). Influencia límite h = 0.008.*

Posteriormente, la calidad de las predicciones de $\log S_w$ proporcionada por la ecuación (4.9) se compara con la obtenida con el módulo WATERNT de EPI Suite, para lo cual las Figuras 4.13 y 4.8A grafican las predicciones

obtenidas mediante este programa. Los 5610 compuestos estudiados conducen a $RMS = 0.89$ en el caso de la ecuación (4.9), mientras que para WATERNT se obtiene $RMS = 1.01$, revelando que se alcanzó un mejor resultado en nuestro modelo propuesto.

Finalmente, los valores numéricos de los mejores descriptores de la solubilidad acuosa para las 5610 estructuras químicas ensayadas se incluyen en la Tabla 4.24.A.

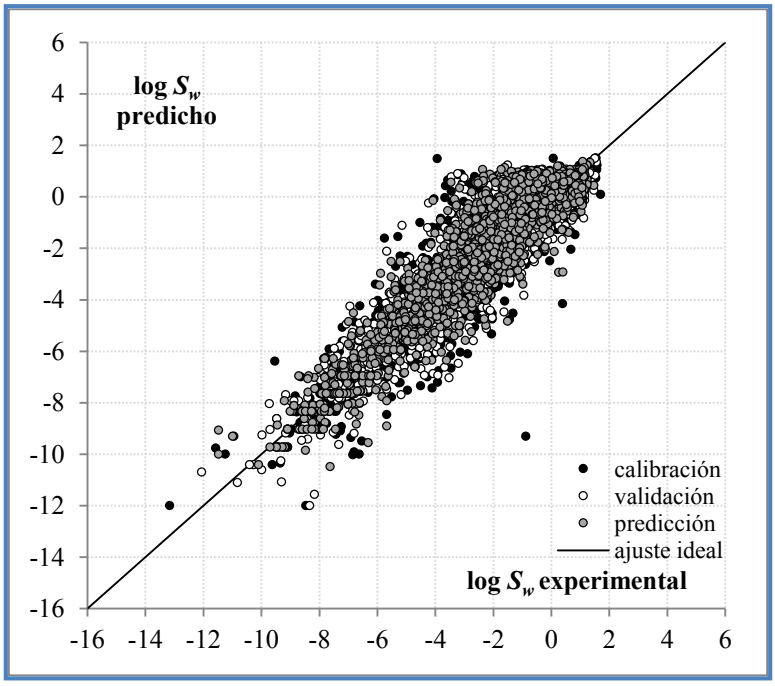


Figura 4.13. Valores de logS_w experimentales y predichos mediante el módulo de EPI Suite.

4.5.3. Conclusiones

La combinación apropiada entre descriptores de naturaleza rígida y flexible permite establecer un modelo QSPR para la solubilidad acuosa de 5610 compuestos químicos de alta diversidad estructural. Los 4 descriptores seleccionados más representativos de la solubilidad mejoran la capacidad predictiva del programa WATERNT de EPI Suite. Se concluye que ambos modelos generan predicciones satisfactorias de la propiedad, a través de metodologías alternativas.

El método de subconjuntos balanceados BSM, técnica de partición molecular desarrollada en el grupo de trabajo, permite establecer modelos matemáticos con un dominio de aplicación que cubre a todas las moléculas analizadas, especialmente a las moléculas del conjunto de predicción.

A través de descriptores independientes de la conformación derivados de la consideración de aspectos constitucionales o topológicos de las estructuras moleculares, es posible establecer modelos predictivos de propiedades fisicoquímicas de interés, en este caso la solubilidad acuosa.

4.6. Estudio QSPR de la constante de la ley de Henry

Los datos característicos de los plaguicidas generalmente se miden de acuerdo con protocolos bien establecidos reconocidos por organismos nacionales e internacionales (directrices de la US EPA, protocolos de la OCDE y de la UE, etc.). La mayoría de los datos fisicoquímicos se miden en el laboratorio en condiciones experimentales bien definidas.

Algunos datos son característicos de la molécula del plaguicida individual, por ejemplo, la solubilidad en agua, presión de vapor, volatilidad, estabilidad en agua, fotodegradación, coeficiente de partición octanol/agua. Bajo condiciones experimentales dadas (temperatura, presión, pH, etc.), los experimentos deben conducir a valores similares.

Otros datos como la vida media en los suelos, o los coeficientes de partición suelo-agua se miden en experimentos de laboratorio y/o de campo y dependen fuertemente de las condiciones experimentales y ambientales, por lo que son solo parcialmente característicos de la molécula pesticida.

Por lo tanto, los valores de los datos que se pueden encontrar en la literatura a veces se extienden en un amplio rango y no siempre es fácil obtener datos “confiables”.

La constante de la ley de Henry (K_H) relaciona las concentraciones en equilibrio líquido y fase vapor de un soluto en soluciones diluidas a presiones moderadas y temperatura constante. En esta ley, la cantidad de un gas dado disuelto en un volumen dado de líquido para formar una mezcla ideal es directamente proporcional a la presión parcial de ese gas en equilibrio con el

líquido. La constante de proporcionalidad se denota como la constante de la ley de Henry (K_H, [L atm mol^{-1}])[119,120] y representa un coeficiente de partición que encuentra aplicaciones notables tanto en Ingeniería Química como en Ingeniería Ambiental.

La tendencia de los pesticidas a volatilizarse de la solución acuosa al aire está determinada en gran parte por sus valores de K_H, un valor alto favorece la volatilización. El coeficiente de reparto aire/agua es importante en los estudios de asociaciones de plaguicidas con la lluvia, el agua de las nubes, agua niebla o en los alvéolos de los pulmones en los seres humanos y de otros animales[121]. Las muestras que contienen dichos compuestos deben manipularse con cuidado para evitar pérdidas, y los pasos de evaporación no deben incluirse en el proceso de muestreo.

Los plaguicidas que tienen grandes valores de K_H pueden analizarse mediante el análisis del espacio de cabeza o mediante la extracción de gases.

Los valores de K_H de los compuestos son indicadores más apropiados de su volatilización, en comparación a si se considera solamente el valor de la presión de vapor, ya que este representa un coeficiente de partición. Incluso, en una primera aproximación, los valores de presión de vapor son útiles para indicar y clasificar compuestos por grupos de volatilidad creciente. Un valor de presión de vapor débil no siempre indica una volatilización despreciable. Como ejemplo, tenemos al DDT que presenta un valor de presión de vapor débil aunque una baja solubilidad en agua, por lo que su volatilización no puede ser totalmente insignificante.

Suntio *et al.* han presentado una recopilación y revisión crítica de la constante de la ley de Henry para muchos plaguicidas[121]. El valor de K_H puede expresarse en forma adimensional o con unidades. En la forma adimensional, las mismas unidades de concentración se usan en las fases de aire y agua. La forma adimensional se puede convertir a la forma dimensional multiplicando por RT, convirtiendo así la concentración de aire en unidades de presión con el uso de la ley de los gases ideales.

En los últimos años, se ha trabajado mucho para medir y tabular K_H. Por ejemplo, Sander recolectó 17350 valores de K_H para 4632 especies,

basado en 689 referencias[122]. La medición experimental de K_H se ha informado utilizando diferentes técnicas, incluida la cromatografía de gases en el espacio de cabeza[123], técnicas de espacio de cabeza modificado[124], variación de la relación de fases[125], método de espacio de cabeza diferencial[126], y técnicas de dilución[127]. Sin embargo, los valores exactos de K_H no están disponibles para muchos compuestos.

Los problemas instrumentales, los límites de detección de bajas concentraciones de compuestos hidrofóbicos, entre otros factores, dificultan y encarecen la determinación experimental de los valores de K_H [128].

Mediante el uso de la ley de Henry, K_H puede calcularse convenientemente como la relación entre la presión de vapor líquida y sólida y la solubilidad. Por lo tanto, K_H se informa a menudo en [Pa m^3 mo1^{-1}].

Los valores de K_H pueden estimarse a partir de las solubilidades determinadas experimentalmente al igual que las presiones de vapor. Los métodos preferidos implican el flujo de aire o agua a través de "columnas generadoras"[129]. Los métodos de cromatografía de gases también se pueden usar para determinar las presiones de vapor[130]. Se demostró que un método de "partición en equilibrio en un sistema cerrado" es adecuado para compuestos que tienen valores altos de K_H (>100 Pa m^3 mo1^{-1})[131].

Otros métodos implican un sistema de flujo en el que la concentración de la sustancia química en agua con una corriente constante de gas se mide como una función del tiempo[132].

Como consecuencia de estas variaciones en los métodos, los valores de K_H informados por diferentes autores muestran amplias discrepancias, como se encuentra para los valores de presión de vapor. Suntio *et al.*, calculó los valores de K_H a partir de la presión de vapor y los valores de solubilidad[121]. En general, se considera que los compuestos con valores de K_H <10-5 Pa m^3/mo1 tienen poca tendencia a volatilizarse.

Los métodos de estimación para K_H con fines ambientales se pueden categorizar como:

(1) métodos de relaciones propiedad-propiedad (PPR);

(2) métodos de contribución de enlaces y grupos;

(3) métodos de solvatación continua;

(4) UNIFAC (coeficiente de actividad del grupo funcional quasi-químico universal) y los métodos de relaciones cuantitativas con descriptores químico cuánticos o relaciones estructura-propiedad (QSPR) basados en descriptores fisicoquímicos[133].

El método PPR más conocido es el método VP/AS (presión de vapor/solubilidad acuosa)[134]. El método VP/AS, también llamado medida indirecta de K_H [134], tiene excelentes resultados, aunque para compuestos con baja solubilidad y baja presión de vapor medidos a una temperatura deseada, puede conducir a grandes errores[135].

El primer método de contribución de enlaces[136] se ha mejorado ampliando el número de definiciones de enlaces de 34 a 59 y con la adición de 15 factores de corrección[137] y finalmente, en revisiones recientes, el método de enlaces contiene 64 definiciones de enlaces y 57 factores de corrección, mientras que el método de contribución de grupos contiene 93 definiciones de grupos[138].

Los modelos de solvatación continua (SMx) se basan en una relación termodinámicamente lineal del logaritmo de K_H y la energía libre de solvatación (DG). Una revisión exhaustiva de los rendimientos de los modelos SMx[139] utilizando un conjunto de datos que incluye 700 valores experimentales de K_H ha revelado que, a pesar del alto costo computacional de estos métodos, se cometieron grandes errores incluso de hasta 30 órdenes de magnitud para poder obtener K_H .

UNIFAC es un modelo semiempírico basado en termodinámica QSPR-PPR y es utilizado para calcular el coeficiente de actividad.

Para aplicaciones ambientales, UNIFAC se usó directamente para calcular el coeficiente de actividad de dilución infinita (c1) para una solución acuosa[140], o indirectamente por extrapolación de datos del equilibrio vapor-líquido obtenidos a concentraciones de solutos más altas[141]. El valor de c1 se usa luego con la presión de vapor y el valor de la relación de presión total (Psat/PT) para calcular K_H .

Aunque el enfoque de UNIFAC puede considerar los efectos de la temperatura, requiere parámetros de interacción que se obtienen del ajuste del modelo a los datos experimentales en fase de equilibrio, que a menudo faltan para los compuestos químicos de interés ambiental.

Por consiguiente, la aplicación de métodos teóricos para la predicción veraz de este importante parámetro para diversos tipos de compuestos es esencial. En la actualidad, los modelos QSPR se han convertido en una herramienta moderna, económica y rápida para predecir las propiedades físicas, biológicas o químicas de los compuestos.[142] De hecho, la oficina Europea de sustancias químicas promueve el uso de modelos QSPR en el marco europeo del Reglamento REACH (CE), Nº 1907/2006.[143]

Tales estudios QSPR han proporcionado modelos satisfactorios para la predicción de la constante de la ley de Henry de conjuntos de datos bastante pequeños y de clases químicas específicas[139]; sin embargo, desarrollar un modelo QSPR integral para una amplia gama de compuestos químicos sigue siendo un desafío para los investigadores.

Un estudio QSPR de Modarresi *et al.*[144] en el sistema aire-agua para 189 hidrocarburos alifáticos encontró una relación lineal entre el logaritmo de K_H y la energía libre de solvatación estándar de Gibbs. Con descriptores moleculares tridimensionales (3D) que incluyen un nuevo descriptor para el factor de forma de la cavidad molecular, el modelo propuesto es valioso para solutos polares y cargados. Este modelo de tres descriptores basados en red neuronal artificial (ANN) conduce a un coeficiente de correlación (*R*) de 0.90 y un error cuadrático medio (*RMS*) de 0.22.

El mismo autor y colaboradores, utilizando un conjunto más grande de 940 compuestos orgánicos, propusieron seis modelos QSPR. En ese estudio, las estructuras moleculares se optimizaron con el método PM3. Informaron que un modelo de red neuronal con función de base radial (RBFN) de diez parámetros presenta el mejor rendimiento con un *R* que varía de 0.88-0.98 en función del conjunto y un *RMS* de 0.564[145]. Utilizando el Método del Reemplazo para 150 hidrocarburos alifáticos, Duchowicz *et al.* diseñaron un modelo QSPR de siete parámetros que llevan a *R*=0.996 y la desviación estándar (*S*) = 0.065[146].

Más tarde, en un estudio QSPR que trabaja con 96 plaguicidas orgánicos, encontraron que el modelo de mejor capacidad predictiva utiliza la función de ponderación de Levenberge-Marquardt con regularización Bayesiana (BR) (*R*=0.74-0.79 y *RMS*=0.93-1.29)[147].

O'Loughlin e English trabajando con varios cientos de compuestos orgánicos en agua, encontraron que los modelos MLR funcionan mejor para conjuntos de compuestos con clases específicas, mientras que los modelos ANN tienen mayor precisión predictiva en conjuntos generales de compuestos[148].

Modelos QSPR anteriores[149–155] han sido desarrollados en base a diferentes puntos tomados en cuenta para ajustar los datos de hasta 495 compuestos[155] y que tienen diferentes grupos funcionales.

Finalmente otros modelos QSPR propuestos en Refs. 133,156,157 aplican diferentes metodologías de trabajo, utilizando descriptores moleculares y métodos de contribución de grupos respectivamente. Los modelos fueron encontrados a su vez mediante la utilización de diferentes técnicas, como son el método de paso a paso, la red neuronal de paso a paso y mediante GA-MLR.

Recientemente, los valores de K_H para seis familias de contaminantes orgánicos persistentes se modelaron utilizando un método de contribución de grupos basado en la teoría de partícula escalada. Los valores de *R* y *RMS* reportados en los conjuntos de calibración y validación son 0.89, 0.22 y 0.88, 0.27, respectivamente[158].

Como es sabido, cada modelo que incluye descriptores tridimensionales generalmente implica altos costos computacionales y largos tiempos durante el cálculo de optimización de la geometría molecular. Por lo tanto, el enfoque QSPR independiente de la conformación[58–61] se puede considerar como una metodología muy útil. Además, también exploramos el rendimiento de los modelos QSPR basados en descriptores óptimos[19].

