UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO
FACULTAD DE QUÍMICA

PESTIMEP, (PESTiside Multiple EndPoint) una base de datos de pesticidas evaluados en múltiples ensayos toxicológicos

T E S I S

QUE PARA OBTENER EL TÍTULO DE

INGENIERO QUÍMICO

PRESENTA

Estibalis Arni Daniel Chávez Gómez

CIUDAD UNIVERSITARIA, CD. MX. 2018

JURADO ASIGNADO:

PRESIDENTE:	**Profesor: Francisco Hernández Luis**
VOCAL:	**Profesor: María Rafaela Gutiérrez Lara**
SECRETARIO:	**Profesor: Karina Martínez Mayorga**
1er. SUPLENTE:	**Profesor: José Luis Medina Franco**
2° SUPLENTE:	**Profesor: Abraham Madariaga Mazón**

SITIO DONDE SE DESARROLLÓ EL TEMA: INSTITUTO DE QUÍMICA, U.N.A.M.

ASESOR DEL TEMA:

Dra. Karina Martínez Mayorga

SUPERVISOR TÉCNICO:

Dr. Abraham Madariaga Mazón

SUSTENTANTE (S):

Estibalis Arni Daniel Chávez Gómez

Agradecimiento:

A CONACyT por haber proporcionado recursos económicos para el desarrollo del presente trabajo de tesis a través del proyecto 220392.

A los Laboratorios Senosiain por proporcionar recursos económicos para la elaboración del trabajo de investigación y tesis.

Índice de contenido

Índice de Figuras

Índice de Tablas

Glosario

Apéndice A - Abreviaturas de las especies en la base de datos (Milne 1995)

Abreviación	Extensión	Traducción
bwd	wild bird species	especies salvajes de aves
cat	cat	gato
chd	child	niño (general)
dck	duck	pato
dog	dog	perro
dom	domestic animal	animal doméstico (perro, gato, etc.)
frg	frog	rana
ger	gerbil	jerbo
gpn	guinea pig	conejillo de indias
ham	hamster	hámster
hmn	human	humano
inf	infant	infante
mam	mam	mamífero
man	man	hombre
mky	monkey	mono
mus	mouse	ratón
pgn	pigeon	pichón
pig	pig	puerco
qal	quail	codorniz
rat	rat	rata
rbt	rabbit	conejo
trk	turkey	pavo
unk	unknown	desconocido
wmn	woman	mujer

Apéndice B - Vías de administración (Milne 1995)

Abreviatura	Extensión	Traducción
ice	intracerebral	intracerebral
ihl	inhalation	inhalación
ims	intramuscular	intramuscular
ipr	intraperitoneal	intraperitoneal
ivn	intravenous	intravenosa
ocu	intraocular	intraocular
orl	oral	oral
par	parenteral	parenteral
scu	subcutaneous	subcutáneo
skn	topically applied to the skin	tópicamente aplicada en piel
unr	unrecorded	sin registro

Apéndice C – Tabla de clasificación de toxicidad de DL_{50} vía oral y cutánea rata de la OMS (World Health Organization 2010).

Clasificación OMS		DL_{50} para rata (mg/Kg en peso corporal)	
		Oral	Cutánea
I A	Extremadamente peligroso	< 5	< 50
I B	Altamente peligroso	5 – 50	50 – 200
II	Moderadamente peligroso	50 – 2000	200 – 2000
III	Ligeramente peligroso	Arriba de 2000	Arriba de 2000
U	Poca posibilidad de presentar peligro agudo	5000 o más	

Apéndice D – Conceptos básicos (Jonh H. y Howard 2007; Hacker, Bachmann, y Messer 2009; Gold et al. 2001; Goyer y Clarkson 2001).

Para poder comprender mejor el trabajo desarrollado durante la generación de la base de datos y su posterior evaluación, es necesario definir ciertos términos utilizados a lo largo de esta tesis, los cuales se listan a continuación:

Toxicología: Es el estudio de los efectos adversos de compuestos químicos en los organismos vivos y ecosistemas, los efectos dañinos ante la exposición de estos compuestos químicos, sus mecanismos de acción, diagnósticos, su prevención y tratamiento ante una intoxicación.

Toxicidad: Capacidad para causar daño a un organismo vivo definida con referencia a la cantidad de sustancia administrada o absorbida, la forma de administración de la sustancia, su distribución en el tiempo (dosis únicas o múltiples), el tipo y la gravedad de la lesión, el tiempo necesario para producir la lesión, la naturaleza de la lesión del organismo afectado y otras condiciones específicas.

Toxicidad aguda: Son los efectos adversos de duración finita que ocurren dentro de un tiempo corto (puede ser de hasta 14 días) después de la administración de una dosis única (o exposición a una concentración dada) de una sustancia de prueba o después de dosis múltiples, por lo general, dentro de las 24 horas del punto de inicio.

Toxicidad crónica: Efectos que persisten sobre un periodo largo de tiempo, pueden ocurrir o no inmediatamente en la exposición a un compuesto químico.

Contaminante: Es todo elemento, compuesto, sustancia, derivado químico o biológico, energía, radiación, vibración, ruido o una combinación de ellos, cura presencia en el ambiente, en ciertos niveles, concentraciones o periodos pueden constituir un riesgo para la salud.

Vía de administración: Camino elegido por el cual una sustancia introducida a un organismo, por ejemplo: vía oral, urinaria, cutánea, intravenosa, inhalada.

Intoxicación: Proceso anómalo que es causado por sustancias químicas y caracterizado por desequilibrio fisiológico secundario a modificaciones bioquímicas en el organismo, manifestado a través de signos, síntomas y exámenes de laboratorio.

Dosis Tóxica: Cantidad de sustancia que produce intoxicación sin llegar a ser letal.

Dosis Letal: Cantidad de sustancia o agente físico que causa la muerte al ser introducida en el organismo.

Dosis letal 50 (DL_{50}): Dosis a la cual un agente químico o físico es capaz de producir la muerte del 50% de los organismos de una población bajo un conjunto de condiciones definidas.

Concentración Letal 50 (CL_{50}): Concentración (en aire o agua) de un agente químico o físico que es capaz de producir la muerte del 50% de los organismos de una población bajo un conjunto de condiciones definidas.

Agroquímicos: Son sustancias químicas que se emplean con recurrencia en la agricultura cuya finalidad es mantener y conservar los cultivos.

Pesticidas: Es cualquier sustancia o mezcla de sustancias destinadas a prevenir, destruir, repeler o mitigar cualquier peste.

QSAR (Quantitative Structure Activity Relationships): Relaciones Cuantitativas Estructura – Actividad, es una técnica que consiste en la construcción de modelos matemáticos con los cuales se relacionan las estructuras químicas con una actividad biológica o química por medio de herramientas estadísticas, para su posterior determinación cuantitativa.

Descriptores moleculares: Son el resultado final de procedimientos matemáticos que permiten transformar la información química, codificada dentro de una representación simbólica de moléculas, en números útiles.

Capítulo I

Introducción

1.1 Introducción

El continuo uso de compuestos químicos en productos básicos tales como alimentos procesados, medicamentos, pinturas y pesticidas, demanda la implementación de medidas de seguridad a nivel de regulación federal e internacional, particularmente para aquellos compuestos en contacto directo con seres humanos. Diversos organismos, nacionales e internacionales han creado reglamentos para el correcto uso de los compuestos químicos que se encuentren en contacto con seres humanos o diversas especies benéficas para la biodiversidad, para así evitar que dichos organismos no se vean afectados por el uso de estos. Para cumplir con los requerimientos regulatorios, existen diversos tipos de ensayos enfocados a determinar la toxicidad de compuestos químicos, entre ellos se encuentran los siguientes: 1) ensayos *in vivo* en los que se hace uso de animales en diferentes vías de administración. 2) ensayos *in vitro* en los que se utilizan tejidos, órganos, bacterias, hongos, algas y cultivos celulares; y por último 3) métodos *in silico* en los cuales se estima alguna actividad biológica, fisicoquímica o climatológica por medio de métodos computacionales (Gozalbes, Ortiz, y López 2014). Los ensayos *in vitro* e *in silico* han cobrado auge debido a que son técnicas ecológicamente más amigables, rápidas y con un costo menor. Dentro de las técnicas más utilizadas en la toxicología computacional, se encuentran los estudios de relaciones cuantitativas estructura - actividad (QSAR por sus siglas en ingles). En estos métodos se relacionan, a través de modelos matemáticos, características de compuestos químicos con una actividad biológica, fisicoquímica o toxicológica que se desea determinar.

México, al ser un país productor de materias primas, continuamente utiliza distintos agentes agroquímicos, los cuales sirven tanto para nutrir la tierra de cultivo (fertilizantes) como para cuidar a las plantas de posibles plagas, estos agroquímicos son conocidos como pesticidas. De esta manera, los usos de agentes agroquímicos ayudan a los agricultores mexicanos en la producción de frutos y vegetales que serán exportados y consumidos en el país.

Dado que los pesticidas son utilizados especialmente para erradicar pestes, son *per se* agentes tóxicos para animales superiores tales como mamíferos, aves, peces e inclusive animales inferiores esenciales para la polinización como las abejas. Por lo tanto, es imperante verificar el efecto de dichos pesticidas al entrar en contacto con el ser humano, tanto en el riesgo que representan por la exposición crónica como aguda. La responsabilidad de salvaguardar a la población y biodiversidad recae tanto en los fabricantes de pesticidas como en las autoridades sanitarias de los diferentes gobiernos.

El uso de metodologías computacionales se deriva de la disponibilidad de información y procedimientos válidos. Gracias a la información toxicológica acumulada a través de los años y al desarrollo de modelos computacionales, es posible la predicción de la toxicidad de compuestos químicos para ensayos particulares. En México y en diversas partes del mundo, se están empleando métodos de cómputo para la predicción de la toxicidad de pesticidas. Esto contribuye a corroborar la seguridad toxicológica de los pesticidas que ya están en uso y determinar la de aquellos que se pretendan introducir al mercado. Los estudios QSAR permiten la predicción de la toxicidad de nuevos compuestos a bajo costo, de manera rápida, y sin el uso adicional de animales. Esto enmarca una alternativa amigable con el medio ambiente que permita seguir impulsando la industria agroquímica mexicana.

1.2 Hipótesis y Objetivos

En México y en diversas partes del mundo se ha comenzado a implementar estudios QSAR como una herramienta para la predicción de ensayos de toxicidad. El empleo de estas metodologías es posible gracias a la información toxicológica acumulada a través de los años y el avance en la implementación y validación de los métodos computacionales.

La hipótesis planteada en la presente tesis es que a partir de una base de datos confiable de pesticidas que contenga evaluaciones toxicológicas en diversas especies, será posible establecer extrapolaciones interespecie. Esta hipótesis será válida en tanto se disponga de un gran número de datos correlacionados.

Como siguiente paso, estos datos permitirán generar modelos matemáticos para la predicción de las evaluaciones toxicológicas contenidas en la base de datos. De ser validadas, dichas predicciones podrán ser utilizadas en la industria agroquímica, con fines regulatorios.

Derivado del planteamiento anterior se establecieron los siguientes objetivos:

1.- Generar una base de datos de pesticidas con información química y toxicológica previamente publicados en libros y manuales, con la finalidad de obtener extrapolaciones que permitan entender cómo cambia la toxicidad entre una especie y otra.

2.- Con la información colectada en la base de datos de pesticidas, realizar modelos QSAR para la predicción de evaluaciones toxicológicas.

Capítulo II

Marco Teórico

2.1 Antecedentes Históricos de los estudios QSAR

A partir de su concepción en los años 60's, los estudios QSAR han evolucionado grandemente, tanto en las metodologías *per se* cómo en la disponibilidad de información y métodos de validación (Cherkasov et al. 2014). Para el año 2014, las predicciones de evaluaciones biológicas y toxicológicas, basadas en modelos QSAR, comenzaron a ser aceptadas por autoridades regulatorias tanto nacionales como internacionales. Esto ha hecho de las predicciones teóricas una alternativa válida en el ámbito regulatorio. Dichas predicciones deben cumplir con algunos de los requisitos establecidos por la OCDE, que se mencionaran más adelante en esta tesis.

En 1962, con la publicación de Hansch *et al* se dio inicio a los estudios QSAR, culminando así con 15 años en búsqueda para comprender las bases de las relaciones estructura-actividad en los reguladores del crecimiento de las plantas (Hansch et al. 1962). Hansch utilizó en un principio las relaciones de Hammlett y los argumentos de Veldstra para obtener modelos que relacionaran los sustituyentes con la reactividad química y los efectos lipofílicos con la potencia biológica (1/C, donde C es la concentración del fármaco) de un fármaco que atraviesa un medio biológico, sin obtener resultados positivos (Veldstra 1953). Posteriormente recurrió a los coeficientes de partición octanol – agua como un sustituto de la medida de la lipofilia. En 1961, Hansch junto con Fujita, quien también se encontraba experimentando con los coeficientes de partición octanol – agua (log P), observaron que el log P es una propiedad aditiva, es decir que la contribución parcial de un sustituyente de una molécula al log P es similar al de otra molécula de un compuesto (Fujita, Iwasa, y Hansch 1964).