4.6.1. Datos experimentales de HENRYWIN (530 moléculas)

La base de datos consiste en 530 compuestos químicos heterogéneos e incluye a pesticidas con valores experimentales de la contante de Henry (K_H), tomados de varias bases de datos confiables y disponibles públicamente, recolectados del módulo HENRYWin de EPI Suite[28].

Los diferentes compuestos de la base de datos presentan valores en la escala logarítmica de K_H, en un intervalo de -12.99 a 1.30. La lista completa de compuestos estudiados aquí está presentada en las Tablas 4.25.A y 4.26.A.

4.6.2. Modelos QSPR de la constante de Henry

Los descriptores moleculares se calcularon en primera instancia mediante los programas PaDEL, Mold[2], DataWarrior y QuBiLs-MAS, obteniéndose así 23723 descriptores.

La Tabla 4.25.A indica los miembros de cada conjunto como validación (^) y predicción (*). De esta manera, los compuestos de calibración y validación constituyen el 66.60 % de toda la base de datos completa.

Luego de realizar la partición molecular en los conjuntos de calibración, validación y predicción mediante la técnica BSM, se procede a la búsqueda de los descriptores más representativos para $\log K_H$. Las mejores regresiones lineales QSPR de 1-7 descriptores obtenidas mediante RM, involucran a los descriptores de la Tabla 4.11 elegidos entre $D = 5838$ variables independientes.

Tabla 4.11. Los mejores descriptores seleccionados para el estudio de la constante de Henry.

d	descriptores	R^2_{cal}	RMS_{cal}	R^2_{val}	RMS_{val}	R^2_{pred}	RMS_{pred}
1	*D274*	0.46	1.73	0.34	1.48	0.45	1.28
2	*MCS130; PSA*	0.63	1.43	0.39	1.39	0.37	1.36
3	*LFEA; D439; clogP*	0.77	1.13	0.73	0.92	0.73	0.90
4	*LFEA; LFEE; MCS130; PSA*	0.78	1.11	0.70	0.99	0.73	0.93

5	*GATS1c; LFEA; LFES; ToPSA; PC540*	0.81	1.02	0.76	0.88	0.79	0.79
6	*MDEC-33; LFEA; LFES; MCS48; D151; clogP*	0.84	0.94	0.81	0.78	0.83	0.72
7	***ATSC1e; LFEA; LFEBH; ToPSA; MCS48; SubC1; D604***	**0.87**	**0.85**	**0.85**	**0.70**	**0.81**	**0.76**
8	*ATSC1e; MDEC33; LFEA; LFEBH; R-TpiPCTPC; ToPSA; MCS48; SubC1*	0.88	0.82	0.87	0.65	0.83	0.71

Con el fin de mantener el tamaño del modelo lo más pequeño como sea posible, seleccionamos el modelo de 7 descriptores como la mejor regresión lineal QSPR, dado que RMS_{cal} no mejora significativamente si se utilizan 8 descriptores:

$$\log K_H = -1.76\,ATSC1e - 3.78\,LFEA - 3.42\,LFEBH - 0.04\,ToPSA + 3.86\,MCS48 + \\ +0.47\,SubC1 - 0.19\,D604 - 2.31 \qquad (4.10)$$

$N_{cal} = 177$, $R^2_{cal} = 0.87$, $RMS_{cal} = 0.85$, $R^2_{ij\max} = 0.29$, $o3 = 1$

$R^2_{aleat} = 0.23$, $RMS^{aleat} = 2.07$ (100000 casos), $R^2_{loo} = 0.85$, $RMS_{loo} = 0.92$

$R^2_{l30\%o} = 0.77$, $RMS_{l30\%o} = 1.13$ (100000 casos), $N_{val} = 176$, $R^2_{val} = 0.85$, $RMS_{val} = 0.70$

$N_{pred} = 177$, $R^2_{pred} = 0.81$, $RMS_{pred} = 0.76$

De estos resultados $R^2_{ij\max}$ es el máximo coeficiente de correlación entre pares de descriptores, indicando la ausencia de serias correlaciones entre los siete descriptores seleccionados.

La ecuación (4.10) predice de manera aceptable la propiedad en los conjuntos de calibración y validación, y lo que es más importante, predice bien a las moléculas del conjunto de predicción no contempladas en el ajuste del modelo. La capacidad predictiva de nuestro modelo QSPR sobre las 177 moléculas de predicción es "buena" según el criterio MAE: para el conjunto de predicción $MAE(100\%) = 0.56$ y $\sigma(100\%) = 0.51$, mientras que si se omite el 5% de los compuestos con altos valores de residuos conduce a $MAE(95\%) = 0.48$ y $\sigma(95\%) = 0.39$.

El gráfico que relaciona las predicciones con los valores experimentales se encuentra en la Figura 4.14. El gráfico de la dispersión de los residuos de la Figura 4.9A tiende a obedecer un patrón aleatorio alrededor de la línea cero.

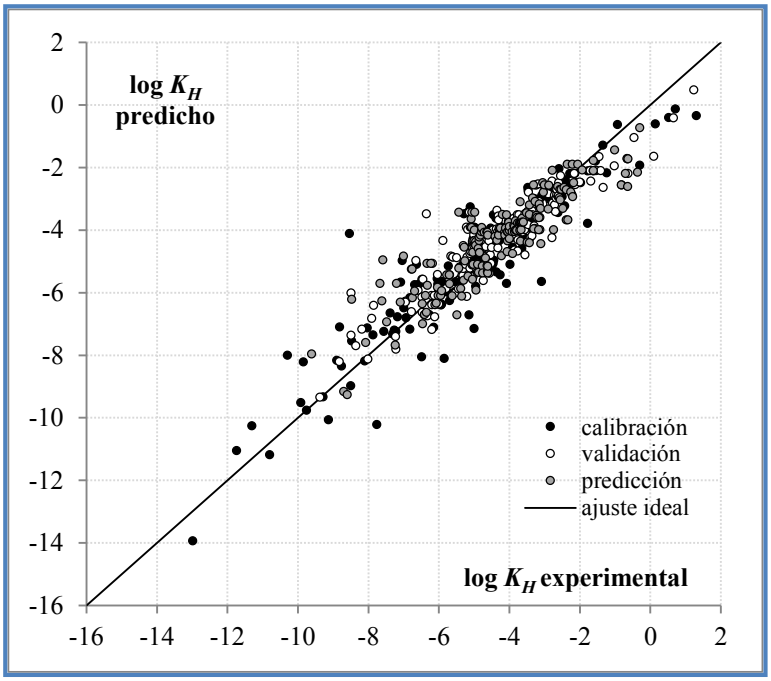


Figura 4.14. Valores de log K_H predichos y experimentales de acuerdo al mejor modelo QSPR de 7-descriptores encontrado.

Según demuestran las Figuras 4.14 y 4.19A, un único compuesto del conjunto de calibración posee un residuo alto mayor a $3.S_{cal}$, tricloroacetaldehído. Este comportamiento puede atribuirse a la heterogeneidad de las estructuras analizadas y que los descriptores elegidos por el modelo no alcancen a predecir acertadamente esta estructura en particular. No obstante, decidimos no remover tal molécula de nuestro estudio.

Un intento de mejorar la calidad estadística de este modelo lineal se hace a partir de la inclusión de un descriptor flexible, que se combine con otros descriptores rígidos seleccionados. Esto no conduce a una mejora en los parámetros de validación. Por tanto, en la base de datos actual no conviene considerar al descriptor flexible pues sobreajusta al conjunto de calibración.

La relación cuantitativa estructura-propiedad representada por la ecuación (4.10) cumple con los parámetros loo, lmo(30%) y aleatorización-Y. Además, se cumplen las siguientes condiciones:

$1-R_0^2/R_{pred}^2 < 0.1\ (0.029)$ o $1-R_0'^2/R_{pred}^2 < 0.1\ (0.087)$; $0.85 \le k \le 1.15\ (1.017)$ o $0.85 \le k' \le 1.15\ (0.957)$; $R_m^2 > 0.5\ (0.77)$

La correlación máxima entre pares de descriptores del modelo no es significativa ($R_{ij\max}^2 = 0.29$). Estos descriptores no-conformacionales se clasifican de la siguiente manera: a) un descriptor de autocorrelación 2D: *ATSC1e*, autocorrelación centrada de Broto-Moreau - distancia 1/ponderada por la electronegatividad de Sanderson; b) dos descriptores de relación lineal de energía libre molecular: *LFEA*, acidez global de enlace hidrógeno del soluto, y *LFEBH*: basicidad global de enlace hidrógeno del soluto; c) un descriptor topológico: *ToPSA*: área superficial polar topológica; d) tres descriptores indicadores: *MCS48*, número de grupos OQ(O)O, donde Q es un heteroátomo diferente a carbono o hidrógeno; *SubC1*: número de carbonos primarios; *D604*: número de carbonos aromáticos sp^2 substituidos.

En este trabajo se propone la siguiente guía QSPR para la búsqueda de estructuras con valores favorables de K_H. Según el signo de los coeficientes de regresión de la ecuación (4.10), cuantos menores sean los valores numéricos de *ATSC1e*, *LFEA*, *LFEBH*, *ToPSA* y *D604*, y mayores sean los valores de *MCS48* y *SubC1* en una estructura considerada, mayor será el valor predicho de su constante de Henry.

El dominio de aplicación del modelo QSPR propuesto demuestra que 6 compuestos del total de 177 del conjunto de predicción tienen un valor de influencia ligeramente superior al límite h^*: nitrato de 3-pentilo, nitrato de isopentilo, 2-nitrooxi-etanol, 2-nitrooxi-3-butanol, 1-nitrooxi-2-butanol y 2-nitrooxi-1-butanol. No obstante, puede considerarse que sus predicciones son confiables, ya que tales moléculas están próximas a moléculas de los conjuntos de calibración y validación, que también exceden h^* en la Figura 4.15.

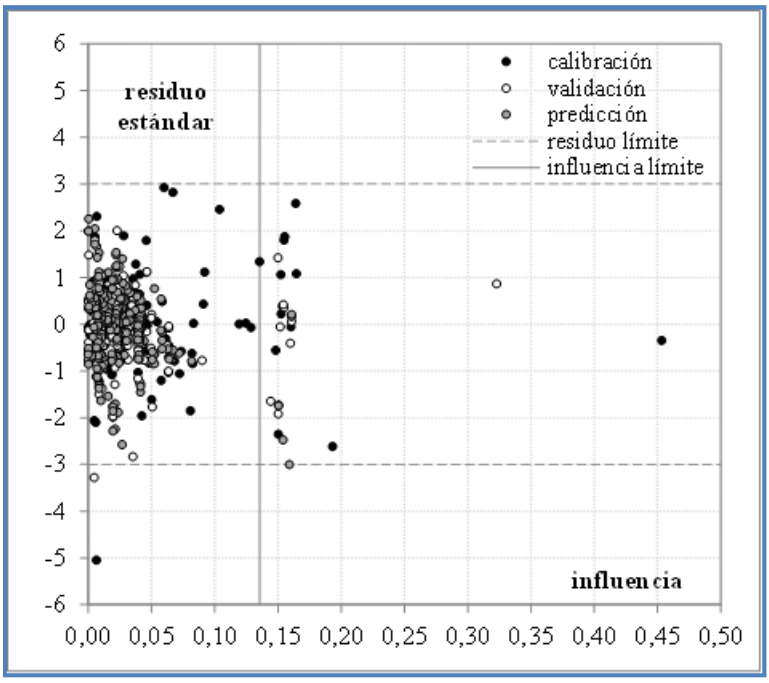


Figura 4.15. Gráfico de Williams para la ecuación (4.10). Influencia límite h=0.136.*

La calidad de las predicciones de $\log K_H$ conseguida mediante la ecuación (4.10) se compara con la obtenida con el módulo HENRYWIN de EPI Suite. Los 530 compuestos estudiados conducen a $RMS = 0.77$ en el caso de la ecuación (4.10), mientras que para HENRYWIN se obtiene $RMS = 0.76$. Se concluye que ambos modelos generan predicciones satisfactorias de la propiedad, a través de metodologías alternativas.

Cabe mencionar que nuestro modelo propuesto de la ecuación (4.10) se diferencia del modelo incluido en el programa HENRYWIN en que este último se calibra utilizando no solo datos experimentales de la constante de Henry, sino que también incorpora estimaciones empíricas de K_H. Dichas estimaciones se calculan con el cociente entre los valores experimentales de presión de vapor (p_v) y solubilidad acuosa (S_w) de cada compuesto. Esto hace que la base de datos empleada por HENRYWIN sea mayor a la utilizada en el estudio actual, generando un modelo calibrado con un mayor número de moléculas y que posee mayor dominio de aplicabilidad, aunque la propiedad que se ajusta en HENRYWIN no es puramente experimental, sino que incluye cierto error procedente de las aproximaciones de K_H.

Un siguiente paso en el estudio actual es aplicar el modelo QSPR de la ecuación (4.10) para predecir la propiedad en compuestos que no poseen valores experimentales de K_H, pero sí poseen estimaciones de K_H basadas en el cociente empírico p_v/S_w extraídas de la base de datos de HENRYWIN. La Tabla 26.A reúne las 809 estructuras que se predicen y que son parte del dominio de aplicabilidad de la ecuación (4.10), con valores de influencia menores al valor límite h^*. El cálculo del error en la predicción de la estimación empírica de $\log K_H$ es el siguiente: $RMS = 1.60$ (ecuación 4.10) y $RMS = 1.49$ (HENRYWIN). Este resultado es de esperarse, pues se entiende que el error provisto por el programa HENRYWIN es menor: las estimaciones p_v/S_w de muchos compuestos fueron utilizadas a la hora de calibrar dicho programa, como se mencionara anteriormente.

Finalmente, los valores numéricos de los mejores descriptores de la propiedad K_H para las 1339 estructuras químicas ensayadas se incluyen en la Tabla 27.A.

4.6.3. Conclusiones

El presente trabajo permitió establecer un modelo QSPR basado en descriptores independientes de la conformación para predecir la constante de Henry. La base de datos está conformada por estructuras químicas muy heterogéneas, aunque las predicciones de la ecuación lineal demostraron ser aceptables al compararse con las obtenidas con el programa HENRYWIN.

En esta propiedad específica estudiada, no fue posible mejorar la calidad predictiva del modelo mediante la inclusión de un descriptor flexible que se combine con otros descriptores rígidos seleccionados, dado que al hacerlo causa un deterioro de los parámetros de validación. No obstante, el modelo QSPR permite predecir la constante de Henry en compuestos que no poseen valores experimentales de K_H y que pertenezcan a su dominio de aplicabilidad.

Bibliografía

1. Lyman WJ. Adsorption coefficient for soils and sediments. *Handb Chem Prop Estim Methods Environ Behav Org Compd Am Chem Soc Washington, DC 1990 p 4 1-4 33 3 fig, 11 tab, 44 ref*. 1990.
2. Sparks DL. Environmental Soil Chemistry. *Acad Press Tokyo, Japan*. 2013:267.
3. Jury WA. Adsorption of organic chemicals onto soil. In Vadose Zone Modeling of Organic Pollutants. *Henn, SC Melancon, SM, Eds; Lewis Publ New York, NY, USA*. 1986:177–189.
4. Gawlik, B.M.; Sotiriou, N.; Feicht, E.A.; Schulte-Hostede, S.; Kettrup A. Alternatives for the determination of the soil adsorption coefficient, KOC, of non-ionic organic compounds-A review. *Chemosphere*. 1997;34:2525–2551.
5. Hansch, C.; Leo A. Exploring QSAR. Fundamentals and Applications in Chemistry and Biology. *Am Chem Soc Washington, DC, USA*. 1995.
6. Kubinyi H. QSAR: Hansch Analysis and Related Approaches. *Wiley-Interscience New York, NY, USA*. 2008.
7. Puzyn, T.; Leszczynski, J.; Cronin MTD. Recent Advances in QSAR Studies: Methods and Applications. *Springer Sci Bus Media BV Houten, Netherlands*. 2010.
8. Katritzky, A.R.; Goordeva EV. Traditional topological indices vs. Electronic, geometrical, and combined molecular descriptors in QSAR/QSPR research. *J Chem Inf Comput Sci*. 1993;33:835–857.
9. Diudea MVE. QSPR/QSAR Studies by Molecular Descriptors. *Nov Sci Publ New York, NY, USA*. 2001.
10. Todeschini, R.; Consonni V. Molecular Descriptors for Chemoinformatics (Methods and Principles in Medicinal Chemistry). *Wiley-VCH Weinheim, Ger*. 2009.
11. Sabljic, A.; Gusten, H.; Verhaar, H.; Hermens J. QSAR modeling of soil sorption. Improvements and systematics of log Koc vs. Log kow correlations. *Chemosphere*. 1995;31:4489–4514.
12. Sabljic A. Predictions of the nature and strength of soil sorption of organic pollutants by molecular topology. *J Agric Food Chem*. 1984;32(2):243-246.
13. Bahnick DA, Doucette WJ. Use of molecular connectivity indices to estimate

soil sorption coefficients for organic chemicals. *Chemosphere*. 1988;17(9):1703-1715.

14. Duchowicz, P.R.; González, M.P.; Helguera, A.M.; Cordeiro, M.N.D.S.; Castro EA. Application of the replacement method as novel variable selection in QSPR. 2. Soil sorption coefficients. *Chemom Intell Lab Syst*. 2007;88:197–203.

15. Goudarzi, N.; Goodarzi, M.; Araujo, M.C.U.; Galvão RKH. QSPR modeling of soil sorption coefficients (Koc) of pesticides using SPA-ANN and SPA-MLR. *J Agric Food Chem*. 2009;57:7153–7158.

16. Shao, Y.; Liu, J.; Wanga, M.; Shi, L.; Yao, X.; Gramatica P. Integrated QSPR models to predict the soil sorption 2014, coefficient for a large diverse set of compounds by using different modeling methods. *Atmos Environ*. 2014;88:212–218.

17. Gramatica, P.; Giani, E.; Papa E. Statistical external validation and consensus modeling: A QSPR case study for Koc prediction. *J Mol Graph Model*. 2007;25:755–766.

18. Duchowicz, P.R.; Comelli, N.C.; Ortiz, E.V.; Castro, E.A. QSAR study for carcinogenicity in a large set of organic compounds. *Curr Drug Saf*. 2012;7:282–288.

19. Toropov, A.A.; Toropova, A.P.; Benfenati, E.; Gini G. OCWLGI descriptors: Theory and praxis. *Curr Comput Aided Drug Des*. 2013;9:226–232.

20. Meylan W, Howard PH, Boethling RS. Molecular topology/fragment contribution method for predicting soil sorption coefficients. *Environ Sci Technol*. 1992;26(8):1560-1567.

21. Ibezim, E.; Duchowicz, P.R.; Ortiz, E.V.; Castro EA. QSAR on aryl-piperazine derivatives with activity on malaria. *Chemom Intell Lab Syst*. 2012;110:81–88.

22. Mullen, L.M.A.; Duchowicz, P.R.; Castro EA. QSAR treatment on a new class of triphenylmethyl-containing compounds as potent anticancer agents. *Chemom Intell Lab Syst*. 2011;107:269–275.

23. Toropov, A.A.; Leszczynska, D.; Leszczynski JP. Predicting water solubility and octanol water partition coefficient for carbon nanotubes based on the chiral vector. *Comput Biol Chem*. 2007;31:127–128.

24. A.A. Toropov, A.P. Toropova, E. Benfenati, G. Gini, T. Puzyn, D. Leszczynska, and J. Leszczynski. Novel application of the coral software to model cytotoxicity

of metal oxide nanoparticles to bacteria Escherichia coli. *Chemosphere*. 2012;89:1098–1102.

25. A.P. Toropova, A.A. Toropov, S.E. Martyanov, E. Benfenati, G. Gini, D. Leszczynska and JL. Coral: QSAR modeling of toxicity of organic chemicals towards Daphnia magna. *Lab, Chemom Intel Syst*. 2012;110:177–181.

26. Todeschini, R.; Consonni V. *Molecular Descriptors for Chemoinformatics (Methods and Principles in Medicinal Chemistry); Wiley-VCH: Weinheim, Germany.*; 2009.

27. Golbraikh, A.; Tropsha A. Beware of q2! *J Mol Graph Model*. 2002;20:269–276.

28. Estimation Program Interface Epi Suite. US EPA. Available online: https://www.epa.gov/tsca-screening-tools/epi-suitetm-estimation-programinterface (accessed on 29 July 2016).

29. ACD/ChemSketch. Available online: http://www.acdlabs.com (accessed on 29 July 2016). 2016.

30. PaDEL. Available online: http://www.yapcwsoft.com/ (accessed on 29 July 2016). 2016.

31. Yap CW. PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. *J Comput Chem*. 2011;32:1466–1474.

32. Duchowicz, P.R.; Castro, E.A.; Fernández FM. Alternative algorithm for the search of an optimal set of descriptors in QSAR-QSPR studies. *MATCH Commun Math Comput Chem*. 2006;55:179–192.

33. Duchowicz, P.R.; Castro, E.A.; Fernández, F.M.; González, M. A new search algorithm of QSPR/QSAR theories: Normal boiling points of some organic molecules. *Chem Phys Lett*. 2005;412:376–380.

34. Duchowicz, P.R.; Talevi, A.; Bruno-Blanch, L.E.; Castro, E.A. QSPR study for the prediction of aqueous solubility of drug-like compounds. *Bioorg Med Chem Lett*. 2008;16:7944–7955.

35. Goodarzi, M.; Duchowicz, P.R.; Wu, C.H.; Fernández, F.M.; Castro, E.A. New hybrid genetic based support Model., vector regression as QSAR approach for analyzing flavonoids-GABA(A) complexes. *J Chem Inf*. 2009;49:1475–1485.

36. Pomilio, A.B.; Giraudo, M.A.; Duchowicz, P.R.; Castro, E.A. QSPR analyses for aminograms in food: Citrus juices and concentrates,. *Food Chem*.

2010;123:917–927.

37. Talevi, A.; Goodarzi, M.; Ortiz, E.V.; Duchowicz, P.R.; Bellera, C.L.; Pesce, G.; Castro, E.A.; Bruno-Blanch, L.E. Prediction of drug intestinal absorption by new linear and non-linear QSPR. *Eur J Med Chem*. 2011;46:218–228.

38. Pasquale, G.; Romanelli, G.P.; Autino, J.C.; García, J.; Ortiz, E.V.; Duchowicz, P.R. Quantitative Chem., structure-activity relationships on chalcone derivatives: Mosquito larvicidal studies. *J Agric Food*. 2012;60:692–697.

39. Matlab 7.0. Available online: http://www.mathworks.com (accessed on 29 July 2016).

40. CORAL, http://www.insilico.eu/coral

41. Wold, S. Eriksson L. Statistical validation of qsar results. In Chemometrics Methods in Molecular Design; van de Waterbeemd, H., Ed.; VCH: Weinheim, Germany. 1995:309–318.

42. Gramatica P. Principles of qsar models validation: Internal and external. *QSAR Comb Sci*. 2007;26:694–701.

43. Eriksson, L.; Jaworska, J.; Worth, A.P.; Cronin, M.T.; McDowell, R.M.; Gramatica P. Methods for reliability QSARS and uncertainty assessment and for applicability evaluations of classification- and regression-based. *Environ Heal Perspect*. 2003;111:1361–1375.

44. ECETOC. The Role of Bioaccumulation in Environmental Risk Assessment: The Webs. *Aquat Environ Relat Food*. 1995.

45. Gobas, F., Morrison, H.A. Bioconcentration and biomagnification in the aquatic environment. In: Boethling, R.S., Mackay, D. (Eds.), Handbook of Property Estimation Methods for Chemicals. *Environ Heal Sci CRC Press Boca Raton, USA*. 2000.

46. ECHA. Guidance on information requirements and chemical safety Assessment., assessment. *Chapter R11 PBT*. 2012.

47. Barron MG. Bioconcentration. Will water-borne organic chemicals accumulate in aquatic animals? *Environ Sci Technol*. 1990;24:1612–1618.

48. HESI I. JRC/SETAC-EU. Workshop on Bioaccumulation Assessments. *Dutch Congr Centre, Hague, Netherlands, 5–6 May*. 2006.

49. R. Todeschini and V. Consonni. Molecular Descriptors for Chemoinformatics,

Methods and Principles. *Med Chem Wiley-VCH, Weinheim,*. 2009.

50. Pavan, M., Worth, A.P., Netzeva TI. Review of QSAR models for bioconcentration. *JRC Rep EUR EN I-21020*. 2006.

51. Veith, G.D., DeFoe, D.L., Bergstedt, B.V. Measuring and estimating the bioconcentration factor of chemicals in fish. J. Fish. *Res Board Can*. 1979;36:1040–1048. http://dx.doi.org/10.1139/f79-146.

52. Mackay D. Correlation of bioconcentration factors. *Environ Sci Technol*. 1982;16:274–278.

53. Bintein, S., Devillers, J., Karcher, W. Nonlinear dependence of fish Environ., bioconcentration on n-octanol/water partition coefficient. *SAR QSAR Res*. 1993;1:29–39.

54. Connell, D.W., Hawker DW. Use of polynomial expressions to describe the 16, bioconcentration of hydrophobic chemicals by fish. *Ecotoxicol Environ Saf*. 1988;16:242–257.

55. Meylan WM, Howard PH, Boethling RS, Aronson D, Printup H, Gouchie S. Improved method for estimating bioconcentration/bioaccumulation factor from octanol/water partition coefficient. *Environ Toxicol Chem*. 1999;18(4):664-672.

56. Zhao, C., Boriani, E., Chana, A., Roncaglioni, A., Benfenati E. A new hybrid system of QSAR models for predicting bioconcentration factors (BCF). *Chemosphere*. 2008;73:1701–1707.

57. A. Gissi, D. Gadaleta, M. Floris, S. Olla, A. Carotti, E. Novelllino, E. Benfenati, and O. Nicolotti. An Purposes, alternative QSAR-based approach for predicting the bioconcentration factor for regulatory. *ALTEX*. 2014;31:23–26.

58. P.R. Duchowicz, N.C. Comelli, E.V. Ortiz and EAC. QSAR study for carcinogenicity in a large set of organic compounds. *Curr Drug Saf*. 2012;7:282–288.

59. P.R. Duchowicz, D.O. Bennardi, D.E. Baselo, E.L. Bonifazi, C. Rios-Luci, J.M. Padrón, G. Burton and RIM. QSAR on antiproliferative naphthoquinones based on a conformation-independent approach. *Eur J Med Chem*. 2014;77:176–184.

60. P.R. Duchowicz, S.E. Fioressi, D.E. Bacelo, L.M. Saavedra, A.P. Toropova and AAT. QSPR studies on refractive indices of structurally heterogeneous polymers. *Chemom Intel Lab Syst*. 2015;140:86–91.

61. J.F. Aranda, J.C. Garro Martinez, E.A. Castro, and P.R. Duchowicz. Conformation-independent QSPR approach for the soil sorption coefcient of heterogeneous compounds. *Int J Mol Sci*. 2016;17:1247.

62. A. Gissi, O. Nicolotti, A. Carotti, D. Gadaleta, A. Lombardo and EB. Integration of QSAR models for bioconcentration suitable for REACH. *I Sci Total Environ*. 2013;456–457:325–332.

63. Open Babel for Windows. Available at https://openbabel.org wiki/Category:Installation, 2017.

64. S.E. Fioressi, D.E. Bacelo, W.P. Cui, L.M. Saavedra, and P.R. Duchowicz. QSPR study on refractive indices of solvents commonly used in polymer chemistry using flexible molecular descriptors. *SAR QSAR Environ Res*. 2015;26:499–506.

65. H. Hong, Q. Xie, W. Ge, F. Qian, H. Fang, L. Shi, Z. Su, R. Perkins, and W. Tong J. Mold2, Molecular descriptors from 2D structures for chemoinformatics and toxicoinformatics. *Chem Inf Model*. 2008;48:1337–1344.

66. RECON version 5. 5/5. 3. Curt M. Breneman, William P. Katt, Martin Martinov, Marlon O. Rhem, N. Sukumar, Tracy R. Thompson, Christopher Whitehead and Dechuan Zhuang. (Rensselaer Polytechnic Institute, 2003). http://reccr.chem.rpi.edu/software.html.

67. B.K. Lavine, C.E. Davidson, C. Breneman, and W. Katt. Electronic van der Waals surface property Databases, descriptors and genetic algorithms for developing structure-activity correlations in olfactory. *J Chem Inf Comput Sci*. 2003;43:1890–1905.

68. R.F.W. Bader. Atoms in Molecules-A Quantum Theory. *Oxford Univ Press Oxford, UK*. 1990.

69. C.M. Breneman and M. Rhem. A QSPR analysis of HPLC column capacity factors for a set of high-energy Transferable, materials using electronic Van der Waals surface property descriptors computed by the atom equivalent method. *J Comput Chem*. 1997;18:182–197.

70. J.R. Valdes-Martini, C.R. García Jacas, Y. Marrero-Ponce, Y. Silveira Vaz 'd Almeida and CM, CAMD-BIR Unit, CENDA Number of register: 2373-2012 2012., 763 D by [190. 191. 137. 201. at 16:51 12 O 2017. QuBiLS-MAS: Free Software for molecular descriptors calculator from Quadratic, Bilinear and Linear Maps based on Graph–Theoretic Electronic-Density Matrices and Atomic

weighting, Version 1.0. *SAR QSAR Environ Res*.