A final de los años 50´s, Taft amplió las relaciones lineales de energía libre para ajustar una ecuación no sólo con efectos electrónicos, sino también con efectos estéricos (Taft 1956). En contraste con esto, los bioquímicos farmacólogos se centraron en los efectos de log P en la absorción de fármacos. A mediados del siglo XX, Fieser demostró gráficamente la relación que existe entre el potencial

antimalárico de los naftoquinones y el coeficiente de distribución éter – agua (Fieser, Ettlinger, y Fawaz 1948). Para el año 1959 Kauzmann respaldó la importancia de la hidrofobicidad para la determinación de la estructura de las proteínas, aumentando así la importancia del log P sobre la potencia biológica de un fármaco (Kauzmann 1959). Al encontrar la relación entre el log P y el potencial biológico, Hansch y Fujita lograron incluir ambos términos en una misma ecuación, demostrando el éxito del enfoque computacional para el modelado de efectos cuantitativos de los sustituyentes en la actividad de un fármaco (Hansch 1969).

Una vez que las aproximaciones de Hansch fueron establecidas, su factibilidad se incrementó mediante la aplicación de cálculos cuantitativos, proporcionando así una alternativa para explorar la actividad en los determinantes electrónicos y estéricos en compuestos químicos estrechamente relacionados. Como ejemplo, Pullmans mostro cómo el potencial cancerígeno de los hidrocarburos aromáticos está relacionado con la estructura electrónica para la predicción del potencial de la región bahía en la activación metabólica de los reactivos intermediarios diol – epóxido en el ADN. A principios de los 80´s, Klopman propuso el rompimiento de una molécula en fragmentos constituyentes 2D, para autogenerar fragmentos de grandes números de moléculas en un conjunto de entrenamiento y con este conjunto correlacionar la frecuencia de estos fragmentos con la actividad biológica (Klopman 1984). La anterior propuesta fue un gran avance en la creación de modelos computacionales eficientes, capaces de representar y correlacionar características estructurales fácilmente interpretables.

Hansch, Kutter y Charton demostraron que los valores E_s de Taft, son parámetros de sustituyentes simétricos relacionados con su radio (Kutter y Hansch 1969; Charton 1969). Hansch además utilizó con la refractividad molar de los sustituyentes como medida de su volumen. Sin embargo, no fue hasta el desarrollo del método CoMFA (análisis comparativo del campo molecular) y otras aproximaciones 3D, que las interacciones energéticas del potencial electrostático a través de una serie de estructuras relacionadas superpuestas fueron tomadas en consideración, siendo así CoMFA la primera demostración exitosa de un QSAR 3D (Cramer, Patterson, y

Bunce 1988). Con el incremento del conjunto de datos y estructuras más diversas, se generaron más descriptores que sirvieran como indicadores para distinguir una serie de moléculas, proponiendo de esta manera el uso de descriptores para modelar directamente la actividad de una serie de moléculas. La aplicación de los modelos QSAR no solo se ha limitado a la generación de correlaciones entre actividades biológicas y los descriptores de una serie de moléculas, también se ha avanzado en métodos de evaluación de confiabilidad usando métodos estadísticos para evitar correlaciones no causales (llamadas *chance correlations*) y usando conjuntos de prueba que no fueron utilizados para desarrollar el modelo predictivo (validación externa) (Klopman y Wang 1991; Hall y Kier 2001).

2.2 Descriptores moleculares

Los descriptores moleculares son fundamentales en el desarrollo de modelos QSAR y pueden ser generados a partir de representaciones estructurales con diferentes niveles de complejidad. Los tipos de descripción estructural comprenden desde las fórmulas moleculares (llamada 1D), fórmulas estructurales bidimensionales (2D), representaciones dependientes de la conformación tridimensional (3D) y en los niveles más altos, los que toman en cuenta la orientación de las moléculas y la dinámica molecular (4D) (Polanski 2009; Todeschini y Consonni 2010). En la práctica, los más usados son los descriptores 2D y 3D.

Los descriptores 2D son obtenidos de representaciones bidimensionales de una molécula, es decir la topología de la molécula, que definen la conectividad de sus átomos en términos de presencia y la naturaleza de los compuestos químicos. Las principales ventajas del uso de descriptores 2D es que contienen información simple y útil sobre la estructura molecular, son invariables en cuanto a la roto-traslación de las moléculas y se pueden calcular si la necesidad de optimizar las estructuras moleculares. Esto hace que su principal desventaja sea que no siempre se pueden reconstruir las estructuras, por lo que es necesario que la secuencias estén correctamente ordenadas, definidas en 2D, para caracterizar moléculas con mayor discriminación. Por otra parte, los descriptores 3D surgen a partir de la necesidad

de hacer un muestreo sistemático de las diferencias espaciales, ya que muchas expresiones QSAR sugieren que las propiedades biológicas dependen de interacciones específicas en procesos de reconocimiento molecular (Cramer, Patterson, y Bunce 1988). El uso de los descriptores 3D puede proporcionar ventajas ya que la representación de las moléculas facilita el uso de representaciones gráficas para el mejoramiento de compuestos químicos: extender cadenas, incluir sustituyentes que mejoren la interacción con la biomacromolécula involucrada. Cabe mencionar que su principal desventaja es la dependencia de la alineación de los ligandos tanto del grupo de prueba como los del grupo de entrenamiento.

2.3 Aplicación de QSAR en toxicología

En el campo de la toxicología el empleo de métodos QSAR se aplica en la predicción de toxicidad en cultivos celulares (*in vitro*) o para pruebas con animales (*in vivo*), donde la evaluación final de la toxicidad puede tener diferentes mecanismos o bien, no tener un mecanismo de interacción (Milan et al. 2011). Para la generación de los modelos, es muy relevante la selección de los compuestos del conjunto de entrenamiento y el espacio químico en el cual el modelo es aplicable, espacio de predicción (van Leeuwen et al. 2009).

Los estudios QSAR utilizados para la predicción de toxicidad deben cumplir con las siguientes características: (I) se prefiere que los compuestos dentro del conjunto de entrenamiento sean estructuralmente similares, lo que hace que la diversidad de mecanismo de acción se reduzca, (II) la evaluaciones toxicológicas modeladas sin objetivo específico o sujetas a la reactividad química deben fundamentarse en principios químicos bien establecidos, (III) las evaluaciones toxicológicas deben estar relacionadas con un objetivo molecular bien definido, (IV) Los datos de toxicidad deben estar disponibles para una cantidad suficientemente grande de compuestos para capturar la mayor cantidad de asociaciones estructura-actividad.

2.4 Diseño de bases de datos de pesticidas

La necesidad de tener información suficiente para los sistemas de modelado y toma de decisiones en evaluaciones de riesgo de pesticidas llamó la atención de organismos internacionales como la IUPAC, quienes se dieron a la tarea de crear una base de datos autorizada de pesticidas con contenido fisicoquímico y toxicológicos suficientemente robusto, publicando así en el 2006 la base de datos de propiedades de pesticidas (PPDB por sus siglas en inglés). Esta base de datos es de dominio público en línea y cimentó las bases para la generación de futuras bases de datos, proporcionando información accesible al garantizar que el contenido de la base de datos fuera confiable (Lewis y Green s/f). Para facilitar el uso de la base de datos para evaluaciones de riesgo y la búsqueda de parámetros particulares para un producto en específico, se almacenó la información en formato MS Access y luego fueron transmitidas a través de filtros de datos para formatearse en páginas HTML ("Environmental Information Sheets (EIS) - Voluntary Initiative" s/f), en su versión en línea. La distribución de la información almacenada en la base de datos de pesticidas se dividió en 5 áreas discretas:

-Información general: estructuras, nombres de identificadores, traducciones en otros idiomas, códigos, etc.

-Datos fisicoquímicos: solubilidad, densidad, índice de refracción, punto de ebullición, punto de fusión, etc.

-Destino ambiental: constante de Henry, tasas de degradación en el suelo, sedimentos en agua, etc.

-Salud humana: clasificación de toxicidad de la OMS, ingestas diarias, toxicidad en mamíferos, evaluación de toxicidad, etc.

-Ecotoxicología: toxicidad aguda, toxicidad crónica para la fauna y flora, bioacumulación, entre otros.

La base de datos garantiza un conjunto de información armonizada y equilibrada al adaptar sus datos a las condiciones particularmente sensibles como lo son el clima y el tipo de suelo. Cuando estas condiciones son muy variables se adaptan a las condiciones de la Unión Europea. Por otra parte, la calidad de los datos contenidos

en la base de pesticidas de la IUPAC cuenta con un medidor de calidad que consta de dos partes. La primera parte se encarga de identificar el tipo de fuente de los datos (publicaciones reguladas, revistas arbitradas, datos de fabricación, etc.) La segunda parte evalúa la confianza de los datos en una escala del 5 al 0. Como ejemplo la calificación de conjuntos de datos regulados normalmente e de 5, mientras que datos de fuentes no estándar o no referenciados reciben calificación de 1 ó 0. El mantenimiento de esta base de datos se hace bajo protocolos estrictos y constantemente se está actualizando y revisando la información que puede entrar a esta base de pesticidas.

2.5 Curado de base de datos

Como parte de la preparación de las bases de datos, es importante definir si existen errores en su creación y/-o en su uso. Estos errores se conocen como errores de categorización y errores de clasificación (Waldman, Fraczkiewicz, y Clark 2015). Los errores de categorización están referenciados a la mala interpretación de la información encontrada en la literatura. Los errores de clasificación son aquellos dados por la mala interpretación de las predicciones. Para la creación de una base de datos solo se contemplan los errores categóricos. Esto conduce a la necesidad de analizar qué tipo de errores categóricos se pueden hallar en la información encontrada y como poder repararlos para así optimizar la calidad de la información. Los errores más típicos son los siguientes:

Errores en la clasificación de sustancias: En el área relacionada a este trabajo, es necesario que las sustancias contenidas en la base de datos sean utilizadas y clasificadas como pesticidas para evitar predicciones erróneas (Miners et al. 2006; Kaivosaari, Finel, y Koskinen 2011).

Errores en las unidades: El error más común en la elaboración de una base de datos es el del tipo de unidades utilizadas, originado por que muchos de los datos proporcionados están reportados en unidades inglesas, del sistema internacional o expresadas en diferentes unidades dentro del mismo sistema, por ejemplo concentraciones expresadas en mM y µM (Kiyoi et al. 2011; Balakin et al. 2004).

Errores en los nombres: Al igual que el error de unidades este es uno de los más comunes al crear una base de datos, ya que un solo pesticida o cualquier otro compuesto químico puede tener diferentes nombres. Para evitar este error es necesario realizar una búsqueda exhaustiva de los posibles nombres para un mismo compuesto y eliminar aquellos que estén duplicados (Weininger 1988; Weininger, Weininger, y Weininger 1989; “Chemical Substances - CAS REGISTRY” s/f).

Errores en las estructuras: Este error es uno de los que no se tiene mucho conocimiento y cuidado. Es por esto que resulta fundamental analizar minuciosamente cada estructura y verificar que este correctamente construida en su formato 2D o 3D, en libros y sitios especializados (Bolton et al. 2008; Durant et al. 1977; Sundriyal et al. 2008).

Desviaciones en propiedades fisicoquímicas: Este error está dado principalmente porque algunas de las propiedades fisicoquímicas son calculadas con diferentes algoritmos, por lo que tendrán variaciones dependiendo del software utilizado. Para evitar errores debidos a estas desviaciones es recomendable utilizar de manera consistente el mismo software para el cálculo de propiedades fisicoquímicas y demás descriptores moleculares (Tetko et al. 2014; Avdeef et al. 1996; Clark et al. 2014).

La mayoría de los errores se corrigen al revisar la información de la base de datos y hacer las comparaciones correspondientes, convirtiendo las unidades, haciendo una agrupación de todos los nombres de la molécula, manteniendo un dato de toxicidad por compuesto, hacer una búsqueda en los registros de pesticidas de la OCDE para reconocer a los compuestos como pesticidas y verificando que las estructuras de los compuestos sean correctas.

2.6 Cálculo de modelos predictivos de evaluaciones entre especies.

Los modelos predictivos a partir de solo evaluaciones toxicológicas entre especies propuestos por la Academia Nacional de la Ciencia (NAS por sus siglas en inglés), parten de análisis estadísticos donde se pueden establecer interpolaciones entre

especies siempre y cuando se cuente con información necesaria (US EPA s/f). Para poder realizar el cálculo de estas interpolaciones se debe seleccionar la especie sustituida y buscar los datos correlacionados que intervendrán en dicho cálculo. La toxicidad prevista se puede calcular como:

$$y = 10^{[a+(b*Log(x))]} \quad (1)$$

Donde *x* es el valor de toxicidad de la especie sustituida, *a* y *b* son la intersección y la pendiente respectivamente.

Los límites de confianza inferior y superior se calculan como:

$$Limite\ inferior = 10^{[Log(y) - t\%CI*\left[MSE*\frac{1}{df+2}+\frac{2(x* -x)}{Sxx}\right]^2]} \quad (2)$$

$$Limite\ superior = 10^{[Log(y) + t\%CI*\left[MSE*\frac{1}{df+2}+\frac{2(x* -x)}{Sxx}\right]^2]} \quad (3)$$

Donde *y* es el valor de toxicidad estimado de los tóxicos predichos, *t%Ci* es el valor de ten el nivel de confianza deseado (90%, 95%, 99%), MSE es el error medio cuadrático y *Sxx* es la suma de las desviaciones cuadradas del sustituto (Dowdy, Weardon, y Chilko 2005).

2.7 Cálculo de modelos QSAR

La predicción de la toxicidad de un compuesto químico involucra la generación de modelos matemáticos en los cuales se relacionan los descriptores moleculares y valores de toxicidad previamente evaluados experimentalmente (Todd M. Martin et al. 2008). Algunos de los métodos utilizados son los siguientes:

Método jerárquico: En este método la toxicidad de un compuesto se puede calcular utilizando el promedio ponderado de las predicciones de varios modelos diferentes. Estos modelos son obtenidos a partir del método de Ward para dividir el conjunto de entrenamiento en una serie de grupos similares (Romesburg 1984). Para este método es necesario determinar la varianza, la desviación estándar, distancia entre

grupos, y el coeficiente de validación, entre otros. A continuación, se describirá como calcular cada parámetro.