71. A.H. Morales, P.R. Duchowicz, M.A. Cabrera Pérez, E.A. Castro, M.N.D.S. Cordeiro and MPG. Carcinogenic, Application of the replacement method as a novel variable selection strategy in QSAR. 1. potential. *Chemom Intel Lab Syst*. 2006;81:180–187.

72. P. Gramatica and A. Sangion. A historical excursus on the statistical validation parameters for QSAR models: A clarifcation concerning metrics and terminology. *J Chem Inf Model* . 2016;56:1127–1131.

73. K. Roy, R.N. Das, P. Ambure, and R.B. Aher. Be aware of error measures. Further studies on validation of predictive QSAR models. *Chemom Intel Lab Syst*. 2016;15:18–33.

74. C. Rojas, P.R. Duchowicz, P. Tripaldi, and R. Pis Diez. Quantitative structure-property relationship analysis for the retention index of fragrance-like compounds on a polar stationary phase. *J Chromatogr*. 2015;A 1422:277–288.

75. P.R. Duchowicz, S.E. Fioressi, E.A. Castro, K. Wróbel, N.E. Ibezim and DEB. Conformation-independent QSAR study on human epidermal growth factor receptor-2 (HER2) inhibitors. *Chem Sel*. 2017;2:3725–3731.

76. L. Kaufman and P.J. Rousseeuw. Finding Groups in Data: An Introduction to Cluster Analysis. *Wiley, New York,*. 2005.

77. K. Roy, S. Kar and PA. On a simple approach for determining applicability domain of QSAR models. *Chemom Intel Lab Syst*. 2015;145:22–29.

78. N.R. Draper and H. Smith. Applied Regression Analysis. *John Wiley Sons, New York*. 1981.

79. A.R. Katritzky and E.V. Goordeva. Traditional topological indices vs. electronic, geometrical, And combined molecular descriptors in QSAR/QSPR research. *J Chem Inf Comput Sci*. 1993;33:835–857.

80. M.V.E. Diudea. QSPR/QSAR Studies by Molecular Descriptors. *Nov Sci Publ New York*. 2001.

81. K. Roy and P.P. Roy. Comparative chemometric modeling of cytochrome 3A4 inhibitory activity of structurally diverse compounds using stepwise MLR, FAMLR, PLS, GFA, G/PLS and ANN techniques. *Eur J Med Chem*. 2009;44:2913–2922.

82. Agency USEP. Finalization of Guidance on Incorporation of Water Treatment Effects on Pesticide Removal and Transformations in Drinking Water Exposure Assessments. 2017.

83. Hamilton, D.; Ambrus, A.; Dieterle, R.; Felsot, A.; Harris, C.; Holland, P.; Katayama, A.; Kurihara, N.; Linders J. Regulatory limits for pesticide residues in water (IUPAC Technical Report). Pure and Applied Chemistry. *Unsworth, J.* 2003;75 (8):1123-1155.

84. Cronin MT. Predicting chemical toxicity and fate. *CRC Press*. 2004.

85. Das, R.N.; Roy K. QSPR with extended topochemical atom (ETA) indices. 4. Modeling aqueous solubility of drug like molecules and agrochemicals following OECD guidelines. *Struct Chem*. 2013;24 (1):303-331.

86. Toropov, A.A.; Toropova, A.P.; Benfenati, E.; Gini, G.; Leszczynska, D.; Leszczynski J. CORAL: QSPR model of water solubility based on local and global SMILES attributes. *Chemosphere*. 2013;90 (2):877-880.

87. Zeng, X.-L.; Wang, H.-J.; Wang Y. QSPR models of n-octanol/water partition coefficients and aqueous solubility of halogenated methyl-phenyl ethers by DFT method. *Chemosphere*. 2012;86 (6):619-625.

88. Wilczyńska-Piliszek, A.J.; Piliszek, S.; Falandysz J. QSAR and ANN for the estimation of water solubility of 209 polychlorinated trans-azobenzenes. *J Environ Sci Heal Part A*. 2012;47 (2):155-166.

89. Bhhatarai, B.; Gramatica P. Prediction of aqueous solubility, vapor pressure and critical micelle concentration for aquatic partitioning of perfluorinated chemicals. *Environ Sci Technol*. 2010;45 (19):8120-8128.

90. Cappelli, C.I.; Manganelli, S.; Lombardo, A.; Gissi, A.; Benfenati E. Validation of quantitative structure–activity relationship models to predict water-solubility of organic compounds. *Sci Total Environ*. 2013;463:781-789.

91. Talevi, A.; Bellera, C.L.; Di Ianni, M.; Duchowicz, P.R.; Bruno-Blanch, L.E.; Castro EA. An integrated drug development approach applying topological descriptors. *Curr Comput Aided Drug Des,*. 2012;8 (3):172-181.

92. Lewis, K.; Green, A.; Tzilivakis, J.; Warner D. The Pesticide Properties DataBase (PPDB) Developed by the Agriculture & Environment Research Unit (AERU). *Univ Hertfordsh*. 2018.

93. Dassault Systèmes BIOVIA. Discovery Studio Modeling Environment. *San*

Diego Dassault Systèmes. 2017.

94. Rojas, C.; Tripaldi, P.; Duchowicz PR. A new QSPR study on relative sweetness. *Int J Quant Struct Relationships*. 2016;1 (1):78-93.

95. Gadaleta, D.; Mangiatordi, G.F.; Catto, M.; Carotti, A.; Nicolotti O. Applicability domain for QSAR models: where theory meets reality. *Int J Quant Struct Relationships*. 2016;1 (1):45-63.

96. Verma, R.P.; Hansch C. An approach toward the problem of outliers in QSAR. *Bioorg Med Chem*. 2005;13 (15):4597-4621.

97. Wishart, D.S.; Feunang, Y.D.; Marcu, A.; Guo, A.C.; Liang, K.; Vázquez-Fresno, R.; Sajed, T.; Johnson, D.; Li, C.; Karu N. HMDB 4.0: the human metabolome database for 2018. *Nucleic Acids Res*. 2017;46 (D1):D608-D617.

98. Roy K. On some aspects of validation of predictive quantitative structure–activity relationship models. *Expert Opin Drug Discov*. 2007;2 (12):1567-1577.

99. Barceló D. Trace Determination of Pesticides and their Degradation Products in Water. *Elsevier*. 1997;Vol. 19.

100. Chau AS. Analysis of Pesticides in Water: Volume I: Significance, Principles, Techniques, and Chemistry of Pesticides. *CRC Press*. 2018.

101. Howard P. Handbook of environmental fate and exposure data: for organic chemicals, volume III pesticides. *Routledge*. 2017.

102. dos Reis, R. R., Sampaio, S. C., & de Melo EB. An alternative approach for the use of water solubility of nonionic pesticides in the modeling of the soil sorption coefficients. *Water Res*. 2014;53:191-199.

103. European Chemical Agency (ECHA). Guidance on information requirements and chemical safety assessment—Chapter R.7a: endpoint specific guidance. 2012.

104. Dedek W. Solubility factors affecting pesticide penetration through skin and protective clothing. *Stud Environ Sci Elsevier*. 1980;Vol. 7:47-50).

105. Cronin Mark TD LDJ. Predicting chemical toxicity and fate. *Boca Raton Florida CRC Press*. 2004.

106. Hansch C. Quantitative approach to biochemical structure-activity relationships. *Acc Chem Res*. 1969;2(8):232-239.

107. Hansch C, Fujita T. p-σ-π Analysis. A method for the correlation of biological

activity and chemical structure. *J Am Chem Soc*. 1964;86(8):1616-1626.

108. Draber W, Fujita T. *Rational Approaches to Structure, Activity, and Ecotoxicology of Agrochemicals*. CRC Press; 1992.

109. Palmer, D. S., O'Boyle, N. M., Glen, R. C., & Mitchell JB. Random forest models to predict aqueous solubility. *J Chem Inf Model*. 2007;47(1):150-158.

110. Delaney JS. ESOL: estimating aqueous solubility directly from molecular structure. *J Chem Inf Comput Sci*. 2004;44(3):1000-1005.

111. Bhhatarai B GP. Prediction of aqueous solubility, vapor pressure and critical Environ, micelle concentration for aquatic partitioning of perfluorinated chemicals. *Env Sci Technol*. 2011;45:8120–8.

112. Raevsky, O. A., Polianczyk, D. E., Grigorev, V. Y., Raevskaja, O. E., & Dearden JC. In silico prediction of aqueous solubility: A comparative study of local and global predictive models. *Mol Inform*. 2015;34(6- 7):417-430.

113. Bouakkadia, A., Haddag, H., Bouarra, N., & Messadi D. QSPR study of the water solubility of a diverse set of agrochemicals: hybrid (GA/MLR) approach. *Synthèse Rev des Sci la Technol*. 2016;32(1):12-21.

114. Klimenko K, Kuz'min V, Ognichenko L, et al. Novel enhanced applications of QSPR models: Temperature dependence of aqueous solubility. *J Comput Chem*. 2016;37(22):2045-2051.

115. Raevsky OA, Grigorev VY, Polianczyk DE, Raevskaja OE, Dearden JC. Six global and local QSPR models of aqueous solubility at pH= 7.4 based on structural similarity and physicochemical descriptors. *SAR QSAR Environ Res*. 2017;28(8):661-676.

116. Cappelli, C. I., Manganelli, S., Lombardo, A., Gissi, A., & Benfenati E. Validation of quantitative structure–activity relationship models to predict water-solubility of organic compounds. *Sci Total Environ*. 2013;463:781-789.

117. Sushko I, Novotarskyi S, Körner R, Pandey AK, Kovalishyn VV, Prokopenko VV et al. Sushko I, Novotarskyi S, Körner R, Pandey AK, Kovalishyn VV, Prokopenko VV, et al. Applicability domain for in silico models to achieve accuracy of experimental measurements. J Chemom 2010;24:202–8. *J Chemom*. 2010;24:202–8.

118. Netzeva TI, Worth A, Aldenberg T, Benigni R, Cronin MT, Gramatica P et al. Current status of methods for defining the applicability domain of (quantitative)

structure– 52, activity relationships. The report and recommendations of ECVAM Workshop. *Altern Lab Anim*. 2005;33:155–73.

119. Brennan RA, Nirmalakhandan N, Speece RE. Comparison of predictive methods for Henrys law coefficients of organic chemicals. *Water Res*. 1998;32(6):1901-1911.

120. Altschuh J, Brüggemann R, Santl H, Eichinger G, Piringer OG. Henry's law constants for a diverse set of organic chemicals: Experimental determination and comparison of estimation methods. *Chemosphere*. 1999;39(11):1871-1887.

121. Suntio, L. R., Shiu, W. Y., Mackay, D., Seiber, J. N., & Glotfelty D. Critical review of Henry's law constants for pesticides. In Reviews of Environmental Contamination and Toxicology. *Springer, New York, NY*. 1988:1-59.

122. Sander R. Compilation of Henry's law constants (version 4.0) for water as solvent. *Atmos Chem Phys*. 2015;15(8).

123. Kieckbusch TG, King CJ. An improved method of determining vapor-liquid equilibria for dilute organics in aqueous solution. *J Chromatogr Sci*. 1979;17(5):273-276.

124. Chai X-S, Zhu JY. Indirect headspace gas chromatographic method for vapor–liquid phase equilibrium study. *J Chromatogr A*. 1998;799(1-2):207-214.

125. Ettre LS, Welter C, Kolb B. Determination of gas-liquid partition coefficients by automatic equilibrium headspace-gas chromatography utilizing the phase ratio variation method. *Chromatographia*. 1993;35(1-2):73-84.

126. Chai XS, Zhu JY. Erratum to"Indirect headspace gas chromatographic method for vapor-liquid phase equilibrium study"[J. Chromatogr. A 799 (1998) 207-214]. *J Chromatogr A*. 2003;1020(2):283-284.

127. Richon D. New equipment and new technique for measuring activity coefficients and Henry's constants at infinite dilution. *Rev Sci Instrum*. 2011;82(2):25108.

128. Staudinger J, Roberts P V. A critical review of Henry's law constants for environmental applications. *Crit Rev Environ Sci Technol*. 1996;26(3):205-297.

129. Spencer, W. F., & Cliath MM. Measurement of pesticide vapor pressures. In Residue reviews. *Springer, New York, NY*. 1983:57-71.

130. Bidleman TF 56. Estimation of vapor pressures for nonpolar organic compounds by capillary gas chromatography. *Anal Chem*. 1984;13:2490-2496.

131. Gossett JM. Measurement of Henry's law constants for C1 and C2 chlorinated hydrocarbons. *Environ Sci Technol*. 1987;21(2):202-208.

132. Yin, C., & Hassett JP. Gas-partitioning approach for laboratory and field studies of mirex fugacity in water. *Environ Sci Technol*. 1986;20(12):1213-1217.

133. Modarresi, H., Modarress, H., & Dearden JC. QSPR model of Henry's law constant for a diverse set of organic chemicals based on genetic algorithm-radial basis function network approach. *Chemosphere*. 2007;66(11):2067-2076.

134. Mackay, D., Shiu, W.S., Ma, K.C. Henry's law constant. In: Boethling, R.S., Mackay, D. (Eds.), Handbook of Property Estimation Lewis, Methods for Chemicals: Environmental and Health Sciences. *Boca Raton, FL, USA*. 2000:69–87.

135. Cramer, R.D. BC (DEF) parameters. 2. An empirical structurebased scheme for the prediction of some physical properties. *J Am Chem Soc*. 1980;102:1849–1859.

136. Hine, J., Mookerjee, P.K. The intrinsic hydrophilic character of Contributions., organic compounds. Correlations in terms of structural. *J Org Chem*. 1975;40:292–298.

137. Meylan, W.M., Howard, P.H. Bond contribution method for estimating Henry's law constants. *Environ Sci Technol*. 1991;10:1283–1293.

138. Meylan, W.M., Howard, P.H. HENRYWIN 3.10. *Syracuse Res Syracuse, NY*. 2000.

139. Dearden, J. C., & Schüürmann G. Quantitative structure- property relationships for predicting henry's law constant from molecular structure. *Environ Toxicol Chem*. 2003;22(8):1755-1770.

140. Shimotori, T., Arnold, W.A. Henry's law constants of chlorinated ethylenes in aqueous alcohol solutions: measurement Estimation, and thermodynamic analysis. *J Chem Eng Data*. 2002;47:183–190.

141. Ornektekin, S., Paksoy, H. O., & Demirel Y. The Performance of UNIFAC and Related Group-Contribution Models. 2. Prediction of Henry Law Constants. *Thermochim Acta,*. 1996;287(2):251-259.

142. Austin ND, Sahinidis N V, Trahan DW. Computer-aided molecular design: An introduction and review of tools, applications, and solution techniques. *Chem*

Eng Res Des. 2016;116:2-26.

143. Worth AP, Bassan A, De Bruijn J, et al. The role of the European Chemicals Bureau in promoting the regulatory use of (Q) SAR methods. *SAR QSAR Environ Res*. 2007;18(1-2):111-125.

144. Modarresi H, Modarress H, Dearden JC. Henry's law constant of hydrocarbons in air–water system: The cavity ovality effect on the non-electrostatic contribution term of solvation free energy. *SAR QSAR Environ Res*. 2005;16(5):461-482.

145. Modarresi H, Modarress H, Dearden JC. QSPR model of Henry's law constant for a diverse set of organic chemicals based on genetic algorithm-radial basis function network approach. *Chemosphere*. 2007;66(11):2067-2076.

146. Duchowicz PR, Garro JCM, Castro EA. QSPR study of the Henry's Law constant for hydrocarbons. *Chemom Intell Lab Syst*. 2008;91(2):133-140.

147. Goodarzi M, Ortiz E V, Coelho L dos S, Duchowicz PR. Linear and non-linear relationships mapping the Henry's law parameters of organic pesticides. *Atmos Environ*. 2010;44(26):3179-3186.

148. O'Loughlin DR, English NJ. Prediction of Henry's Law Constants via group-specific quantitative structure property relationships. *Chemosphere*. 2015;127:1-9.

149. Abraham, M.H., Andonian-Haftvan, J., Whiting, G.S., Leo, A. T, R.S. Hydrogen bonding. Part 34. The factors that influence the solubility of gases and vapours in water at 298 K, and a new method for its determination. *J Chem Soc Perkin Trans*. 1994;2:1777–1779.

150. Katritzky, A.R., Mu, L., Karelson, M. A QSPR study of the 36, solubility of gases and vapors in water. *J Chem Inf Comput Sci*. 1996:1162–1168.

151. Dearden, J. C., Cronin, M. T. D., Sharra, J. A., Higgins, C., Boxall, A. B. A., Watts, C. D., & Schuüürmann GF. The prediction of Henry's law constant: a QSPR from fundamental considerations. Quantitative Structure− Activity Relationships. *Environ Sci VII,*. 1997:135-142.

152. Dearden, J. C., Ahmed, S. A., Cronin, M. T., & Sharra JA. QSPR prediction of Henry's law constant: improved correlation with new parameters. *Mol Model Predict Bioactivity Springer, Boston, MA*. 2000:273-274.

153. English, N.J., Carroll, D.G. Prediction of Henry's law constants by J., a

quantitative structure property relationship and neural networks. *Chem Inf Comput Sci*. 2001;41:1150–1161.

154. Yao, X., Liu, M., Zhang, X., Hu, Z., Fan, B. Radial basis Relationship, function network-based quantitative structure–property for the prediction of Henry's law constant. *Anal Chim Acta*. 2002;462:101– 117.

155. Yaffe, D., Cohen, Y., Espinosa, G., Arenas, A., Giralt, F. A fuzzy (QSPR), ARTMAP-based quantitative structure–property relationship Inf., for the Henry's law constant of organic compounds. *J Chem Comput Sci*. 2003;43:85–112.

156. Gharagheizi, F., Abbasi, R., & Tirandazi B. Prediction of Henry's law constant of organic compounds in water from a new group-contribution-based model. *Ind Eng Chem Res*. 2010;49(20):10149-10152.

157. Gharagheizi, F., Ilani-Kashkouli, P., Mirkhani, S. A., Farahani, N., & Mohammadi AH. QSPR molecular approach for estimating Henry's law constants of pure compounds in water at ambient conditions. *Ind Eng Chem Res*. 2012;51(12):4764-4767.

158. Razdan NK, Koshy DM, Prausnitz JM. Henry's Constants of Persistent Organic Pollutants by a Group-Contribution Method Based on Scaled-Particle Theory. *Environ Sci Technol*. 2017;51(21):12466-12472.

159. Idorsia Pharmaceuticals Ltd.Thomas Sander. OSIRIS DataWarrior. Versión 4.7.3.

Capítulo 5. Estudios QSAR de la toxicidad de pesticidas

5.1. Estudio QSAR de la toxicidad aguda en la lombriz *Eisenia foetida*

Los plaguicidas se utilizan ampliamente en todo el mundo en la agricultura para la protección de las plantas y para aumentar los rendimientos de producción y la calidad de los productos agrícolas. Sin embargo, tienen una gran desventaja: la toxicidad[1]. Como resultado de su uso excesivo, se encuentran como residuos en el medio ambiente[2].

Numerosos estudios subrayan la contaminación ambiental causada por los pesticidas. Crean un riesgo para el medio ambiente, los seres humanos, los animales y otros organismos[3]. La evaluación de riesgos para los plaguicidas puede proporcionar una protección contra la contaminación. Una de las formas de medir el riesgo para el medio ambiente y los seres humanos es determinar la toxicidad aguda de los plaguicidas.

Estos se prueban contra una variedad de animales, incluidos mamíferos, aves, peces e invertebrados[4]. Las lombrices representan el 60-80% de la biomasa total del suelo de los invertebrados en el suelo y juegan un papel importante en la estructura y en el aumento de los nutrientes de los suelos.

Se han desarrollado varios protocolos para evaluar los efectos de los plaguicidas en las lombrices de tierra (*Eisenia foetida*), entre los cuales el más conocido es la guía 207 de la OCDE, una prueba de suelo artificial de 14 días[5,6]. Por lo tanto, las lombrices como *Eisenia foetida* pueden usarse con éxito como bioindicadores para la evaluación de la toxicidad aguda de los plaguicidas en los ecosistemas terrestres[7].

Desde estos puntos de vista, los efectos subletales de los plaguicidas sobre el crecimiento, la reproducción y el comportamiento de la lombriz de tierra por la exposición del suelo deben ser examinados en su registro de la UE, junto con la evaluación del riesgo de bioacumulación en las aves y los mamíferos, cuando *logKow* del pesticida es mayor a 3.

Deben considerarse las propiedades fisicoquímicas de los plaguicidas, que controlan la adsorción/desorción, y el secuestro en las partículas del suelo o su asociación con la materia orgánica disuelta[8–10].

Para comprender mejor el potencial tóxico y la posible bioacumulación de plaguicidas debe estudiarse el metabolismo y la información sobre los metabolitos[11].

Con el fin de proteger el medio ambiente de manera efectiva, las agencias reguladoras (por ejemplo, la Directiva de la Comunidad Europea: Directiva del Consejo 2009) requieren una evaluación de la toxicología de todos los plaguicidas. Lamentablemente, la determinación experimental de la toxicidad requiere un tiempo considerable, es costosa y genera un dilema ético (demandas para reducir o abolir las pruebas en animales)[4,12]. Por lo tanto, es necesario desarrollar métodos para predecir la actividad biológica a partir de las características estructurales.

El uso de métodos predictivos, basados en herramientas computacionales, es una opción rápida, económica y ética para evaluar la toxicidad de los plaguicidas en animales[13]. Estos métodos comprenden a las Relaciones Cuantitativas Estructura-Actividad (QSAR).

Hasta ahora, se ha desarrollado una gran cantidad de modelos QSAR para predecir la toxicidad aguda de los plaguicidas[12,14–19]. Sin embargo, no hay estudios dedicados a predecir la toxicidad aguda de plaguicidas en *Eisenia foetida*.

5.1.1. Datos experimentales de PPDB (79 moléculas)

Los datos de toxicidad aguda se obtienen de la base de datos de propiedades de pesticidas (PPDB), actualizada al 10-3-2017, y que fuera desarrollada por la Unidad de Investigación de Agricultura y Medio Ambiente en la Universidad de Hertfordshire[20]. La actividad se expresa como la concentración letal media (LC_{50}): corresponde a la concentración ($mg.kg^{-1}$) de pesticida que conduce a la muerte del 50% de la lombriz *Eisenia foetida*. Los valores de toxicidad aguda se convierten en $\log LC_{50}$.

Se extraen de la base de datos PPDB a 143 pesticidas de estructuras químicas heterogéneas y se dividen en 2 grupos. El primer grupo de 79 pesticidas incluye datos verificados por la base PPDB[20], que tienen mayor grado de confianza, por lo que se utiliza dicho conjunto para establecer el modelo QSAR. El segundo grupo de compuestos es el conjunto de ensayo: contiene 21 plaguicidas con datos no verificados o menos confiables. Ambos grupos de compuestos se encuentran en la Tabla 5.1.A.

5.1.2. Desarrollo del modelo QSAR

Se dibujan las estructuras químicas con ACD ChemSketch[21] en formato molecular MDL mol (V2000). La conversión de las moléculas se realizó con el programa de Windows Open Babel[22]. Cuando se tienen sales como 2 estructuras desconectadas se las dibuja tal cual, sin conectar, pe. maneb (no se remueven contraiones).

Los descriptores moleculares se calcularon mediante los programas PaDEL (14464), Recon (248), Mold^2 (777), DataWarrior (34) y QuBiLs-MAS (8448), por lo que se obtiene 23971 descriptores. El análisis sobre el conjunto de calibración de la dependencia lineal, exclusión de descriptores de valores únicos, y descriptores vacíos, conduce a $D = 2055$ variables estructurales.

Luego de realizar algunas pruebas preliminares sobre las 79 moléculas para establecer regresiones lineales, no se consigue establecer ajustes satisfactorios de los datos con los descriptores rígidos ($R^2 < 0.5$). Por tanto, se

analiza el rango de pesos moleculares estudiado y se opta por trabajar en el rango entre 73 y 296 ($g\,mol^{-1}$) que comprende a 58 moléculas. En la Tabla 5.1.A se incluyen los detalles de los compuestos estudiados, y se emplean los siguientes símbolos para su identificación: ^ conjunto de validación; * conjunto de predicción; ** conjunto de estimación.

Debido al tamaño reducido del presente conjunto molecular, consistente en 58 moléculas, se decide emplear el 70 % de las moléculas para la calibración del modelo, y partes iguales para los dos conjuntos restantes (15%). Las mejores regresiones lineales QSAR de 1-5 descriptores, obtenidas mediante RM y descriptores de naturaleza rígida, involucran a los descriptores de la Tabla 5.1.

Tabla 5.1. Modelos lineales para la toxicidad aguda en *Eisenia foetida.*

d	descriptores	R^2_{cal}	RMS_{cal}	R^2_{val}	RMS_{val}	R^2_{pred}	RMS_{pred}
1	*D017*	0.21	0.69	0.14	0.72	0.54	0.44
2	*GATS4m; APC8NX*	0.48	0.56	0.09	0.65	0.60	0.48
3	*GATS4m; SHBint4; APC8NX*	0.61	0.48	0.67	0.40	0.60	0.42
4	***GATS4m; nBondsD; SHBint4; APC8NX***	**0.72**	**0.41**	**0.70**	**0.38**	**0.87**	0.31
5	*HybRatio; VE1D; MCS133; PC590; APC8NX*	0.82	0.33	0.07	0.88	0.76	0.31

Al tratarse de pocas moléculas estudiadas, el mejor modelo elegido de la Tabla 5.1 debe cumplir con la relación práctica $N/d = 6$; una relación lineal de 4 descriptores cumple con este requisito y posee además los mejores resultados de validación:

$$\log LC_{50} = -0.56\ GATS4m + 0.15\ nBondsD - 0.17\ SHBint4 - 0.88\ APC8NX + 2.36 \quad (5.1)$$

$N_{cal} = 41$, $R^2_{cal} = 0.72$, $RMS_{cal} = 0.41$

$R^2_{ij\max} = 0.03$, $o2.5 = 0$, $R^2_{aleat} = 0.52$, $RMS^{aleat} = 0.54$ (100000 casos)

$R^2_{loo} = 0.66$, $RMS_{loo} = 0.46$, $N_{val} = 8$, $R^2_{val} = 0.70$, $RMS_{val} = 0.38$

$N_{pred} = 9$, $R^2_{pred} = 0.87$, $RMS_{pred} = 0.31$

La correlación entre los descriptores no es significativa ($R^2_{ij\max} = 0.03$). En las Figuras 5.1 y 5.1.A se grafican los resultados de las predicciones de $\log LC_{50}$ y el gráfico de dispersión de residuos que describe de manera aceptable a los conjuntos de validación y predicción con estos 4 descriptores moleculares. Según sugieren estas figuras, ningún compuesto de calibración posee un residuo mayor a $2.5.S_{cal}$.

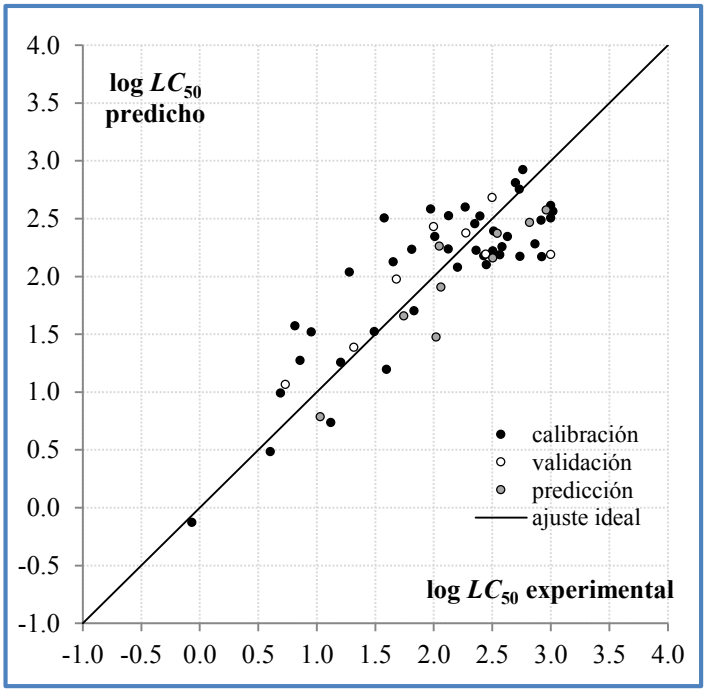


Figura 5.1. Valores de $\log LC_{50}$ *experimentales y predichos según la ecuación (5.1).*

En el presente estudio no es conveniente el empleo de descriptores flexibles, debido al tamaño reducido de los conjuntos de validación y predicción para analizar el desempeño predictivo de tales variables optimizables. Esto se demuestra con el cálculo de 21 descriptores flexibles

basados en atributos individuales, y 37 flexibles basados en atributos múltiples; la Tabla 5.2 resume los mejores modelos encontrados de 1-3 atributos estructurales. Puede apreciarse que un mejor resultado en el conjunto de validación no garantiza un mejor resultado en el conjunto de predicción.

La combinación de alguno de estos descriptores flexibles con descriptores rígidos en una regresión lineal causaría el deterioro de su capacidad predictiva y el sobreajuste del conjunto de calibración, por lo que se excluye a los descriptores flexibles del presente análisis.