La varianza general está definida como la suma de las variaciones individuales de cada grupo:

$$V(l) = \sum_{k=1}^{m} v(k,l) \quad (4)$$

Donde *v(k,l)* es la varianza (en términos de los descriptores moleculares) para el grupo *k* en el paso *l*:

$$v(k,l) = \sum_{i=1}^{n_k} \sum_{j=1}^{d} (x_{ij} - C_j)^2 \quad (5)$$

Donde n_k es el número de compuestos químicos en el grupo, d es el número de descriptores en el conjunto total de descriptores, x_{ij} es el descriptor normalizado *j* para el compuesto químico *i* y *Cj* es el valor promedio del descriptor:

$$C_j = \frac{1}{n_k} \sum_{i=1}^{n_k} x_{ij} \quad (6)$$

Cada paso de este método conjunta dos grupos en uno, incrementando la varianza en todos los grupos, a partir de esto se puede minimizar esta varianza de la siguiente manera:

$$\Delta V_{min}(l+1) = V(l+1) - V(l) = v(k', l+1) - v(k_1.l) - v(k_2, l) \quad (7)$$

Donde el conjunto k_1 y k_2 se juntan en el paso *l* para generar el conjunto *k'* en el paso *l + 1*. El proceso de combinaciones conjuntos continua hasta que todos los compuestos químicos se juntan en un solo conjunto.

Después de que el conjunto está completo, cada conjunto es analizado para determinar si se puede generar un modelo QSAR. Cada conjunto es evaluado

usando algoritmos genéticos para determinar los descriptores óptimos para caracterizar los valores de toxicidad de los compuestos químicos de ese conjunto. El máximo número de descriptores permitido para cada conjunto esta dado por la relación $n_k/5$ (Eriksson et al. 2003; J G Topliss y Edwards 1979), siendo al menos de 5 el número de compuestos por descriptor, es decir un quíntuple.

Los algoritmos genéticos son usados para maximizar el ajuste del quíntuple, que se quedó fuera del coeficiente de validación cruzada ($q^2_{adj,\,LMO}$):

$$q^2_{adj,LMO}\left[\frac{\frac{\sum_{i=1}^{n_k}(\hat{y}_i - y_{\exp\,j})^2}{n_k - p - 1}}{\frac{\sum_{i=1}^{n_k}(y_{exp\,j} - \bar{y}_{exp})^2}{n_k - 1}}\right] \quad (8)$$

Donde yi y $y_{exp\,j}$ son las toxicidades predicha y experimental, respectivamente. $\bar{Y}_{exp}$ es el promedio de la toxicidad experimental de los compuestos químicos del grupo y p es el número de parámetros en el modelo (Witten y Frank 2005).

La toxicidad predicha ($\hat{y}$) para un compuesto químico a evaluar está dada por el promedio ponderado de todas las predicciones validadas:

$$\hat{y} = \frac{\sum_{j=1}^{nvc} w_j \hat{y}_j}{\sum_{j=1}^{\#grupos\ validos} w_j} \quad (9)$$

Donde yj y wj es la predicción y el peso para el modelo y nvc es el número de grupos validos de predicción (David, Dennis, y Thomas 2011).

El valor del peso está dado por:

$$w_j = \frac{1}{se_j^2} \quad (10)$$

Donde se_j es el error estándar j de la predicción y está dado por:

$$se_j = \sqrt{\sigma_j^2(1 + h_{00})} \quad (11)$$

σ_j^2 se define de la siguiente manera:

$$\sigma_j^2 = \frac{\sum_{i=1}^{nj}(\hat{y}_i - y_{\exp j})^2}{n_j - p_j - 1} \quad (12)$$

donde n_j es el número de compuestos químicos en el modelo de grupo j, p_j es el número de parámetros del modelo j y h_{00} es el apalancamiento (medida para saber que tanto se alejan los descriptores de una predicción de los descriptores de otra predicción) del compuesto de prueba dado por:

$$h_{00} = X_o^T (X^T X)^{-1} X_0 \quad (13)$$

donde X_0 es el vector del modelo de valores de los descriptores para el compuesto de prueba.

El cuadrado de la desviación estándar para la predicción de múltiples modelos puede aproximarse a:

$$\sigma_\mu^2 = \frac{\overline{\sigma^2}}{nvc} = \left(1/nvc\right) \frac{\sum_{j=1}^{nvc} w_j se_j^2}{\sum_{j=1}^{nvc} w_j} = \left(1/nvc\right) \frac{\sum_{j=1}^{nvc} \left(1/se_j^2\right) se_j^2}{\sum_{j=1}^{nvc} \left(1/se_j^2\right)}$$

$$= \frac{1}{\sum_{j=1}^{nvc} \left(1/se_j^2\right)} \quad (14)$$

La incertidumbre en la predicción general del compuesto de prueba es:

$$\hat{u} = t_{1-\alpha/2,nvc}\sigma_\mu = t_{1-\frac{\alpha}{2},nvc-1} \sqrt{1/\sum_{j=1}^{nvc} \frac{1}{se_j^2}} \quad (15)$$

Donde t es el valor en el estadístico t de Student, α = 0.1(intervalo de confianza del 90%) y sej es el error estándar de la predicción. El intervalo de predicción se obtiene al sumar y restar la incertidumbre de la predicción de toxicidad:

$$\hat{y} - \hat{u} \leq Toxicidad \leq y + \hat{u} \quad (16)$$

El intervalo de predicción indica que tiene el 90% de confianza de que la toxicidad real está entre $\hat{y} - \hat{u}$ y $y + \hat{u}$.

La incertidumbre de la predicción para un grupo modelo está dada por (Montgomery, Peck, y Vining 2001):

$$u_j = t_{1-\frac{\alpha}{2},nj-p-1}\sqrt{\sigma^2(1+h_{00})} \quad (17)$$

La incertidumbre es una función de la calidad del modelo de regresión (σ^2) y la distancia (en el espacio descriptivo del modelo) entre el químico de prueba y los químicos usados en el grupo para generar el modelo (h_{00}).

Para definir el dominio de aplicabilidad en el método jerárquico se hace uso de dos restricciones. En la primera se verifica si la sustancia problema se encuentra dentro de la elipse multidimensional definida por los rangos de los valores de los descriptores del grupo con el que se generó el modelo, la restricción se satisface si el aplacamiento del compuesto de prueba (h_{00}) es menor al valor máximo de aplacamiento de los compuestos usados en el modelo (Montgomery, Peck, y Vining 2001). La segunda restricción (Rmax) verifica si la distancia del compuesto de prueba al centro del conjunto de compuestos químicos es menor a la distancia máxima de cualquier compuesto dentro del conjunto al centro de este conjunto. La distancia está definida en términos de la tabla completa de descriptores:

$$distancia\ i = \sum_{j=1}^{d}(x_{ij} - C_j)^2 \quad (18)$$

Donde la distancia i es la distancia del compuesto químico i al centro del grupo.

Método FDA: La predicción para cada compuesto químico de prueba se hace usando un nuevo modelo que ajusta los compuestos químicos similares al compuesto químico de prueba. Este método solo utiliza un conjunto de compuestos generado en el transcurso de ejecución.

En el método FDA, propuesto por Contrera y sus colaboradores, se seleccionan de 15 a 20 compuestos para el conjunto de entrenamiento con una similitud del 75% con el compuesto de prueba (Contrera, Matthews, y Benz 2003). El coeficiente del coseno de similitud está definido por:

$$SC_{i,k} = \frac{\sum_{j=1}^{\#descritor} x_{ij} x_{kj}}{\sqrt{\sum_{j=1}^{\#descriptor} x_{ij}^2 \sum_{j=1}^{\#descriptor} x_{kj}^2}} \quad (19)$$

Donde x_{ij} es el valor de j descriptor normalizado por el compuesto *i*.

Método de un solo modelo: La predicción en este modelo se hace a través de una regresión lineal múltiple (usando descriptores moleculares como variables independientes) usando un enfoque de algoritmo genético (Benigni y Richard 1996).

Método del grupo de contribución: Las predicciones se realiza usando un modelo de regresión lineal múltiple que ajusta al conjunto de entrenamiento (usando fragmentos moleculares como variables independientes) (Todd M. Martin y Young 2001).

Método de vecinos cercanos: La toxicidad predicha está estimada por el promedio de los tres compuestos con mayor índice de similitud al compuesto de prueba dentro del conjunto de entrenamiento (Todd M. Martin et al. 2008).

Método consenso: Esta toxicidad esta predicha al tomar el promedio de la toxicidad predicha en los métodos anteriores (Zhu et al. 2008).

Método "bosques aleatorios" (random forest): La toxicidad es predicha utilizando un árbol de decisión que almacena un compuesto químico en una determinada

puntuación de toxicidad usando un conjunto de descriptores moleculares como variables de decisión (Cassano et al. 2010).

Método de modo de acción: la predicción de toxicidad se estima usando un proceso de dos pasos, en el primero se determina a partir del modelo de análisis de discriminante lineal. Para el segundo la toxicidad se estima usando un modelo de regresión lineal múltiple (T. M. Martin et al. 2015; Todd M. Martin et al. 2013).

Método de redes neuronales artificiales: Este modelo usa la retropropagación para clasificar descriptores y a su vez generar predicciones. Esta red, inspirada en una red neuronal de un cerebro humano, se puede construir a mano, creada por un algoritmo o ambos (Kubat 1999). La red se puede monitorear y modificar durante el tiempo de entrenamiento del modelo QSAR. Los nodos en esta red son todos sigmoides.

Para entender mejor el comportamiento de una red neuronal artificial, es necesario conocer las partes que conforman una neurona en este modelo, las cuales se encuentran en la **Figura 1**.

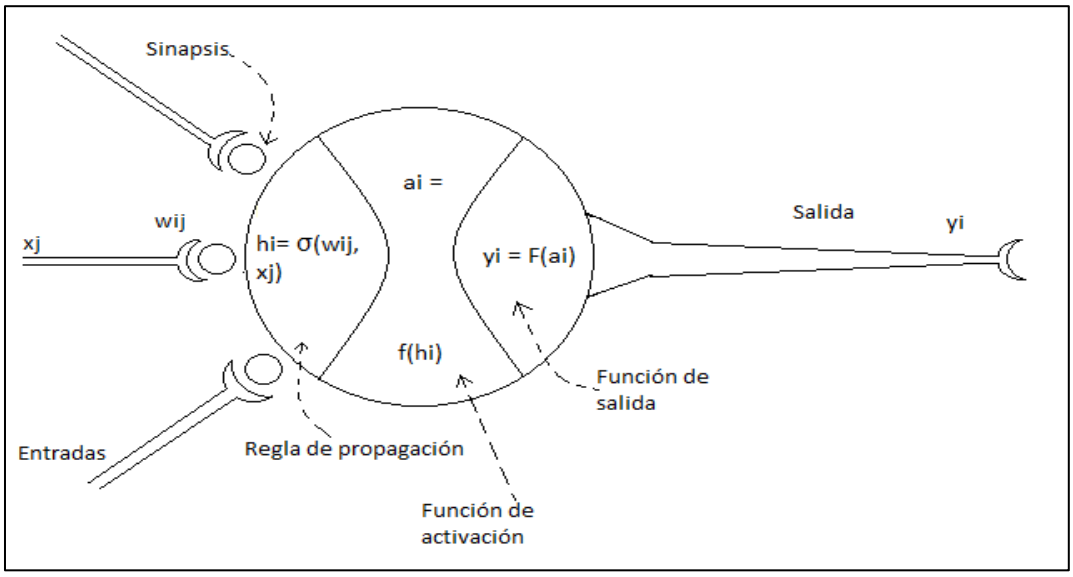


Figura 1. Representación de una red neuronal artificial (Rumelhart, McClelland, y The PDP Research Group 1988)

Entradas: $x_j\ (t)$. Son las variables de entrada de la neurona, pueden ser binarias (digitales) o continuas (analógicas) dependiendo del modelo de aplicación.

Pesos sinápticos w_{ij}*:* Representa la intensidad de interacción entre cada neurona presináptica *j* y la neurona postsináptica *i*.

Regla de propagación: $\sigma(w_{ij}, x_j\ (t))$. Proporciona el valor del potencial postsináptico, $h_i\ (t)$, de la neurona 1 en función de sus pesos y sus entradas:

$$h_i(t) = \sigma\left(w_{ij}, x_j(t)\right) \qquad (20)$$

La función de propagación más utilizada es la lineal, basada en la suma ponderada de las entradas con los pesos sinápticos.

$$h_i(t) \sum_j w_{ij} x_j = w_i^r x \qquad (21)$$

Siendo así el peso sináptico el responsable de excitación (pesos positivos) o de la inhibición (pesos negativos) de la neurona postsináptica. Para poder pasar al estado de activación, la suma de los pesos multiplicados por las entradas $(h_i(t))$ debe superar el valor umbral (θ), el cual es el valor en el que la neurona se activa y aporta información a la predicción, en caso de no superarlo esta neurona no se activa y por lo tanto no aporta información.

Función de activación o transferencia: $f_i(a_i(t\text{-}1), h_i(t))$. Proporciona el estado activado a_i, de la neurona i en función de su anterior a_i(t-1) y su potencial postsináptico actual:

$$a_i(t) = f_i\left(a_i(t-1), h_i(t)\right) \qquad (22)$$

Algunos modelos de redes neuronales dependen únicamente el estado actual de la neurona.