Tabla 5.2. Calidad estadística de los mejores descriptores flexibles de la toxicidad aguda en *Eisenia foetida.*

atributos estructurales	R^2_{cal}	RMS_{cal}	R^2_{val}	RMS_{val}	R^2_{pred}	RMS_{pred}
hsg(ec1)	0.47	0.56	0.36	0.60	0.05	0.64
hsg(ec1); hard	0.63	0.47	0.57	0.47	0.07	0.74
hsg(ec1); hard; c5	0.77	0.37	0.65	0.49	0.14	0.76

La relación cuantitativa estructura-actividad representada por la ecuación (5.1) cumple con los parámetros loo y aleatorización-Y. Además, se cumplen las siguientes condiciones:

$1 - R_0^2 / R_{pred}^2 < 0.1\ (0.024)$ o $1 - R_0'^2 / R_{pred}^2 < 0.1\ (3.63.10^{-5})$; $0.85 \le k \le 1.15\ (1.11)$ o $0.85 \le k' \le 1.15\ (0.89)$; $R_m^2 > 0.5\ (0.74)$

Los descriptores no-conformacionales de la ecuación (5.1) se clasifican de la siguiente manera:

a) un descriptor de autocorrelación 2D: *GATS4m*, autocorrelación de Geary-distancia 4 / ponderada por la masa

b) un descriptor constitucional: *nBondsD*, el número de enlaces dobles

c) un descriptor electrotopológico: *SHBint4*, suma de descriptores de fuerza del estado-E para enlaces hidrógeno potenciales de camino de longitud 4

d) un descriptor indicador de pares de átomos: *APC8NX*, la cuenta de enlaces N-X a la distancia topológica 8

Los valores de los descriptores de los pesticidas se incluyen en la Tabla 5.2.A.

Dado que un pesticida es más tóxico en la lombriz *Eisenia foetida* cuanto menor sea su valor de concentración letal media, es posible proponer la siguiente guía QSAR para la búsqueda de estructuras con valores determinados de esta propiedad. Según el signo de los coeficientes de regresión de la ecuación (5.1), cuantos mayores sean simultáneamente los valores numéricos de *GATS4m*, *SHBint4*, *APC8NX* y menor el valor de *nBondsD*, mayor toxicidad aguda presentará el pesticida.

Con respecto al DA de la ecuación (5.1), la Figura 5.2 indica que una única molécula de predicción posee una influencia ligeramente superior al límite: imidacloprid, aunque su predicción puede considerarse confiable.

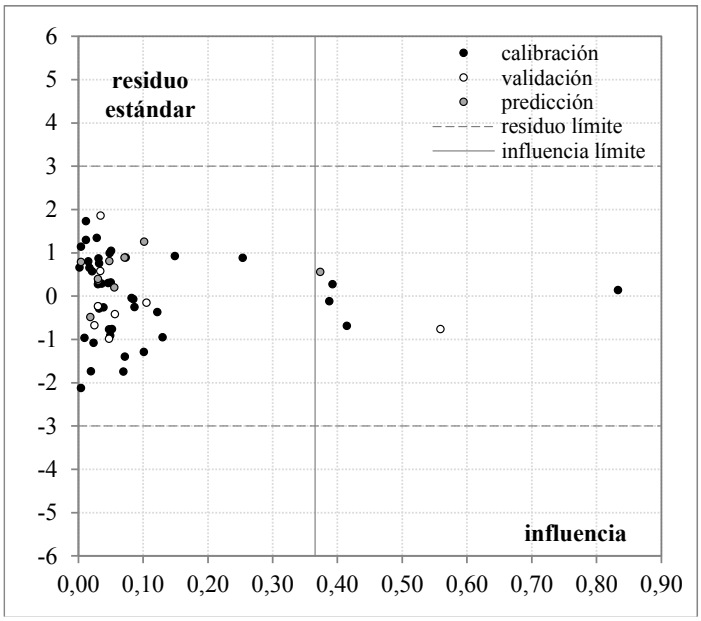


Figura 5.2. Gráfico de Williams para la ecuación (5.1). Influencia límite $h^* = 0.366$.

El siguiente paso del estudio actual consiste en predecir el conjunto de ensayo de 21 plaguicidas con datos no verificados o menos confiables. Según la Tabla 5.1.A, las predicciones de las moléculas **59**-**79** son por lo general próximas a los valores reportados no verificados. Solo un plaguicida está fuera del dominio de aplicación de la ecuación (5.1) (aclonifen), mientras que 2 moléculas poseen predicciones que se diferencian más del valor reportado no verificado (1,2-dicloropropano y tiobencarb).

5.1.3. Conclusiones

A partir de datos verificados y confiables de la base de datos PPDB, se consigue establecer un modelo QSAR predictivo para la toxicidad aguda de 58 pesticidas evaluada en la lombriz *Eisenia foetida*. La mejor relación lineal establecida basada en 4 descriptores no-conformacionales constituye una guía QSAR valiosa para la búsqueda de estructuras con valores determinados de la concentración letal media. Además, la aplicación del modelo para la predicción de 21 plaguicidas con datos no verificados o menos confiables puede considerarse aceptable.

El presente estudio de pesticidas resulta novedoso, pues no se han reportado previamente investigaciones QSAR de toxicidad aguda en la especie *Eisenia foetida*.

Se verifica en el actual trabajo que, debido al número reducido de compuestos empleados para calibrar y validar los datos, no resulta apropiado el uso de descriptores flexibles basados en la optimización numérica de pesos de correlación.

5.2. Estudio QSAR de la toxicidad aguda en ratas

La Regulación Europea No.1272/2008 en clasificación, etiquetado y empaquetado (CLP) requiere una evaluación de la dosis letal media (LD_{50}), especificando que la toxicidad aguda significa aquel efecto adverso que ocurre seguidamente de una administración oral o dermal de una simple dosis de una sustancia o de una mezcla, o múltiples dosis dada dentro de las 24 horas, o una exposición a la inhalación de 4 horas[42].

La dosis letal media hace referencia a la dosis que mata al 50 % de los animales dentro de las 24 horas[43]. Sin embargo, la regulación europea está actualmente promoviendo métodos alternativos a las medidas de toxicidad con el fin de evitar el uso de animales[44]; esto tiene un impacto en las regulaciones específicas, por ejemplo, REACH[45].

Antes de admitir cualquier valor experimental nuevo dentro de REACH, la industria debe cuidadosamente confirmar la información que está disponible, y tomar en cuenta fuentes de datos alternativas.

Saganuwan realizó una revisión en la literatura del uso de LD_{50} en el desarrollo y evaluación de fármacos y compuestos químicos. Se pudo determinar que, en el pasado, muchos animales se habían utilizado para la determinación de LD_{50}. La OCDE ha reducido el número de animales de prueba de 5-15 y actualmente se ha reducido aún más de 2-6.

Aunque la aplicación de LD_{50} se ha reducido drásticamente, todavía se aplica y se acepta en algunas partes del mundo. Además, se debe permitir que los animales en los que se realizan las pruebas LD_{50} mueran para ver el efecto final del fármaco o producto químico que se está evaluando, ya que la eutanasia de los animales de ensayo puede enmascarar algunos signos de toxicidad de los agentes de prueba. Por lo tanto, el estudio de la toxicidad de drogas y productos químicos es un proceso científico necesario para el descubrimiento y desarrollo de fármacos, así como para la identificación de tóxicos potenciales[46].

Entre los métodos alternativos, la Regulación REACH menciona los modelos QSAR. Para conocer los requerimientos regulatorios para la reducción de pruebas con animales, y la evaluación de LD_{50}, se examinaron un número de métodos QSAR para estimar el valor de LD_{50} [47–50].

Cronin y Dearden[51] realizaron una revisión crítica de los modelos QSAR aplicados a la toxicidad en mamíferos. Ellos encontraron que las predicciones en base a modelos QSAR han sido desarrolladas en forma escasa, y han determinado que la hidrofobicidad correlaciona bien con dicha actividad estudiada. Además, consideran que cuando se tiene en cuenta el modo de acción, la toxicidad puede ser útil para generar una fuente de información valiosa y aplicar a diferentes tipos de especies que se emplean para realizar las pruebas.

Un estudio QSAR sobre la toxicidad aguda en mamíferos sobre un conjunto de plaguicidas organofosforados es presentado por Eldred y Jurs[52]. Para encontrar dichos modelos utilizaron diferente tipo de descriptores moleculares, basados en las características topológicas, electrónicas y geométricas. La selección de variables se realizó mediante algoritmos genéticos para encontrar el mejor subconjunto de descriptores que respaldarían un modelo de red neuronal computacional de alta calidad (CNN), y que vincula la estructura molecular con los valores de $-\log\left(\text{mmol.kg}^{-1}\right)$ de los compuestos bajo estudio. Los mejores siete descriptores del modelo CNN no lineal encontrado, conducen a valores de *RMS*=0.22 unidades logarítmicas para el conjunto de calibración y *RMS*=0.25 unidades logarítmicas para el conjunto de predicción.

Un trabajo propuesto por García-Domenech[53] predice la toxicidad en agua en mamíferos para compuestos pesticidas organofosforados usando descriptores topológicos. En el mismo, se emplearon modelos con 6 variables para la predicción de LD_{50}-intraperitoneal ($R = 0.85$, $Q^2 = 0.61$) y se seleccionaron 8 variables para LD_{50}-oral ($R = 0.91$, $Q^2 = 0.70$). Para realizar la validación cruzada y analizar el conjunto de predicción externo, se aplicó la metodología LMO, esto permitió evaluar la estabilidad y el rendimiento de predicción de los modelos topológicos seleccionados.

Un modelo de relación cuantitativa estructura-toxicidad ha sido propuesto por Devillers[54] para estimar la toxicidad agua en ratas machos y hembras por plaguicidas organofosforados. Los conjuntos fueron particionados en 51 compuestos en el conjunto de calibración y 9 compuestos en el conjunto de predicción. Los modelos seleccionados se describieron a partir de vectores de autocorrelación que codifican la lipofilicidad, la refractividad molar, la capacidad aceptora de uniones (*HBA*) y la capacidad dadora de uniones (*HBD*) de las moléculas. Se emplearon dos técnicas para seleccionar las mejores características mediante PLS y red neuronal artificial, con el fin de que los modelos tengan en cuenta el sexo de los organismos a la hora de estimar la toxicidad de los plaguicidas.

Los mejores resultados se obtuvieron con un modelo de ANN de 8/4/1 calibrado con los algoritmos de contra propagación y descenso de gradiente conjugado. Los valores de *RMS* para el conjunto de calibración y el conjunto de predicción externo fueron de 0.29 y 0.26, respectivamente.

En otro estudio[55], se compiló un conjunto de datos de 7385 compuestos con muchos valores conservadores de LD_{50}. Se empleó un enfoque QSAR combinatorio para desarrollar modelos sólidos y predictivos de toxicidad aguda en ratas causada por la exposición oral a productos químicos. Para permitir una comparación equitativa entre el poder predictivo de los modelos generados en este estudio con un programa comercial para predecir la toxicidad TOPKAT, se seleccionó un subconjunto de datos de todo el conjunto completo que incluía 3472 compuestos utilizados en el conjunto de calibración de TOPKAT. Los 3913 compuestos restantes, que no estaban presentes en el conjunto de calibración TOPKAT, se usaron como el conjunto de predicción externo.

Se desarrollaron modelos QSAR de 5 tipos diferentes para el conjunto de calibración. El uso del umbral de dominio de aplicabilidad implementado en la mayoría de los modelos, mejoró la precisión de predicción externa, pero se esperaba que condujera a la disminución en la cobertura del espacio químico; dependiendo del umbral de dominio de aplicabilidad, R^2 varió de 0.24 a 0.70. Finalmente, se desarrollaron varios modelos de consenso promediando el valor predicho de LD_{50} para cada compuesto usando los cinco modelos. Los modelos de consenso proporcionaron una mayor precisión de predicción para el conjunto de datos de validación externa con la cobertura más alta en comparación con los modelos individuales. Los modelos LD_{50} de consenso validados y desarrollados en este estudio pueden usarse como herramientas computacionales confiables de la toxicidad aguda *in vivo*.

Finalmente, Hamadache *et al.*[15]. desarrollaron un modelo QSAR validado para predecir la toxicidad oral aguda de 329 plaguicidas en ratas. Este modelo QSAR se basa en 17 descriptores moleculares que se obtuvieron con un modelo ANN 17/9/1 calibrado con el algoritmo de contra propagación

Quasi Newton (BFGS). La eficacia de predicción para el conjunto de validación externo se estimó mediante Q^2_{ext} y *RMS* que son iguales a 0.95 y 0.20, respectivamente. El 98.6% del conjunto de validación externo se predijo correctamente.

5.2.1. Datos experimentales de T.E.S.T. (7413 moléculas)

Los valores de LD_{50} aguda oral en ratas representa la cantidad del producto químico (masa del producto químico por peso corporal de la rata) que cuando se ingiere por vía oral mata al 50% de las ratas. El conjunto de datos para este punto final se obtuvo de la base de datos del programa de estimación de toxicidad de libre acceso T.E.S.T versión 4.0[56].

La lista de compuestos químicos de dicha base de datos fue filtrada previamente utilizando los siguientes criterios: i. solo se usaron compuestos químicos con valores discretos de LD_{50} (es decir, los valores de LD_{50} con ">" o "<" fueron eliminados); ii. los compuestos solo pueden contener los siguientes símbolos: C, H, O, N, F, Cl, Br, I, S, P, Si o As; iii. los compuestos deben representar un solo componente puro (es decir, sales, mezclas isoméricas indefinidas, polímeros o mezclas fueron eliminados).

En la versión de T.E.S.T. 4.0 y posteriores, todos los isómeros se mantuvieron porque la presencia de isómeros tuvo un impacto insignificante en la estadística del conjunto de predicción externo. El conjunto final de datos de LD_{50} oral en rata contiene 7413 sustancias químicas. Los valores del punto final modelado se convierten en $-\log LD_{50}\,[\mathrm{mol\,kg^{-1}}]$. El intervalo de valores de toxicidad aguda oral en ratas se sitúa entre 0.291 y 7.207. En la Tabla 5.3.A se incluyen los detalles de los compuestos estudiados.

5.2.2. Metodología QSAR

Se dibujan las estructuras químicas con ACD ChemSketch[21] en el formato molecular MDL mol (V2000). La conversión de las moléculas se realizó con el programa para Windows Open Babel[22].

Los 23690 descriptores moleculares se calcularon mediante los programas PaDEL (14464), Mold2 (777), QuBiLs-MAS (8448) y el mejor descriptor flexible mediante el programa CORAL. Se obtienen 23690 descriptores moleculares.

El análisis sobre el conjunto de calibración de la dependencia lineal, exclusión de descriptores de valores únicos, y descriptores vacíos, conduce a $D = 12525$ variables estructurales.

El objetivo principal del trabajo actual es comparar el desempeño predictivo del modelo QSAR desarrollado aquí con aquel reportado en el programa T.E.S.T. Para ello, se utiliza el mismo conjunto de predicción en ambos casos, conformado por 1482 compuestos químicos heterogéneos, y los modelos se calibran a través de las restantes 5931 moléculas. La técnica BSM conduce a un conjunto de calibración de 2964 moléculas y un conjunto de validación de 2967 moléculas.

Se aplica la técnica RM para hallar los descriptores moleculares más representativos de la toxicidad aguda. Las mejores regresiones lineales QSAR de 1-9 descriptores conducen a los resultados mostrados en la Tabla 5.3.

Tabla 5.3. Los mejores descriptores para el estudio de la toxicidad oral en ratas.

d	Descriptores	R^2_{cal}	RMS_{cal}	R^2_{val}	RMS_{val}	R^2_{pred}	RMS_{pred}
1	*DCW*	0.51	0.67	0.41	0.72	0.41	0.74
2	*KR2683; DCW*	0.52	0.66	0.42	0.72	0.44	0.73
3	*KR2683; KR3632; DCW*	0.52	0.66	0.43	0.71	0.44	0.72
4	*KR2683; KR3632; D389; DCW*	0.53	0.66	0.44	0.71	0.45	0.72
5	*KR2683; KR4659; D389; Qub1; DCW*	0.54	0.65	0.44	0.71	0.46	0.71
6	*PC35; KR2683; KR4659; D389; Qub2; DCW*	0.54	0.65	0.44	0.70	0.46	0.71

7	*Sub295; KR2683; KR4659; APC3CS; D389; Qub3; DCW*	0.55	0.64	0.44	0.70	0.47	0.71
8	*PC31; Sub295; KR2683; KR4659; D389; Qub4; Qub3; DCW*	0.55	0.64	0.45	0.70	0.46	0.71
9	*MCS69; PC35; Sub295; KR2683; KR4659; D389; D697; Qub2; DCW*	0.56	0.64	0.45	0.70	0.47	0.71

Se aprecia de la Tabla 5.3 que los modelos encontrados en esta base de datos de alta heterogeneidad estructural no tienen buena calidad estadística, pues el parámetro R^2_{cal} apenas logra superar el límite generalmente aceptado de 0.5. Si bien se ha analizado una gran cantidad de descriptores rígidos junto al mejor descriptor flexible para establecer la relación estructura-propiedad, no se consigue un buen ajuste lineal en las 2964 moléculas de calibración. Además, se observa de la Tabla 5.3 que el incremento del número de descriptores presentes en el modelo no conduce a una mejora significativa de RMS en los tres conjuntos moleculares, sino que este parámetro varía de manera homogénea con el aumento de la dimensión del modelo.

Si se elige el modelo de 5 descriptores para predecir la toxicidad aguda,

$$-\log LD_{50} = 0.70\ KR2683 + 0.41\ KR4659 - 0.031\ D389 + 0.20\ Qub1 + 0.048\ DCW + 1.54 \qquad (5.2)$$

$N_{cal} = 2964$, $R^2_{cal} = 0.54$, $RMS_{cal} = 0.65$
$R^2_{ij\max} = 0.061$, $o3 = 30$, $R^2_{aleat} = 0.011$, $RMS^{aleat} = 0.95$ (100000 casos)
$R^2_{loo} = 0.54$, $RMS_{loo} = 0.65$, $R^2_{l30\%o} = 0.53$, $RMS_{l30\%o} = 0.66$ (100000 casos)
$N_{val} = 2967$, $R^2_{val} = 0.44$, $RMS_{val} = 0.71$
$N_{pred} = 1482$, $R^2_{pred} = 0.46$, $RMS_{pred} = 0.71$

El gráfico de las predicciones en función de los valores experimentales de $-\log LD_{50}$ se muestra en la Figura 5.3, mientras que la dispersión de residuos aparece en la Figura 5.2.A.

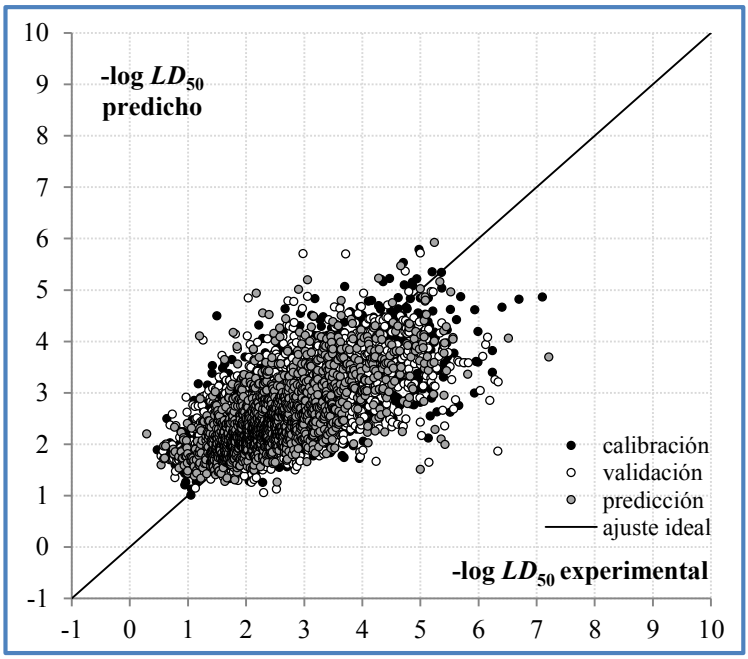


Figura 5.3. Valores de $-\log LD_{50}$ *experimentales y predichos según la ecuación (5.2).*

La relación cuantitativa estructura-propiedad representada por la ecuación (5.2) cumple con los parámetros loo, lmo(30%) y aleatorización-Y. También se cumplen las condiciones:

$1-R_0^2/R_{pred}^2 < 0.1\ (0.008)$; $0.85 \le k \le 1.15\ (0.998)$ o $0.85 \le k' \le 1.15\ (0.934)$, aunque no se cumple con $R_m^2 > 0.5\ (0.43)$

El criterio MAE indica que la capacidad predictiva del modelo QSAR de 5 descriptores sobre las 1482 moléculas de predicción es 'moderada': para el conjunto de predicción $MAE(100\%) = 0.53$ y $\sigma(100\%) = 0.47$, mientras que si se omite el 5% de los compuestos con altos valores de residuos conduce a $MAE(95\%) = 0.46$ y $\sigma(95\%) = 0.34$.

Los 5 descriptores del modelo son prácticamente ortogonales entre sí:

a) dos descriptores indicadores Klekota-Roth: *KR2683*, presencia del fragmento OC(=O)[NH][CH3] y *KR4659*, presencia del fragmento OC(=O)Cc1ccccc1.

b) un descriptor topológico: *D389*, suma de distancias topológicas entre los vértices P y Cl

c) un descriptor topológico de QuBiLS: *Qub1*, índice algebraico obtenido de la matriz de densidad electrónica del pseudografo molecular y que considera pesos atómicos.

d) un descriptor flexible: *DCW*, basado en *hsg-pt2* y SMILES-*sss*

Los valores numéricos de estos descriptores se incluyen en la Tabla 5.4.A.

A la hora de comparar la calidad de las predicciones de $-\log LD_{50}$ según la ecuación (5.2), con las halladas mediante el modelo QSAR de consenso del programa T.E.S.T., este último conduce a los mejores resultados según se observa de las Figuras 5.4 y 5.3.A. Los 1482 compuestos químicos del conjunto de predicción se predicen mejor con T.E.S.T.: $R^2_{pred} = 0.62$ y $RMS_{pred} = 0.59$; mientras que con la ecuación (5.2): $R^2_{pred} = 0.46$ y $RMS_{pred} = 0.71$.

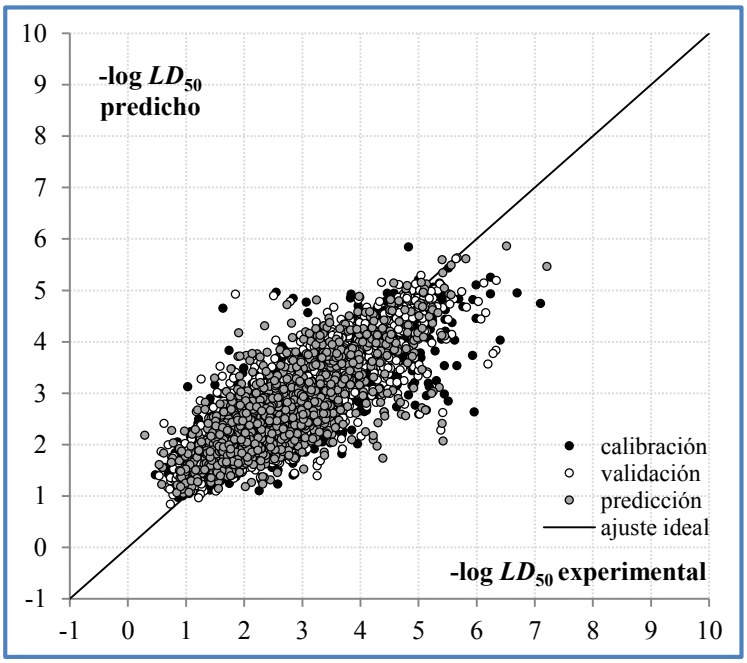


Figura 5.4. Valores de $-\log LD_{50}$ experimentales y predichos según el modelo QSAR de consenso de T.E.S.T.

5.2.3. Conclusiones

El trabajo desarrollado consistió en la búsqueda de un modelo QSAR para el estudio de la toxicidad aguda oral en ratas, sobre 7413 compuestos químicos heterogéneos y pesticidas. Para este fin se analizaron 12525 descriptores moleculares de naturaleza rígida y el mejor flexible, aunque los

modelos de regresión lineal encontrados no alcanzaron una calidad estadística aceptable. El mejor modelo de consenso propuesto en el programa T.E.S.T. conduce a las mejores predicciones de la dosis letal media.

La principal ventaja de ambas metodologías, sea la propuesta aquí o la implementada en T.E.S.T, consiste en el empleo de descriptores moleculares independientes de la conformación y la utilización de programas de cálculo de descriptores de libre acceso.

El estudio actual resulta ser una primera etapa de investigación, para poder continuar con la búsqueda de modelos predictivos de la toxicidad aguda en ratas en el presente conjunto de datos de alta diversidad estructural. Una manera de mejorar los resultados encontrados es a través de la incorporación de nuevas definiciones de descriptores moleculares; utilización de esquemas de consenso para la predicción de la propiedad; empleo de interpolaciones lineales de datos; u otras herramientas alternativas para el análisis de los datos.

Bibliografía

1. Wan, M. T. (2013). Ecological risk of pesticide residues in the British Columbia environment. *J. Environ. Sci. Heal.* Part B 48, 344–363
2. Müller, K., Tiktak, A., Dijkman, T. J., Green, S. & Clothier, B. Advances in pesticide risk reduction. *Encycl. Agric. Food Syst. van Alfen, ed.). Acad. Press. USA* 17–34 (2014).
3. Gopi, R. A., Satyavani, G., Shanmugasundaram, R. & Murthy, B. P. Acute Toxicity Evaluation of Expired Pesticides on Earthworms Eisenia fetida. *Int. J. Environ. Sci.* **4,** 1121 (2014).
4. Price, N. R. & Watkins, R. W. Quantitative structure-activity relationships (QSAR) in predicting the environmental safety of pesticides. *Pestic. Outlook* **14,** 127–129 (2003).
5. Gao, M., Song, W., Zhang, J. & Guo, J. Effect on enzymes and histopathology in earthworm (Eisenia foetida) induced by triazole fungicides. *Environ. Toxicol. Pharmacol.* **35,** 427–433 (2013).
6. Pelosi, C., Joimel, S. & Makowski, D. Searching for a more sensitive earthworm species to be used in pesticide homologation tests–A meta-analysis. *Chemosphere* **90,** 895–900 (2013).
7. Zou, Xiaoming; Xiao, Xiaoyu; Zhou, Hanfeng; Chen, Feng; Zeng, Jianjun; Wang, Wenbiao;; Feng, Guangping; Huang, X. Effects of soil acidification on the toxicity of organophosphorus pesticide on Eisenia fetida and its mechanism. *J. Hazard. Mater.* (2018).

8. Lanno, R., Wells, J., Conder, J., Bradham, K. & Basta, N. The bioavailability of chemicals in soil for earthworms. *Ecotoxicol. Environ. Saf.* **57,** 39–47 (2004).

9. Belfroid, A. C., Sijm, D. & Van Gestel, C. A. M. Effect of aging of chemicals in soil on their biodegradability and extractability. *Environ. Rev* **4,** 276–299 (1996).

10. Alexander, M. Aging, bioavailability, and overestimation of risk from environmental pollutants. *Environ. Sci. Technol.* **34,** 4259–4265 (2000).

11. Edwards, C. A. & Bohlen, P. J. The effects of toxic chemicals on earthworms. in *Reviews of environmental contamination and toxicology* 23–99 (Springer, 1992).

12. Golbamaki, A. *et al.* Comparison of in silico models for prediction of Daphnia magna acute toxicity. *SAR QSAR Environ. Res.* **25,** 673–694 (2014).

13. Sullivan, K. M., Manuppello, J. R. & Willett, C. E. Building on a solid foundation: SAR and QSAR as a fundamental strategy to reduce animal testing. *SAR QSAR Environ. Res.* **25,** 357–365 (2014).

14. Cheng, F., Li, W., Liu, G. & Tang, Y. In silico ADMET prediction: recent advances, current challenges and future trends. *Curr. Top. Med. Chem.* **13,** 1273–1289 (2013).

15. Hamadache, M. *et al.* A quantitative structure activity relationship for acute oral toxicity of pesticides on rats: Validation, domain of application and prediction. *J. Hazard. Mater.* **303,** 28–40 (2016).

16. Hamadache, M., Benkortbi, O., Hanini, S. & Amrane, A. QSAR modeling in ecotoxicological risk assessment: application to the prediction of acute contact toxicity of pesticides on bees (Apis mellifera L.). *Environ. Sci. Pollut. Res.* **25,** 896–907 (2018).

17. Martin, T. M., Lilavois, C. R. & Barron, M. G. Prediction of pesticide acute toxicity using two-dimensional chemical descriptors and target species classification. *SAR QSAR Environ. Res.* **28,** 525–539 (2017).

18. Mazzatorta, P., Smiesko, M., Lo Piparo, E. & Benfenati, E. QSAR model for predicting pesticide aquatic toxicity. *J. Chem. Inf. Model.* **45,** 1767–1774 (2005).

19. Toropov, A. A. *et al.* QSAR models for predicting acute toxicity of pesticides in rainbow trout using the CORAL software and EFSA's OpenFoodTox database. *Environ. Toxicol. Pharmacol.* **53,** 158–163 (2017).

20. Agriculture & Environment Research Unit (AERU) at the University of Hertfordshire. PPDB: Pesticide Properties DataBase. (2015). Available at: https://sitem.herts.ac.uk/aeru/footprint/es/.