La función de activación por lo general toma la forma *y = f(x),* donde *x* es el potencial postsináptico mientras que *y* es el estado de activación. En la **Tabla 1** se muestran las funciones de activación más empleadas:

Tabla 1. Funciones de activación de red neuronal artificial (Sarangapani 2006)

	Función	Rango
Identidad	$y = x$	$[-\infty, +\infty]$
Escalón	$y = sign(x)$ $y = H(x)$	$\{-1, +1\}$ $\{0,+1\}$
Lineal a tramos	$y = \begin{cases} -1, & si\ x < -1 \\ x, & si\ +1 \leq x \leq -1 \\ +1, & si\ x > +1 \end{cases}$	$[-1,1]$
Sigmoidea	$y = 1/(1+e^{-x})$ $y = tgh(x)$	$[0, +1]$ $[-1,+1]$
Gaussiana	$y = Ae^{Bx^2}$	$[0, +1]$
Sinusoidal	$y = Asen(\omega x + \varphi)$	$[-1,+1]$

Función de salida: $F_i(a_i(t))$. Establece la salida actual $y_i\ (t)$, de la neurona i en función de su activación actual $a_i\ (t)$. Aunque frecuentemente la función de salida es igual a la función de activación de la neurona quedando de la siguiente manera:

$$y_i(t) = F_i\big(a_i(t)\big) = a_i(t) \qquad (23)$$

De modo que, la operación de la neurona *i* se puede expresar de la siguiente manera:

$$y_i(t) = F_i\left(f_i\left(a_i(t-1), \sigma_i\left(w_{ij}, x_j(t)\right)\right)\right) \quad (24)$$

En resumen, en el marco teórico se encuentran las definiciones que se utilizaron a lo largo del trabajo de tesis. Asimismo, se incluyen los antecedentes históricos en el desarrollo de la metodología QSAR con el uso de descriptores moleculares y la introducción de esta metodología en la determinación de la toxicidad, el diseño de una base de datos, así como del curado de la información contenida en esta, además del cálculo de modelos predictivos interespecie y de modelos QSAR a partir de diferentes metodologías. Con la investigación realizada a lo largo del marco teórico se generó una base de datos de pesticidas, se desarrollaron modelos de predicción interespecies y modelos QSAR para el presente trabajo.

Capítulo III

Metodología

La construcción de bases de datos requiere una búsqueda de información de manera extensa y detallada. Además, dicha información debe ser colectada cuidadosamente para evitar errores. A continuación, se presentan los pasos seguidos en este trabajo para la elaboración, curado y evaluación de la base de datos PESTIMEP.

Para la generación de la base de datos de pesticidas se utilizó el libro *"CRC Handbook of Pesticides".* Este libro contiene información de estructuras químicas correspondientes a pesticidas, nombres comunes de los mismos y sus sinónimos, propiedades fisicoquímicas básicas, además de contar con una gran cantidad de determinaciones toxicológicas de diferentes especies. Debido a lo anterior, el uso de este libro en la obtención de correlaciones de toxicidad entre especies y la construcción de modelos QSAR es fundamental.

3.1 – Almacenamiento de información de pesticidas

El primer paso para la generación de la base de datos PESTIMEP consistió en la construcción (representación) de estructuras en dos dimensiones (2D) de los pesticidas reportados en el libro "*CRC Handbook of Pesticides*". Las representaciones moleculares en dos dimensiones se elaboraron en ChemAxon Marvin Sketch, el cual permite convertirlos a cadenas ASCII cortas conocido como SMILES (*Simplified Molecular Input Line Entry Specification*). Con las estructuras en lenguaje SMILES se generó a una hoja de datos de MS Excel conteniendo el nombre del pesticida, la estructura en formato SMILES, las propiedades fisicoquímicas experimentales (densidad, punto de fusión, punto de ebullición, solubilidad y coeficiente de partición), la toxicidad en diferentes especies de animales encontradas en el libro (DL50, CL50) y el número de identificación del pesticida.

3.2 – Preparación de estructuras químicas y datos

El siguiente paso en la generación de la base de datos fue la preparación los datos y las estructuras químicas. Esto se debe a que la información que se almacenó

anteriormente no siempre se encuentra en las mismas unidades, además es necesario optimizar las estructuras y generar los modelos 3D a partir de los modelos 2D, a estas etapas se les conoce como (I) curado de datos y (II) optimización.

Etapa I – En la curación de la base de datos se verificó que los datos contenidos fueran únicamente de pesticidas; que los valores de toxicidad estuvieran en las mismas unidades; que no hubiera pesticidas duplicados y que todas las estructuras estuvieran correctamente construidas.

Etapa II – Para terminar la preparación de las estructuras se optimizaron las estructuras, esta etapa se realizó con el programa MOE (Molecular Operating Environment Software). Esta etapa se realizó en dos fases las cuales son:

1.- Se asignó el estado de protonación de las especies protonables, considerando pH 7.2. Además, se agregaron todos los átomos de hidrógeno requeridos.

2.-El segundo paso consistió en hacer que todas las geometrías estuvieran en un mínimo de energía, en la cual también quedaron definidos los centros quirales.

3.3 – Análisis de la base de datos PESTIMEP

La base de datos se analizó en dos etapas. La primera etapa consistió en un análisis estadístico únicamente de las diferentes evaluaciones toxicológicas de las sustancias encontradas en el libro *¨CRC Handbook of Pesticides"* con mayor número de datos. La segunda etapa consistió en la generación de un modelo predictivo de toxicidad usando la metodología QSAR con los descriptores calculados y seleccionados.

3.3.1 Análisis estadístico de evaluaciones toxicológicas

La primera evaluación de la base de datos se ejecutó mediante los programas Statistica StatSoft TIBCO y Spotfire TIBCO. En los cuales mediante el uso de herramientas de análisis estadístico se construyó histogramas, diagramas de caja, regresiones lineales, una matriz de correlación de las evaluaciones toxicológicas

(DL_{50}), así como gráficos de barra de la mediana de R^2 y la mediana de correlaciones entre especies. Adicionalmente se realizó un conteo entre las evaluaciones toxicológicas de diferentes especies, tomando en cuenta solo aquellas con una cantidad de datos mayores o iguales a diez. Posteriormente se seleccionó la evaluación toxicológica con mayor cantidad de datos como la columna objetivo para quitar los pesticidas sin datos de DL_{50}. La base de datos depurada de los pesticidas sin datos toxicológicos se utilizó en la segunda etapa de evaluación.

3.3.2 Generación de modelo QSAR para la predicción de toxicidad

La segunda evaluación consistió en la obtención de un modelo predictivo QSAR, para el cual fue necesario utilizar el programa Dragon Descriptor (“Molecular descriptors calculation - Dragon - Talete srl” s/f) en cálculo de los descriptores moleculares, el programa Knime para la selección de descriptores y finalmente el programa WEKA con el que se generó el modelo predictivo. Por lo tanto, esta etapa se dividió en 3 pasos presentados a continuación:

3.3.2.1 – Descripción del perfil fisicoquímico de las moléculas utilizando descriptores moleculares.

Una vez que se realizó la curación de la base de datos, así como la optimización de las estructuras moleculares, el siguiente paso fue la obtención de los descriptores moleculares, los cuales se calcularon utilizando el programa Dragon Descriptor. Algunos de los descriptores calculados son el número de átomos de carbono y de nitrógeno, el giro de spin, logP, la energía de ionización, el número de anillos, el índice de ramificación, el coeficiente de partición, entre otros.

3.3.2.2 – Selección de variables

Los descriptores que se utilizaron en el modelo se seleccionaron a través de metodologías de selección de variables. El número de descriptores debe estar acorde al número de moléculas que contenga la base de datos, esta proporción es empírica y corresponde aproximadamente a cinco moléculas por cada descriptor

utilizado, esto permite evitar el sobreajuste y sobre-especificación del modelo (John G. Topliss y Costello 1972).

La selección de los mejores descriptores para la generación de modelos QSAR se realizó a través de determinaciones estadísticas con el programa KNIME, un programa especializado en minería de datos. La minería de datos es un campo de la estadística en el que por medio aprendizaje automatizado y bases de datos, se pueden determinar patrones en un conjunto de datos (Nisbet, Elder IV, y Miner 2009). La selección de descriptores se estableció mediante las siguientes fases:

Fase 1.- Se generó una matriz de correlación lineal (de X por X) pareada considerando todos los descriptores calculados.

Fase 2.- Por medio de un filtro de correlación se eliminaron los datos que tuvieran una correlación de 0.9, 0.7 y 0.5 en diferentes procesos, ya que descriptores correlacionados generarían redundancia y disminuirían el poder predictivo de los modelos generados. De esta manera se mantuvo un descriptor por cada par de descriptores correlacionados.

Fase 3.- La siguiente etapa fue determinar filtro de baja varianza, esto se hizo por medio un método iterativo a través de un metanodo en Knime y consistió en establecer un lazo paramétrico de inicio enlazado con un metanodo y este a su vez fue conectado a un lazo final de parametrización. Este procedimiento permitió establecer el mejor parámetro para el filtro de baja varianza.

Fase 4.- Con el mejor parámetro establecido se generó un filtro de baja varianza para eliminar los descriptores que tuvieran baja dispersión, es decir aquellos con alta homogeneidad. Este paso es fundamental para evitar descriptores que puedan contener información distractora para algoritmos de aprendizaje (como las redes neuronales artificiales) que están basados en la distancia entre los descriptores moleculares (“KNIME Analytics Platform | KNIME” s/f).

Fase 5.- Posteriormente, se normalizaron todos los datos para que estuvieran en las mismas dimensiones y se agregó la información toxicológica (DL_{50}) con mayor número de datos.

Fase 6.- Una vez agregada la información toxicológica, se quitaron los pesticidas que estructuralmente no se encontraran dentro del dominio de aplicabilidad y los que tenían un valor alejado completamente del resto del conjunto que puedan afectar al modelo predictivo (Jiawei y Kamber 2001).

Fase 7.- Por último, se realizó una selección manual para eliminar los descriptores de conteo, con ayuda de gráficas de DL_{50} contra cada uno de los descriptores, ya que el uso de descriptores de conteo hace que los datos sean repetidos múltiples veces aportando mucho peso a un punto dentro del modelo.

3.3.2.3 – Generación de modelo predictivo QSAR y predicción.

En este paso se utilizó el programa WEKA para automatizar el cálculo de los modelos predictivos. Para poder llevar a cabo este proceso se siguieron las siguientes fases:

Fase 1.- Se dividió la base de datos en dos grupos, el grupo de entrenamiento con un porcentaje de 90% y el grupo de prueba con un porcentaje de 10%, seleccionando los compuestos de manera aleatoria.

Fase 2.- Con el grupo de entrenamiento se seleccionó el clasificador (o metodología) con la que se generó el modelo y se realizó la validación interna de este mediante las siguientes opciones (Bouckaert et al. 2016):

Usando el grupo de entrenamiento: El clasificador es evaluado en su eficiencia predictiva usando las instancias en las que fue entrenado.

Validación cruzada: El clasificador es evaluado utilizando la cantidad de grupos ingresados. En esta última, la información ingresada es dividida aleatoriamente en 10 partes, proporcionalmente iguales. Durante el proceso, una de las partes se

mantiene a la vez, mientras que las otras nueve partes pasan por el esquema del entrenamiento de aprendizaje. Por lo tanto, el procedimiento de aprendizaje se ejecuta un total de 10 veces en diferentes conjuntos de entrenamiento. Finalmente, las 10 estimaciones de error se promedian para arrojar la estimación de error global (Witten y Frank 2005).

Porcentaje de separación: em donde el clasificador es evaluado en su capacidad predictiva para un determinado porcentaje de los datos que se mantienen para la prueba. La cantidad depende del valor ingresado.

Fase 3.- Después de haber generado el modelo se realizó una validación externa mediante un grupo de prueba que no fue usado en la generación del modelo, mediante la adición de un grupo de prueba

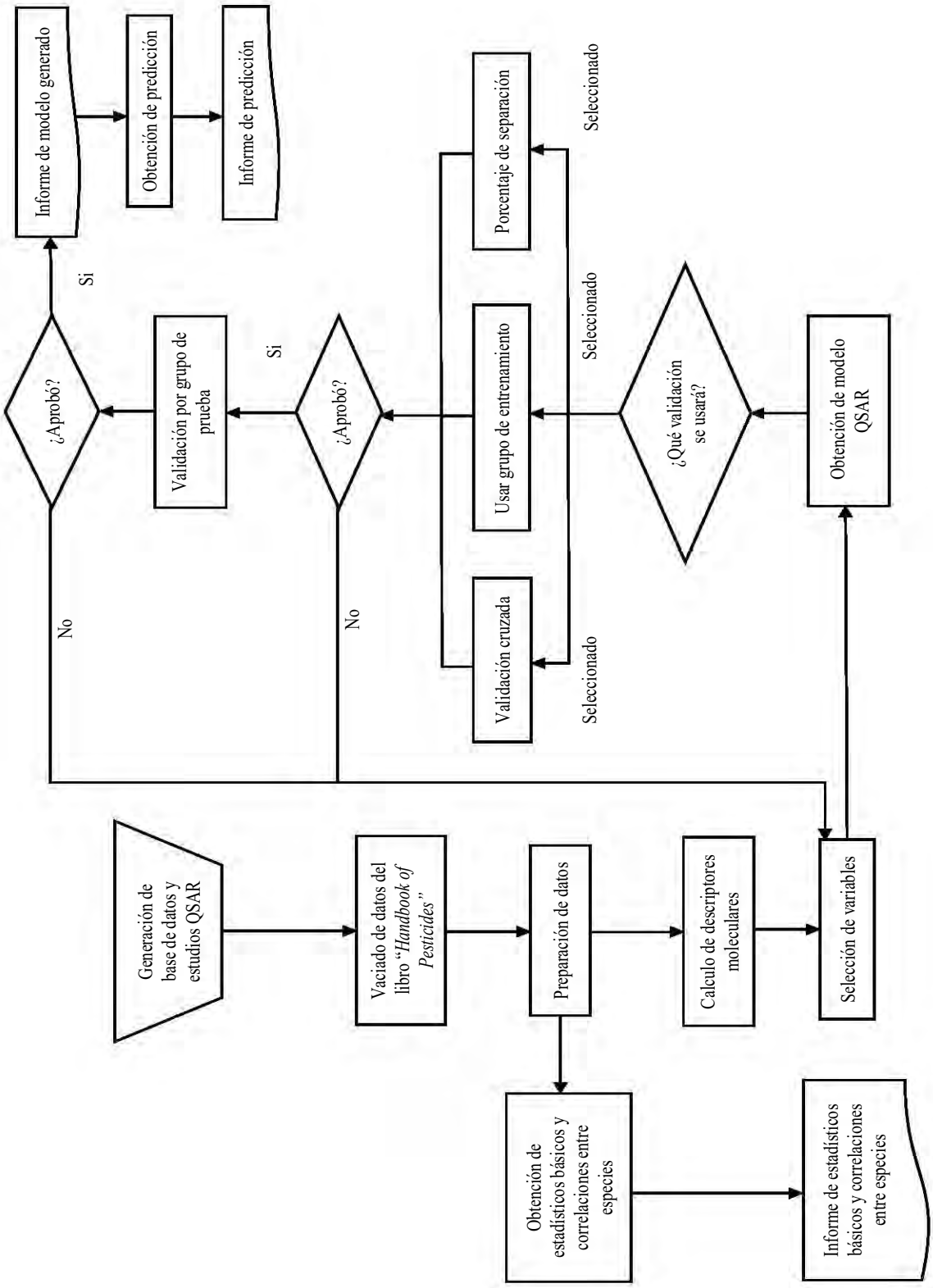
En este se evalúa la capacidad predictiva del clasificador usando un conjunto de datos externos y diferentes al conjunto de entrenamiento.

Para evaluar los modelos generados se hace uso de parámetros estadísticos que son los coeficientes de validación usando el grupo de entrenamiento (q^2) y el coeficiente de validación usando el grupo de prueba (r^2). Valores de estos dos coeficientes mayores a 0.5 y 0.6 respectivamente, indican que los modelos y la predicción son aceptables (Golbraikh et al. 2003).

Fase 4 – En esta última fase se calcularon los descriptores de un pesticida ejemplo (con DL_{50} conocido) correspondientes al modelo anteriormente generado, posteriormente se realizó la predicción de DL_{50} usando el modelo QSAR y se calculó el error relativo porcentual para comparar el DL_{50} experimental del predicho. El error relativo porcentual se calculó mediante la siguiente ecuación (Michael y Timothy 2005):

$$\%Error = \left(\frac{\left|DL_{50\ experimental} - DL_{50\ predicho}\right|}{\left|DL_{50\ experimental}\right|}\right) * 100 \quad (25)$$

3.4 Diagrama de Flujo



Capítulo IV

Resultados y Discusión de Resultados

4.1 Generación de la base de datos

La base de datos incluyó inicialmente 158 moléculas correspondientes a pesticidas que fueron recopiladas del libro *"CRC Handbook of Pesticides".* La tabla generada contenía cinco columnas de identificación: Estructura, CAS RN, Nombre Común, Sinónimos y número en el índice de Merck (**Figura 2**). Las siguientes cinco columnas fueron destinadas para las propiedades fisicoquímicas encontradas en el libro: Punto de Fusión (°C), Punto de Ebullición (°C), Densidad (g/ml), Solubilidad (g/L) y coeficiente de partición Octanol – Agua (**Figura 2**).

					Propiedades Fisicoquimicas				
Estructura	CAS RN	Nombre Común	Sinonimos	Merck Index No.	Punto de Fusion (PF) (°C):	Punto de Ebullicion (PE):	Densidad (g/ml):	Solubilidad:	Octanol/Water PC:
1	3383-96-8	Abate	Phodphorothioic ac	9057	30 - 30.5	NA	1.32	Insol. Water	80900
COP(=S)(OC)OC1=CC=C(SC2=CC=C(OP(=S)(OC)OC)C=C2)C=C1			Phosphorothioic acid, O,O' -(thiodi-p-phenylene) O,O,O',O' -tetramethyl ester (8CI)					hexane	
COP(=S)(OC)OC1=CC=C(SC2=CC=C(OP(=S)(OC)OC)C=C2)C=C1			Difenphos					sol. Acetronitrile, carbon tetrachloride, diethyl et	
COP(=S)(OC)OC1=CC=C(SC2=CC=C(OP(=S)(OC)OC)C=C2)C=C1			Temephos						
COP(=S)(OC)OC1=CC=C(SC2=CC=C(OP(=S)(OC)OC)C=C2)C=C1									
COP(=S)(OC)OC1=CC=C(SC2=CC=C(OP(=S)(OC)OC)C=C2)C=C1									
COP(=S)(OC)OC1=CC=C(SC2=CC=C(OP(=S)(OC)OC)C=C2)C=C1									
COP(=S)(OC)OC1=CC=C(SC2=CC=C(OP(=S)(OC)OC)C=C2)C=C1									
2	2227-13-6	Tetrasul	Sulfide, p-chloroph	NA	NA	NA	NA	NA	NA
ClC1=CC=C(SC2=CC(Cl)=C(Cl)C=C2Cl)C=C1			p-Chlorophenyl 2,4,5-trichlorophenyl sulfide (ACN)						
ClC1=CC=C(SC2=CC(Cl)=C(Cl)C=C2Cl)C=C2			Animert						
ClC1=CC=C(SC2=CC(Cl)=C(Cl)C=C2Cl)C=C3			Animert V-10						
3	30560-19-1	Acephate	Phosphoramidothi	26	82-93	NA	1.35	790g/L water 20°C	NA
CSP(OC)(NC(C)=O)=O			Acetylphosphoramidothioic acid est		impuro			151 g/L acetone	
CSP(OC)(NC(C)=O)=O			Orthene		64-68			>100 g/L ethanol	
CSP(OC)(NC(C)=O)=O								35g/L ethyl acetate	
CSP(OC)(NC(C)=O)=O								16 g/L benzene	
CSP(OC)(NC(C)=O)=O								0.1 g/L hexane	
4	15972-60-8	Alachlor (ACN)	Acetamide, 2-chlor	193	40-41	100	1.133	140 mg/L water 23°C	NA

Figura 2. Vista parcial de la base de datos PESTIMEP

Por último, los datos de DL_{50} (mg/kg) y de CL_{50} (mg/m^3) encontrados se capturaron en 465 columnas separadas por especie animal y a su vez subdivididas en diferentes vías de administración (las abreviaturas de las especies y vías de administración se incluyen en el Apendice B): orl, ihl, skn, ice, ims, ipr, ivn, ocu, par, scu y unr (**Figura 3**).

Rat										
orl [mg/kg]	ihl	skn	ice	ims	ipr	ivn	ocu	par	scu	unr
1000		1370			912				2302	8600
8600										
13000										
3960					6810					

Figura 3. Vías de administración de las evaluaciones toxicológicas en rata contenidas en la base de datos PESTIMEP: oral, inhalatoria, piel, intracerebral, intramuscular, intraperitoneal, intravenosa, ocular, parenteral, subcutáneo, urinaria

Para cada pesticida de creó una estructura 2D en ChemAxon Marvin Sketch. Todas las estructuras fueron verificadas en bases de datos electrónicas tales como Chemicalize, Chemspider y Pubchem ("Chemicalize - Instant Cheminformatics Solutions" s/f, "ChemSpider | Search and share chemistry" s/f, "The PubChem Project" s/f) para confirmar su correcta representación (**Figura 4**). Una vez generada la base de datos PESTIMEP se realizó el curado de datos, en el cual se corrigieron los errores de unidades hallados en las evaluaciones toxicológicas, se separó la información de DL_{50} de CL_{50}, debido a que para este trabajo la DL_{50} de las especies es el principal objeto de interés por tener mayor número de datos.

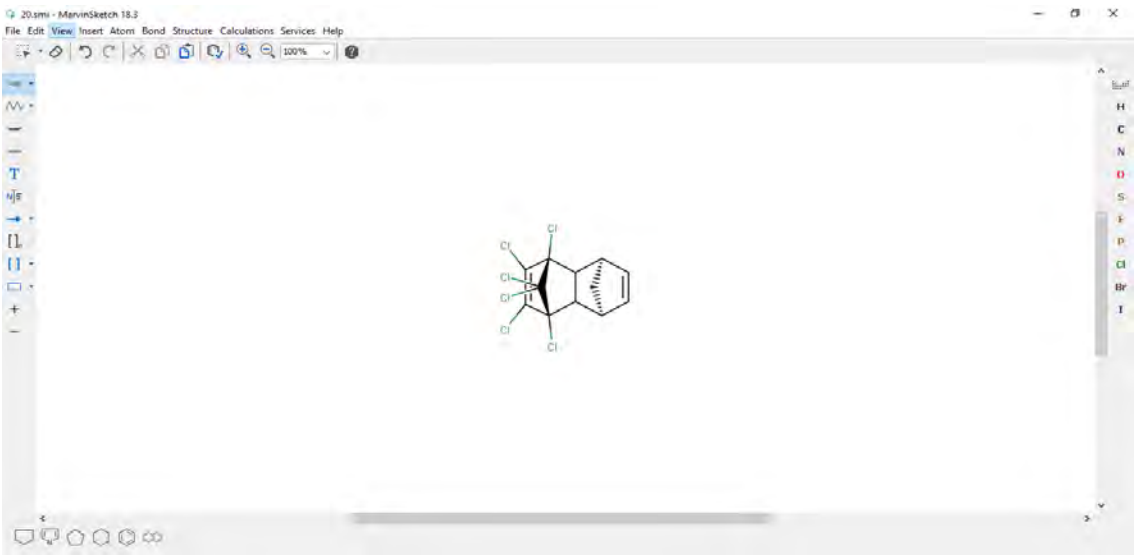


Figura 4. Estructura 2D del pesticida Aldrin construida en ChemAxon Marvin Sketch

4.2 Correlaciones entre especies

Posterior al curado de la base de datos, se hizo un conteo de evaluaciones toxicológicas con el fin de establecer la columna objetivo para la evaluación de las bases de datos. El resultado de este conteo arrojó 24 evaluaciones con 10 o más datos toxicológicos (ver **Tabla 2**), siendo las cinco más representativas: rata oral con 146 datos, seguido por ratón oral con 101, conejo piel con 83, rata piel con 77 y ratón intraperitoneal con 55. Al contener mayor número de datos para poder ser analizados, se tomó la evaluación de rata vía oral como la columna objetivo. Con la columna objetivo ya seleccionada se realizó un procedimiento de depuración donde se eliminaron aquellos compuestos que no tuvieran información de DL_{50} en rata vía oral, reduciendo así la base de 158 compuestos a 146.

Tabla 2. Evaluaciones toxicológicas con una cantidad de datos mayor o igual a 10, tambien se muestran los valores promedio, mínimo. máximo y desviación estandar.

Variable	Estadísticas Descriptivas				
	Valid N	Promedio	Mínimo	Máximo	Dev.Std.
orl rat	146	1963.312	0.79000	10001.0	2411.41
skn rat	77	2436.352	2.40000	23000.0	3282.75
ipr rat	39	766.592	0.28000	6810.0	1516.28
ivn rat	12	21.789	0.30000	87.0	30.46
scu rat	23	1835.642	0.27900	15001.0	3539.12
unr rat	32	2035.417	0.93000	10001.0	2737.60
orl mus	101	1569.472	0.30000	15001.0	2402.06
skn mus	13	3681.231	8.00000	10001.0	3694.38
ipr mus	55	688.131	0.83000	6811.0	1216.35
ivn mus	15	75.712	0.20000	320.0	91.55
scu mus	22	2570.879	0.25000	23800.0	5991.36
unr mus	14	1064.143	27.00000	4490.0	1321.29
orl dck	26	1564.969	0.60000	11300.0	3008.59
orl rbt	46	1152.739	10.00000	7100.0	1506.38
skn rbt	83	3339.611	4.70000	22601.0	3761.66
orl cat	12	262.750	2.00000	802.0	231.44
orl bwd	24	43.938	0.75000	400.0	83.88
orl ckn	25	1780.896	8.00000	7951.0	2692.77
orl mam	14	2322.286	10.00000	12600.0	3418.97
unr mam	33	1749.303	40.00000	10000.0	2222.03
orl dog	30	1645.263	3.00000	10001.0	2739.41
orl pgn	10	30.216	2.37000	110.0	41.90
orl qal	29	1938.273	1.20000	16001.0	4129.34
orl gpg	40	8419.565	2.30000	300000.0	47300.65

Como se observa en la **Tabla 2** la mayoría de las evaluaciones toxicológicas presenta una dispersión de datos, evidenciado por la desviación estándar, mayor a 1000, exceptuando la evaluación rata intravenosa, ratón intravenoso, gato oral, ave salvaje oral y pichón oral, en donde el número de datos es menor a 30. La desviación estándar en las evaluaciones toxicológicas muestra qué tan alejados se encuentran los datos del promedio de valores de DL_{50} y no debe confundirse con desviación estándar debida a error experimental. El hecho de tener desviaciones estándar muy grandes en las evaluaciones con mayor número de datos, en comparación a las que contienen menor cantidad, denota la variabilidad en la información toxicológica de estas evaluaciones.