21. ACD/ChemSketch. Available online: http://www.acdlabs.com (accessed on 29 July 2016). (2016).

22. Open Babel for Windows. Available at https://openbabel.org wiki/Category:Installation, 2017.

23. PaDEL. Available online: http://www.yapcwsoft.com/ (accessed on 29 July 2016). (2016).

24. Yap, C. W. PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. *J. Comput. Chem.* **32,** 1466–1474 (2011).

25. RECON, version 5. 5/5. 3. Curt M. Breneman, William P. Katt, Martin Martinov, Marlon O. Rhem, N. Sukumar, Tracy R. Thompson, Christopher Whitehead and Dechuan Zhuang. (Rensselaer Polytechnic Institute, 2003).

26. B.K. Lavine, C.E. Davidson, C. Breneman, and W. Katt. Electronic van der Waals surface property Databases, descriptors and genetic algorithms for developing structure-activity correlations in olfactory. *J. Chem. Inf. Comput. Sci.* **43,** 1890–1905 (2003).

27. R.F.W. Bader. Atoms in Molecules-A Quantum Theory. *Oxford Univ. Press. Oxford, UK* (1990).

28. C.M. Breneman and M. Rhem. A QSPR analysis of HPLC column capacity factors for a set of high-energy Transferable, materials using electronic Van der Waals surface property descriptors computed by the atom equivalent method. *J. Comput. Chem.* **18,** 182–197 (1997).

29. H. Hong, Q. Xie, W. Ge, F. Qian, H. Fang, L. Shi, Z. Su, R. Perkins, and W. Tong, J. Mold2, Molecular descriptors from 2D structures for chemoinformatics and toxicoinformatics. *Chem. Inf. Model.* **48,** 1337–1344 (2008).

30. Thomas Sander. Idorsia Pharmaceuticals Ltd. OSIRIS DataWarrior. Versión 4.7.3. (2014).

31. J.R. Valdes-Martini, C.R. García Jacas, Y. Marrero-Ponce, Y. Silveira Vaz 'd Almeida, and C. M., CAMD-BIR Unit, CENDA Number of register: 2373-2012, 2012. & 763, D. by [190. 191. 137. 201. at 16:51 12 O. 2017. QuBiLS-MAS: Free Software for molecular descriptors calculator from Quadratic, Bilinear and Linear Maps based on Graph–Theoretic Electronic-Density Matrices and Atomic weighting, Version 1.0. *SAR QSAR Environ. Res.*

32. Duchowicz, P.R.; Castro, E.A.; Fernández, F. M. Alternative algorithm for the search of an optimal set of descriptors in QSAR-QSPR studies. *MATCH Commun. Math. Comput. Chem.* **55,** 179–192 (2006).

33. Duchowicz, P.R.; Castro, E.A.; Fernández, F.M.; González, M. A new search algorithm of QSPR/QSAR theories: Normal boiling points of some organic molecules. *Chem. Phys. Lett.* **412,** 376–380 (2005).

34. Matlab 7.0. Available online: http://www.mathworks.com (accessed on 29 July 2016).

35. P. Gramatica and A. Sangion. A historical excursus on the statistical validation parameters for QSAR models: A clarifcation concerning metrics and terminology. *J. Chem. Inf. Model.* . **56,** 1127–1131 (2016).

36. K. Roy, R.N. Das, P. Ambure, and R.B. Aher. Be aware of error measures. Further studies on validation of predictive QSAR models. *Chemom. Intel. Lab. Syst.* **15,** 18–33 (2016).

37. Golbraikh, A.; Tropsha, A. Beware of q2! *J. Mol. Graph. Model.* **20,** 269–276 (2002).

38. Wold, S. Eriksson, L. Statistical validation of qsar results. In Chemometrics Methods in Molecular Design; van de Waterbeemd, H., Ed.; VCH: Weinheim, Germany. 309–318 (1995).

39. Gramatica, P. Principles of qsar models validation: Internal and external. *QSAR Comb. Sci.* **26,** 694–701 (2007).

40. K. Roy, S. Kar, and P. A. On a simple approach for determining applicability domain of QSAR models. *Chemom. Intel. Lab. Syst.* **145,** 22–29 (2015).

41. Eriksson, L.; Jaworska, J.; Worth, A.P.; Cronin, M.T.; McDowell, R.M.; Gramatica, P. Methods for reliability QSARS and uncertainty assessment and for applicability evaluations of classification- and regression-based. *Environ. Heal. Perspect.* **111,** 1361–1375 (2003).

42. European parliament and of the council 16, December 2008 on classification, labelling and packaging of substances and, mixtures, amending and repealing Directives 67/548/EEC and 1999/45/EC, A., Considerations, amending R. (EC) N. 1907/2006 – A. I. 3. 1. 2. . "Specific & Toxic", for classification of substances as acutely. Regulation (EC) No 1272/2008.

43. J.W. Trevan. The Error of Determination of Toxicity. *Proc. R. Soc. Lond. B.* **101,** 483–514 (1927).

44. COUNCIL REGULATION (EC) No 440/2008 of 30 May 2008 laying down test. Methods pursuant to Regulation (EC) No 1907/2006 of the European Parliament of the Council on the Registration, Evaluation, Authorisation and Restriction of (REACH), Chemicals.

45. REGULATION (EC) No 1907/2006 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 18 December 2006 concerning the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH), establishing a European Chemicals Agency, amending Directive 1999/4.

46. Saganuwan, S. A. Toxicity studies of drugs and chemicals in animals: an overview. *Bulg. J. Vet. Med.* **20,** (2017).

47. J. Devillers. Prediction of mammalian toxicity of organophosphorus pesticides from QSTR modeling. *SAR QSAR Environ. Res.* **15,** 501–510 (2004).

48. R. García-Domenech, P. Alarcón-Elbal, G. Bolas, R. Bueno-Marí, F. A. C.- & Olmos, S.A. Delacour, M.C. Mouriño, A. Vidal, and J. Gálvez. Prediction of acute toxicity of organophosphorus pesticides using topological indices. *SAR QSAR Environ. Res.* **18,** 745–755 (2007).

49. T. Martin. Toxicity Estimation Software Tool (TEST). *Softw. available Http//www.epa.gov/nrmrl/std/qsar/qsar.html*

50. TerraBase Inc., Hamilton, C. TerraQSAR. *Softw. available Http//www.terrabase-inc.com/*

51. Cronin, M. T. D. & Dearden, J. C. QSAR in toxicology. 2. Prediction of acute mammalian toxicity and interspecies correlations. *Quant. Struct. Relationships* **14,** 117–120 (1995).

52. Eldred, D. V & Jurs, P. C. Prediction of acute mammalian toxicity of organophosphorus pesticide compounds from molecular structure. *SAR QSAR Environ. Res.* **10,** 75–99 (1999).

53. García-Domenech, R. *et al.* Prediction of acute toxicity of organophosphorus pesticides using topological indices. *SAR QSAR Environ. Res.* **18,** 745–755 (2007).

54. Devillers, J. Prediction of mammalian toxicity of organophosphorus pesticides from QSTR modeling. *SAR QSAR Environ. Res.* **15,** 501–510 (2004).

55. Zhu, H. *et al.* Quantitative structure− activity relationship modeling of rat acute toxicity by oral exposure. *Chem. Res. Toxicol.* **22,** 1913–1921 (2009).

56. U.S. EPA. User's Guide for T.E.S.T. (version 4.2) (Toxicity Estimation Software Tool): A Program to Estimate Toxicity from Molecular Structure. (2016).

Conclusiones generales y proyecciones futuras

En la presente tesis doctoral se ha estudiado y aplicado la teoría QSPR-QSAR para generar relaciones cuantitativas que permitan predecir actividades y propiedades de interés agronómico.

A lo largo de la tesis el desarrollo de modelos predictivos ha seguido el protocolo recomendado por la Organización para la Cooperación Económica y el Desarrollo (OECD), el cual se basa en cinco principios básicos: actividad/propiedad/toxicidad definida, uso de un algoritmo inequívoco, definición del dominio de aplicabilidad del modelo, medida apropiada de la calidad del modelo (ajuste, robustez y predictividad) e interpretación del mecanismo de los descriptores del modelo, y que ayuden a tomar decisiones en trabajos posteriores.

De acuerdo a los objetivos planteados en la el trabajo de tesis, se han aplicado los mejores descriptores moleculares rígidos y flexibles, que contemplen los aspectos constitucionales, topológicos, electrónicos y lipofílicos de las moléculas investigadas, y que se calculan con programas computacionales diseñados de acceso libre, tales como PaDEL, Mold^2, RECON, EPI Suite, QuBiLs-MAS, CORAL y Datawarrior.

Los modelos lineales fueron obtenidos principalmente mediante regresión lineal multivariable a partir del Método del Reemplazo, con el fin de desarrollar relaciones que nos permitan establecer una interpretación directa de la estructura molecular estudiada con énfasis en los pesticidas.

En cada modelo de regresión se ha definido el dominio de aplicabilidad mediante el criterio del valor de influencia, y en la técnica basada en estandarización. Finalmente, la interpretación del mecanismo de los descriptores moleculares empleados en nuestros modelos se realizó en base a los coeficientes de regresión y su estandarización.

La preparación y pretratamiento de los datos ha sido un aspecto fundamental para garantizar la confiabilidad en la construcción de los modelos QSPR-QSAR. Para este propósito aplicamos una opción que viene integrada al programa CORAL, que permite la identificación de compuestos con la misma codificación SMILES, o a partir de su número de CAS.

Hemos explorado y aplicado el concepto de descriptores óptimos o flexibles de CORAL, el cual basa su cálculo en la representación de la estructura molecular a partir de grafos o SMILES. Genera una combinación lineal de los mejores atributos estructurales propios de la molécula, según la propiedad/actividad analizada. En este sentido, la búsqueda de la mejor combinación de atributos se realizó empleando la técnica del Método de Inclusión de a Pasos, mediante la inclusión de un atributo estructural a la vez para alcanzar los mejores resultados.

El trabajo realizado logró establecer modelos matemáticos de importancia agronómica a partir de propiedades fisicoquímicas, tales como el coeficiente de sorción en suelo, el factor de bioconcentración, la solubilidad acuosa, la constante de la ley de Henry, y finalmente, a través de toxicidad aguda en lombrices y en ratas. La predicción de este tipo de propiedades permitiría conocer de antemano su valor antes de evaluar experimentalmente los compuestos químicos, y resulta de gran importancia tanto por los efectos ambientales, como por la toxicidad en animales y en el hombre.

La obtención de modelos QSPR-QSAR, a partir de un enfoque de representación estructural independiente de la conformación, es válido y útil cuando se trabaja con bases de datos de interés en propiedades de pesticidas. Este enfoque permitió explorar grandes bases de datos con compuestos heterogéneos que incluían pesticidas, sin la necesidad de considerar los aspectos tridimensionales de las moléculas. La combinación de descriptores de naturaleza rígida y flexible basados en dicho enfoque no-conformacional ha permitido obtener resultados satisfactorios en la mayoría de las propiedades de interés agronómico estudiadas.

Existen aún muchos temas adicionales por investigar en el contexto del trabajo de tesis doctoral:

1) empleo de base de datos más específicas para las diferentes clases de compuestos pesticidas que se encuentran actualmente en el mercado: insecticidas, fungicidas, y herbicidas;

2) análisis de los residuos de plaguicidas en diferentes alimentos, para poder establecer modelos que sirvan como herramienta para su estimación;

3) establecer modelos QSPR en base a otras propiedades de interés de pesticidas, tales como la presión de vapor, el punto de ebullición o el coeficiente de reparto octanol/agua;

4) desarrollar modelos QSAR a partir de nuevas actividades de interés en pesticidas, en otras especies animales como *Trucha arcoiris*, *Daphnia magna*, toxicidad dietaria en codorniz, y toxicidad en abejas;

5) aplicar las estimaciones de los modelos QSPR/QSAR en pesticidas empleados en diferentes áreas y cultivos de Argentina, junto con las condiciones ambientales y de manejo de cada cultivo, con el fin de proponer estimaciones de contaminación ambiental de dichos compuestos en los ecosistemas agrícolas;

6) uso de técnicas quimiométricas alternativas para el análisis de los datos, por ejemplo, regresión de mínimos cuadrados parciales, redes neuronales artificiales, y también ensayar el uso de predicciones consensuadas de la propiedad y predicciones basadas en interpolaciones lineales;

7) aplicación de nuevos programas de acceso libre para el cálculo de los descriptores moleculares, para lograr una mejor representación estructural.

Publicaciones y trabajos presentados en eventos científicos

El presente trabajo de tesis doctoral dio lugar a las siguientes publicaciones:

1. Aranda, J. F., Duchowicz, P. R. y Castro, E. A. (2014). Estudio, desarrollo y aplicación de modelos de la Teoría QSPR-QSAR para la correlación de propiedades de interés agronómico. Revista Electrónica Investigación Jóven de la UNLP, 1, 56-56. ISSN 2314-3991.
2. Aranda, J. F., Duchowicz, P. R. y Castro, E. A. (2016). Avances en estudios QSAR/QSPR con aplicaciones agronómicas. Revista Electrónica Investigación Jóven de la UNLP, 3, 85-85. ISSN 2314-3991.
3. Aranda, J. F., Garro Martinez, J. C., Castro, E. A. & Duchowicz, P. R. (2016). Conformation-independent QSPR approach for the soil sorption coefficient of heterogeneous compounds. International Journal of Molecular Sciences, 17(8), 1247-1255.
4. Aranda, J. F., Bacelo, D. E., Leguizamón Aparicio, M. S., Ocsachoque, M. A., Castro, E. A. & Duchowicz, P. R. (2017). Predicting the bioconcentration factor through a conformation-independent QSPR study. SAR and QSAR in Environmental Research, 28(9), 749-763.
5. Fioressi, S. E., Bacelo, D. E., Rojas, C., Aranda, J. F. & Duchowicz, P. R. (2019). Conformation-Independent QSPR Study on Water Solubility of Pesticides. Ecotoxicology and Environmental Safety, 171, 47-53.

Los siguientes trabajos fueron publicados en congresos:

1. Estudio, desarrollo y aplicación de modelos de la Teoría QSPR-QSAR para la correlación de propiedades de interés agronómico, J. F. Aranda, Jornadas de becarios del INIFTA, 14-17 de octubre 2014, La Plata. Participante como expositor.

2. Predicciones QSPR en coeficientes de sorción en suelo. J. F. Aranda, P. R. Duchowicz y E. A. Castro. XXX Congreso Argentino de Química,

Asociación Química Argentina, Sánchez de Bustamante 1749, Ciudad Autónoma de Buenos Aires, 22-24 de Octubre de 2014. Presentación de póster.

3. Avances en estudios QSAR/QSPR con aplicaciones agronómicas, J. F. Aranda, Jornadas de becarios del INIFTA, 13-16 de octubre 2015, La Plata. Participante como expositor.

4. Estudio, desarrollo y aplicación de modelos de la teoría QSPR-QSAR para la correlación de propiedades de interés agronómico. J. F. Aranda, P. R. Duchowicz y E. A. Castro. Primeras Jornadas de Tesistas de la Facultad de Ciencias Exactas, 28, 29 y 30 Octubre 2015, Universidad Nacional de La Plata. Presentación de póster.