Para hacer un análisis más completo de la dispersión de los datos en las diferentes evaluaciones, se indica en las **Figuras 5 y 6** mediante histogramas y diagramas de cajas, los cuales fueron generados en el programa Statistica StatSoft TIBCO y Spotfire TIBCO. En estos histogramas se puede apreciar cómo a pesar de tener una gran dispersión en los datos, la mayor parte de estos se encuentra distribuido con mayor frecuencia en intervalos de 0 a 1000 mg/kg y de 0 a 2000 mg/kg, por lo que se esperaría que los datos correlacionados se encontrarán en estos intervalos. La afirmación previa puede ser corroborada con el uso de los diagramas de caja, en los cuales se observa que la mediana es menor al promedio de DL_{50} en cada especie, La interpretación estadística sería que, si la mediana es menor al promedio de un conjunto de datos presenta una asimetría del tipo positiva, es decir que los datos se agrupan en mayor proporción a la izquierda (cuadril 1 a 2) y la dispersión más grande se encuentra a la derecha (cuadril 2 a 3). Trasladando esta interpretación a la base de pesticidas PESTIMEP, esta asimetría es de tipo positiva por tener mayor agrupación en intervalos de 0 a 1000 mg/kg y de 0 a 2000 mg/kg (Paolo 2003).

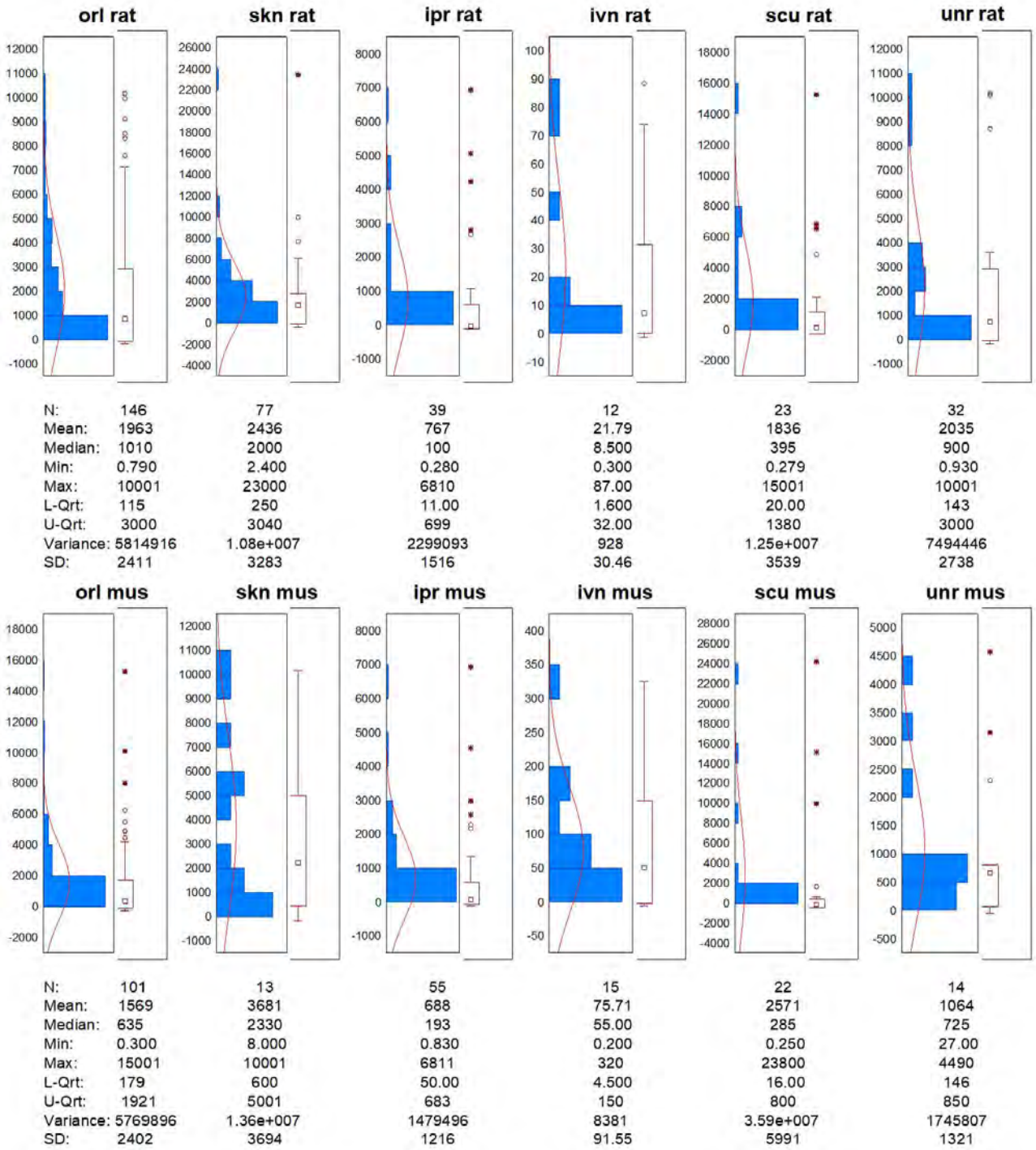


Figura 5. Histogramas y diagramas de caja de las evaluaciones toxicológicas: rata oral, rata piel, rata intraperitoneal, tata intravenosa, rata subcutaneo, rata sin registro, ratón oral, ratón piel, ratón intraperitoneal, ratón intravenosa, ratón subcutaneo , ratón sin registro

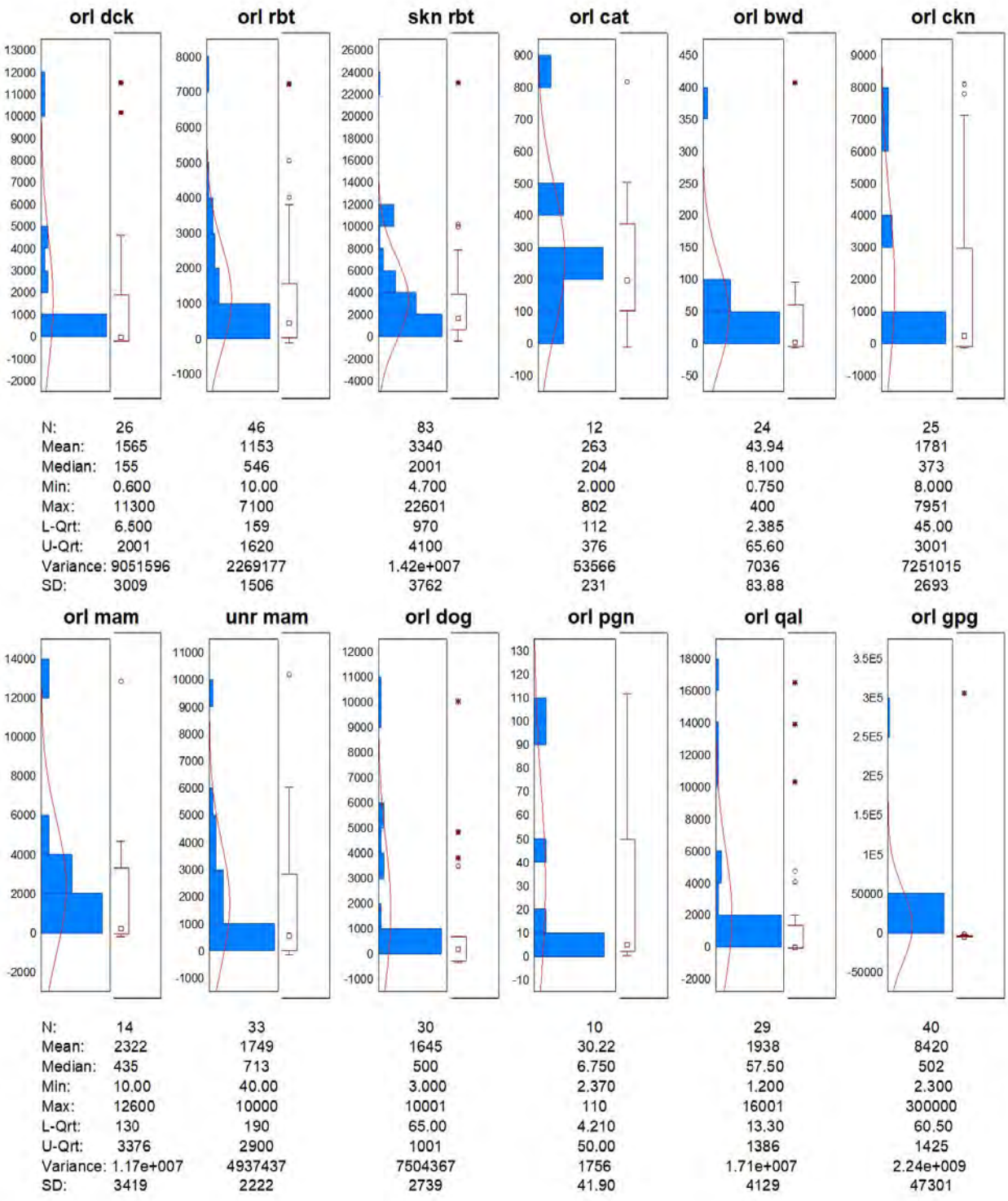


Figura 6. Histogramas y diagramas de caja de las evaluaciones toxicológicas: pato oral, conejo oral, conejo piel, gato oral, pajaro salvaje oral, pollo oral, mamífero oral, mamifero sin registro, perro oral, pichón oral, codorniz oral, conejillo de indias oral

En contraste con este análisis, los diagramas de caja también sirven para delimitar el uso de la base de pesticidas en la obtención de modelos predictivos QSAR, pues proporciona un panorama real de la distribución de los datos que, en principio permitiría saber dentro de qué intervalos de toxicidad funcionan los modelos predictivos. Por ejemplo, si se desea usar el conjunto datos de la evaluación rata oral para generar una predicción, solo puede ser usado para predecir toxicidades en un intervalo de 115 a 3000 mg/kg (pesticidas moderadamente y poco tóxicos según la tabla de toxicidad de la OMS incluida en la **Apéndice C**) por tener una concentración mayor de datos dentro de los cuadriles del diagrama de caja. Para ilustrar mejor cómo se relacionan las evaluaciones entre las distintas especies, se calculó una matriz de correlación en donde los ejes *x* y *y* son las evaluaciones en la base de pesticida con mayor cantidad de información toxicológica (**Figura 7**).

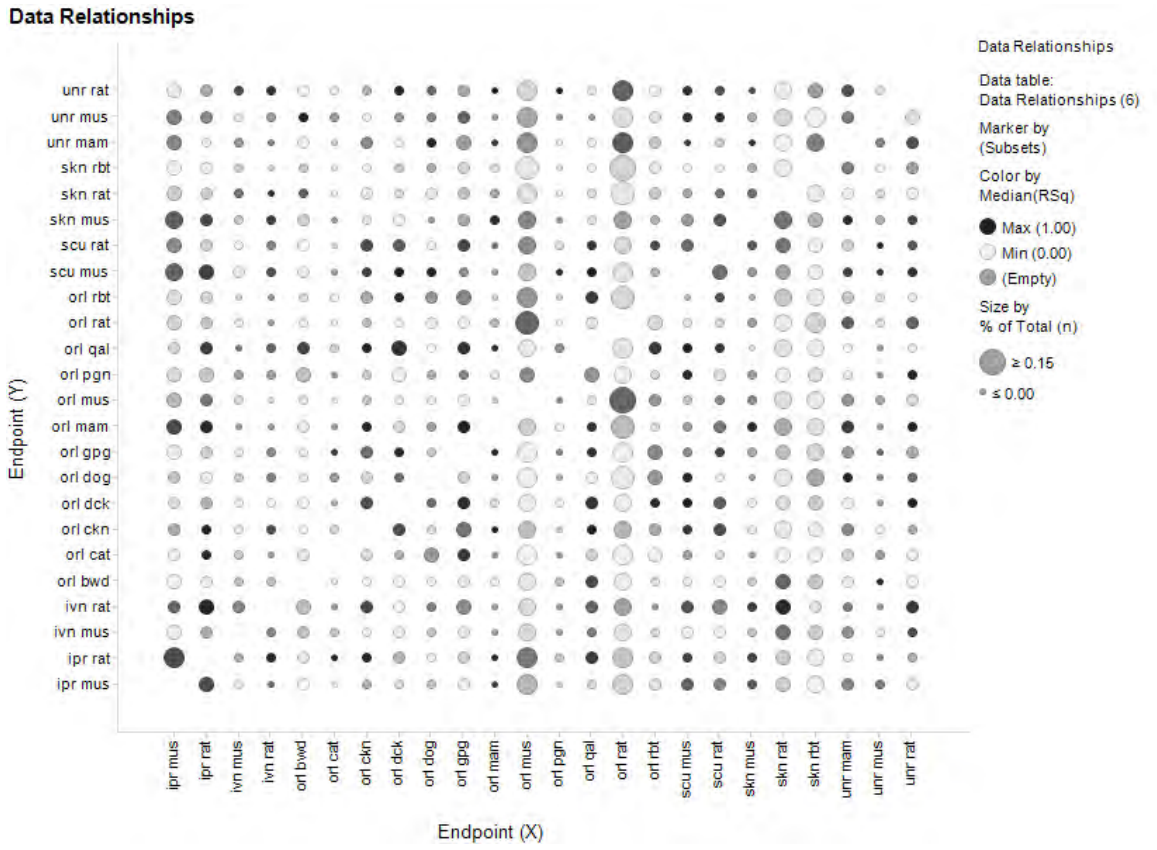


Figura 7. Matriz de correlación de las evaluaciones toxicológicas con una cantidad mayor o igual a 10

En esta matriz se visualiza la cantidad de datos correlacionados mediante el uso de áreas circulares; entre mayor sea el área, la cantidad de datos relacionados será más grande. También es posible observar el grado de correlación entre los valores contenidos dentro de la circunferencia, ya que entre más obscura sea ésta tendrá valores con una correlación cercana a 1. De este modo se observa que los puntos con más valores correlacionados son los que se relacionan con la evaluación rata oral, no obstante, el punto con mejor correlación y número de datos es el de rata oral con ratón oral. Por otra parte, los puntos con mejor correlación en la matriz son los de pato oral con ratón subcutáneo y rata intraperitoneal con pollo oral.

Con los puntos más relevantes de la matriz de correlaciones anteriormente mencionados se hicieron regresiones lineales para conocer el comportamiento de estas correlaciones, las cuales pueden ser vistas en las **Figuras 8, 9 y 10.**

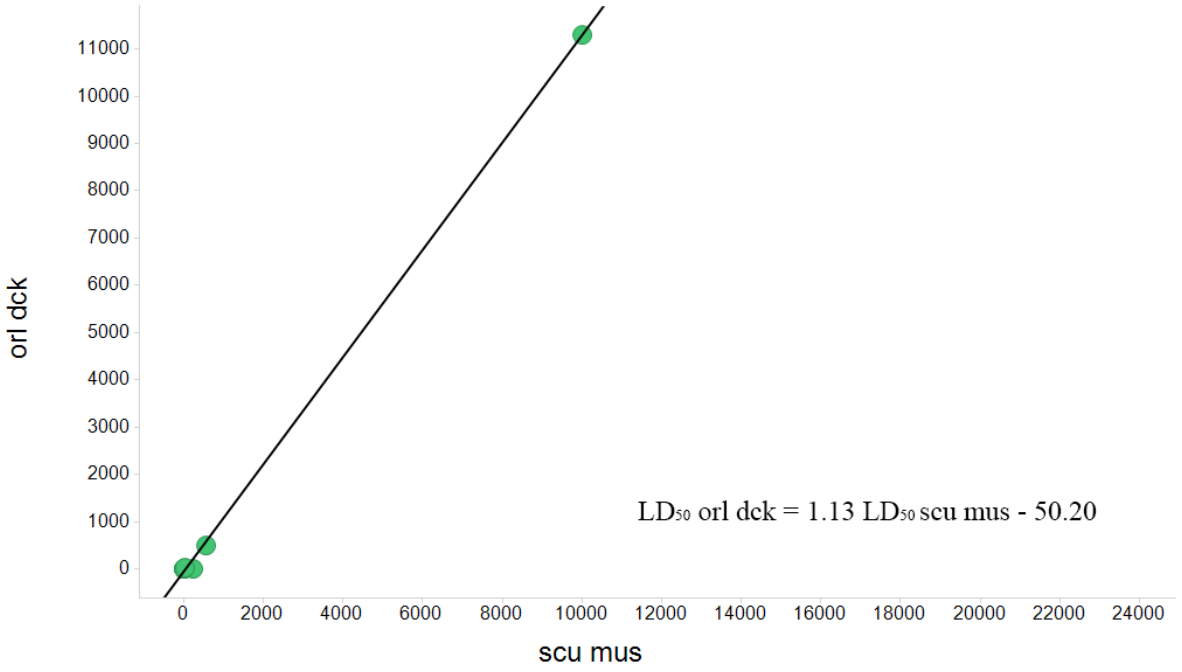


Figura 8. Regresión lineal entre la evaluación pato oral y subcutáneo ratón, se incluye la ecuación de regresión

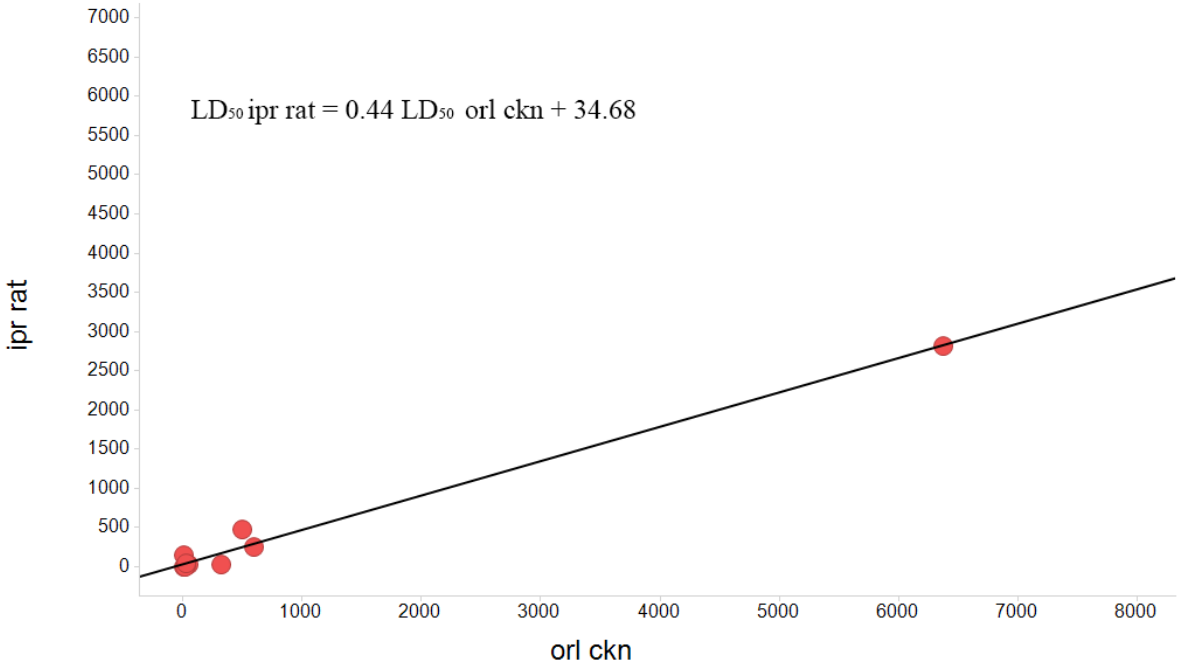


Figura 9. Regresión lineal entre la evaluación rata intraperitoneal y pollo oral, con ecuación de regresión

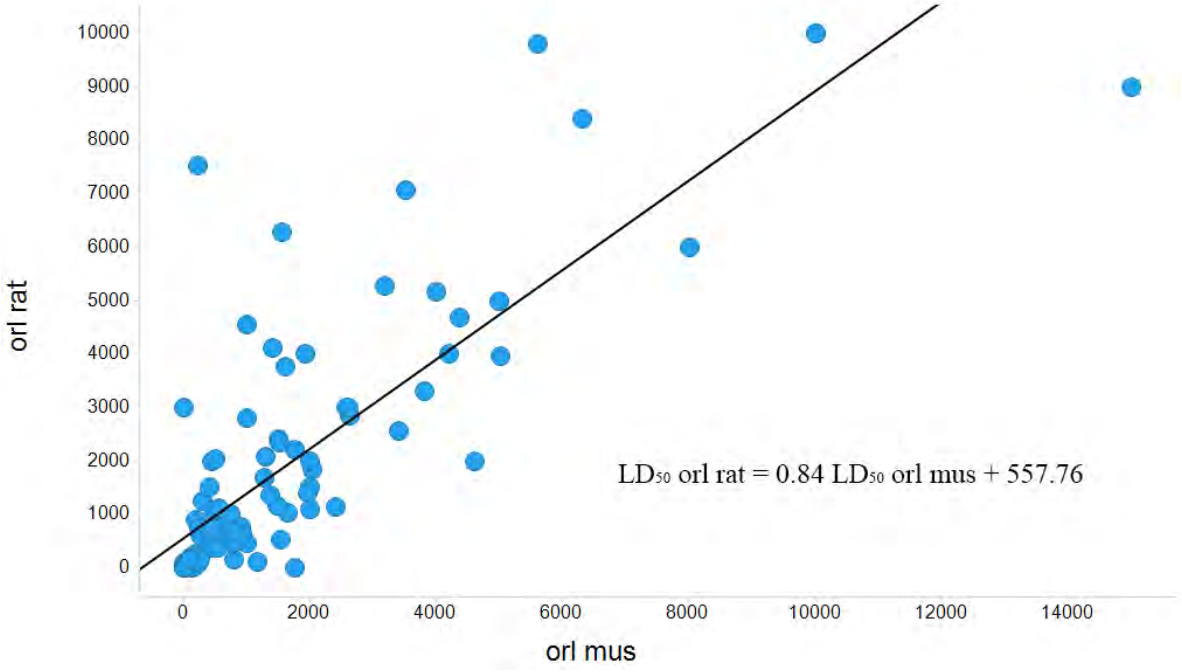


Figura 10. Regresión lineal entre la evaluación rata oral y ratón oral, con ecuación de regresión

En la regresión lineal de las evaluaciones pato oral *vs.* ratón subcutáneo se puede apreciar que la mayor parte de los datos se correlaciona en un intervalo de 0 a 1000 mg/kg en pato oral con el intervalo 0 a 2000 mg/kg en ratón subcutáneo, teniendo un valor de correlación muy alejado del conjunto de datos con coordenadas [10000,11000], lo que hace que su valor de r^2 sea de 0.9995. Sin embargo, el número de datos en este cálculo es de tan solo siete (ver **Tabla 3**), lo que hace que una interpolación entre estas evaluaciones no sea tan confiable. Esto último se puede apreciar especialmente entre los valores [501, 545] y [10000,11000].

Tabla 3. Regresiones lineales entre las especies con mejor R^2 y la de mayor número de datos correlacionados

Endpoint (Y)	Endpoint (X)	R^2	R	N
orl dck	scu mus	0.999496	0.999748	7
ipr rat	orl ckn	0.986449	0.993202	9
orl rat	orl mus	0.665019	0.815487	101

Este comportamiento es parecido en la regresión lineal rata intraperitoneal *vs.* pollo oral en la cual la mayor parte de los datos se encuentran ubicados en intervalos de 0 a 1000 mg/kg en ambas evaluaciones, teniendo un punto correlacionado alejado del conjunto en las coordenadas [6375, 2823] teniendo un valor de $r^2 = 0.9864$ esta vez con un número de datos de nueve. Mientras que en la regresión lineal rata oral *vs.* ratón oral la distribución de los datos es más dispersa. Para este par de evaluaciones los datos se encuentran correlacionados en los intervalos de 0 a 2000 mg/kg en la evaluación ratón oral y de 0 a 1000 mg/kg en la evaluación rata oral, con un valor de $r^2 = 0.6650$, usando un conjunto de datos de ciento uno. En este caso la interpolación entre estas evaluaciones es más confiable a pesar de contar con un valor de r^2 menor, pues al tener una cantidad más grande de datos es posible establecer de mejor manera la tendencia en la distribución de los datos. Sin embargo, es importante mencionar que es necesario hacer otro tipo de ajuste

estudios para describir de manera más adecuada el comportamiento en la distribución de la toxicidad de los pesticidas contenidos en la base de datos.

Finalmente, es importante conocer el comportamiento que tienen la correlación entre las evaluaciones toxicológicas a medida que aumenta o disminuye el número de relaciones que se establecen entre estos datos. Esto se puede analizar en la **Figura 11** en la cual, por medio de una comparación entre el gráfico de barras de la mediana de r^2 y el gráfico de barras de la mediana de cantidad de datos correlacionados entre cada especie, como se observa en la **Figura 11**. En esta comparación se aprecia que las evaluaciones que establecen más relaciones con otras evaluaciones son las de ratón intraperitoneal, ratón oral, rata oral, rata piel y conejo piel, mientras que las evaluaciones con menos relaciones de datos entre especies son ratón intravenoso, rata intravenosa, gato oral, mamífero oral, pichón oral, ratón subcutáneo, ratón piel y rata sin registro. No obstante, las evaluaciones con mayor número de relaciones entre especie son las que tiene valores de correlación menor, en contraste con esto las evaluaciones con menor número de relaciones tienen valores de r^2 mayores.

Este comportamiento se debe principalmente a que las evaluaciones con más relaciones entre especies cuentan con mayor disponibilidad de datos que pueden usarse para establecer interpolaciones, los cuales suelen estar más dispersos. Por otra parte, las evaluaciones que establecen menos relaciones disponen de menor cantidad de datos, a su vez estos datos se agrupan en intervalos pequeños haciendo que la dispersión en las correlaciones que establecen con otras evaluaciones sea menor y por lo tanto haciendo que los valores de r^2 sean más grandes en comparación con las evaluaciones con mayor cantidad de datos.

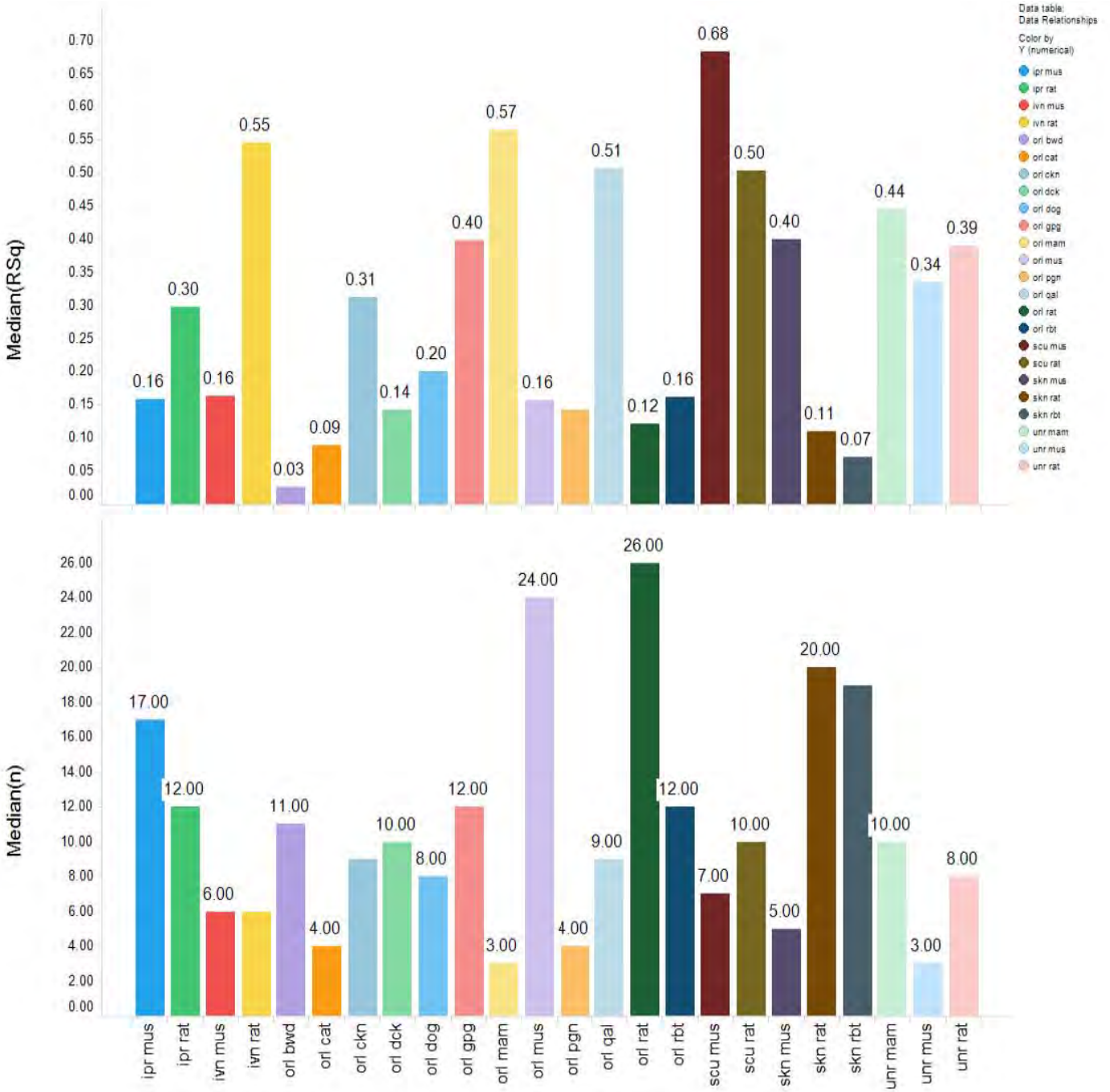


Figura 11. Graficas de barras de los promedios de R^2 y promedio de datos correlacionados de las evaluaciones toxicológicas

4.3 Modelo predictivo QSAR y predicción

Después de haber curado la base de datos, se prepararon las estructuras usando el programa Molecular Operating Environment (MOE, v.16.08). En este proceso se depositaron todas las moléculas contenidas en la base de pesticidas que se encontraban en 2D y fueron transformadas a 3D (**Figura 12**). Posteriormente se limpiaron las estructuras y se optimizaron utilizando el campo de fuerza MMFF, con la opción de conservación de la quiralidad. Con las estructuras limpias y optimizadas se trasladaron al programa Dragon Descriptor, en el cual se hizo el cálculo de los descriptores moleculares 2D y 3D. El resultado de este cálculo arrojó 3502 descriptores, de los cuales se conservaron 1200 al quitar todos aquellos que no contenían valores disponibles

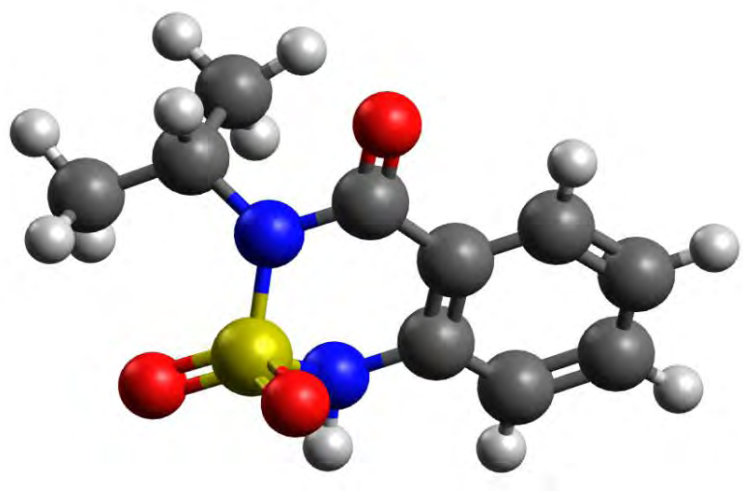


Figura 12. Estructura tridimensional del pesticida Bentazon después de haber sido optimizada y conservada su quiralidad

Antes de obtener los modelos predictivos con los descriptores calculados se definió el dominio de aplicabilidad y se seleccionaron las variables necesarias para el modelo QSAR. Inicialmente se contaba con 1200 descriptores disponibles, sin embargo, no todos pueden ser usados para generar un modelo predictivo debido a que no todos estos descriptores tienen una aportación significativa en el cálculo de dicho modelo. El dominio de aplicabilidad está definido por las características

estructurales de los pesticidas y su toxicidad. La selección de variables está delimitada estadísticamente por la regla de Topliss y Costello la cual consiste en minimizar las correlaciones fortuitas causadas por un sobreajuste del modelo predictivo al tener una relación de máximo 5 moléculas por cada descriptor (John G. Topliss y Costello 1972). Para poder realizar la selección de variables y definir del dominio de aplicabilidad se utilizó el programa Knime, el cual permite automatizar este proceso por medio de nodos a los cuales se les puede asignar tareas específicas. El total de descriptores se pasó por un nodo de correlación lineal en el cual se hizo una matriz de correlación entre descriptores y posteriormente se separaron los descriptores con mayor correlación a través de un filtro de correlación.

Para este filtro se tomaron aleatoriamente valores de correlación con los que se diferenció la información original; este proceso se realizó de manera independiente tres veces, obteniendo los resultados mostrados en la **Tabla 4**. De estas correlaciones se seleccionó la de 0.7 por que este permite trabajar con una mayor cantidad de descriptores, quitando los que presentan una alta correlación.

Tabla 4. Descriptores moleculares calculados originalmente con Dragon Descriptor, descriptores después de quitar aquellos sin valor numérico e introducidos a Knime antes de aplicar cualquier filtro y finalmente descriptores obtenidos luego de ser filtrados

Programa utilizado	Tratamiento		Número de descriptores
Dragon	Descriptores calculados originalmente		3502
Knime		Descriptores con valores numéricos	1200
	Filtro de Correlación	0.9	741
		0.7	342
		0.5	154
	Filtro de Baja Varianza	(0.891)	75
		Filtros de Conteo	25

Dado que el modelo fue satisfactorio no se requirió aplicar el mismo proceso en la selección del valor de correlación. La siguiente parte en la selección de variables consistió en quitar todos los descriptores que fueran altamente homogéneos con un filtro de baja varianza. Para este paso se usó un método iterativo, calculado que el mejor valor para el filtro de baja varianza es de 0.891. Después de aplicar este filtro de baja varianza se redujeron los descriptores de 342 a 75. Posteriormente, se normalizó la información y se agregó la columna de DL_{50} de la evaluación rata oral, seleccionada en la primera etapa de la evaluación de la base de datos, para pasar a la última parte del proceso donde quedó definido el dominio de aplicabilidad y se quitaron los descriptores de conteo.

En este último paso se quitaron todos los pesticidas que se encontraran fuera del dominio de aplicación estructural o biológico, tales como el Boro y todos los pesticidas cuyo valor de DL_{50} fueran superiores a 7000 mg/kg (Jiawei y Kamber 2001). Estos últimos corresponden a valores atípicos de la evaluación rata oral en la **Figura 5**, dejando así 137 pesticidas de los 146 iniciales. Además, se eliminaron los descriptores de conteo, que al igual que los descriptores correlacionados sobre ajustan los modelos: número de átomos de carbono, número de átomos de azufre, número de átomos de nitrógenos, entre otros. Con este último paso se redujo a un total de 25 descriptores moleculares disponibles para generar el modelo predictivo QSAR. En la **Figura 13** se muestra la hoja de trabajo en Knime.

En la **Figura 14** se representa la matriz final de correlación de los 137 pesticidas que se usaron para la construcción del modelo predictivo. Como se observa la correlación entre los descriptores es mínima, después de haber pasado por todo el proceso de selección, tomando como objeto de comparación la línea diagonal de la matriz. Esto permite que el modelo QSAR generado con los 25 descriptores y las 137 moléculas no se encuentre sobre ajustado. La lista completa de los descriptores moleculares finales se muestra en la **Tabla 5**.

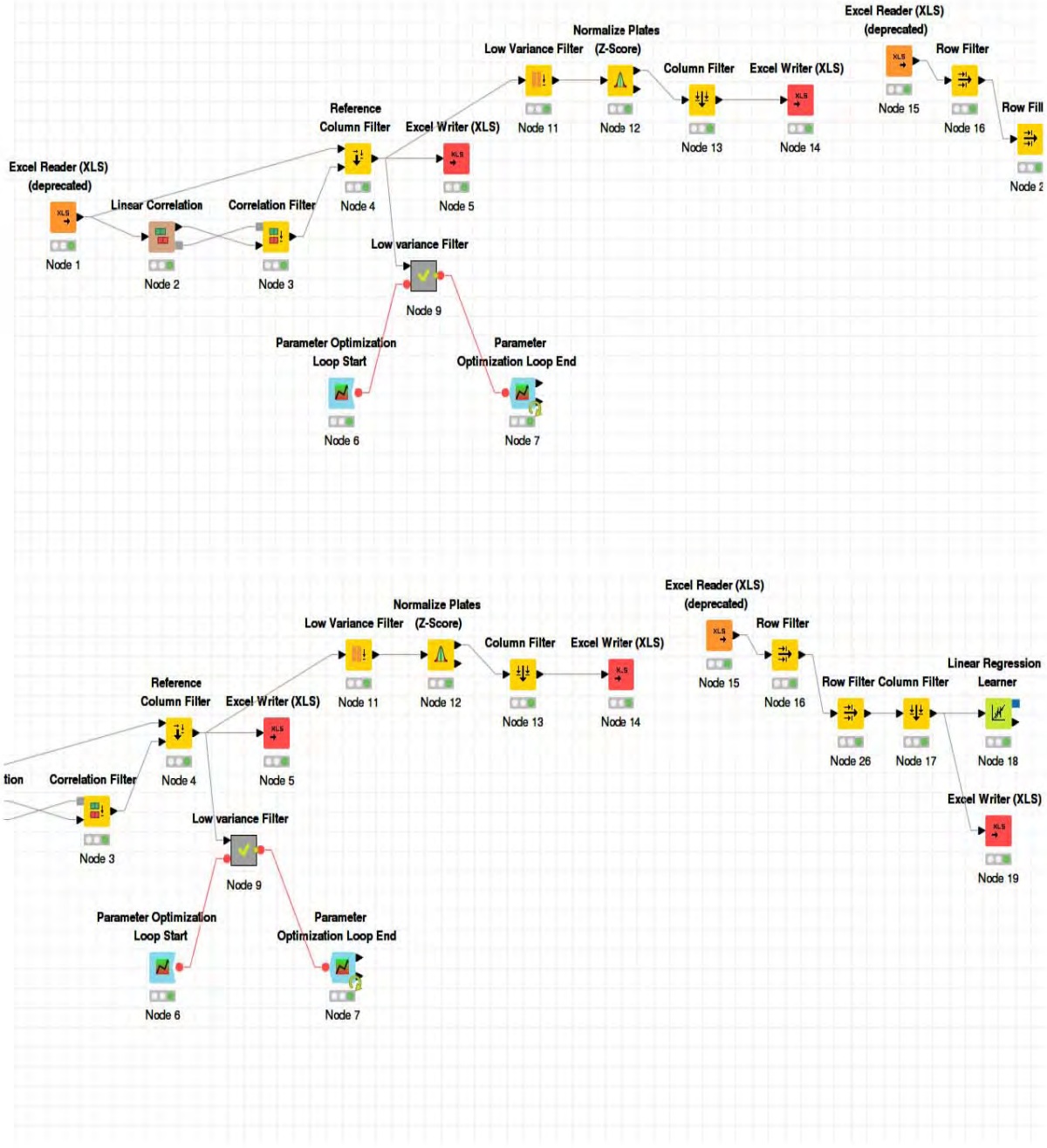


Figura 13. Hoja de trabajo en Knime para la selección de variables usadas en el modelo QSAR

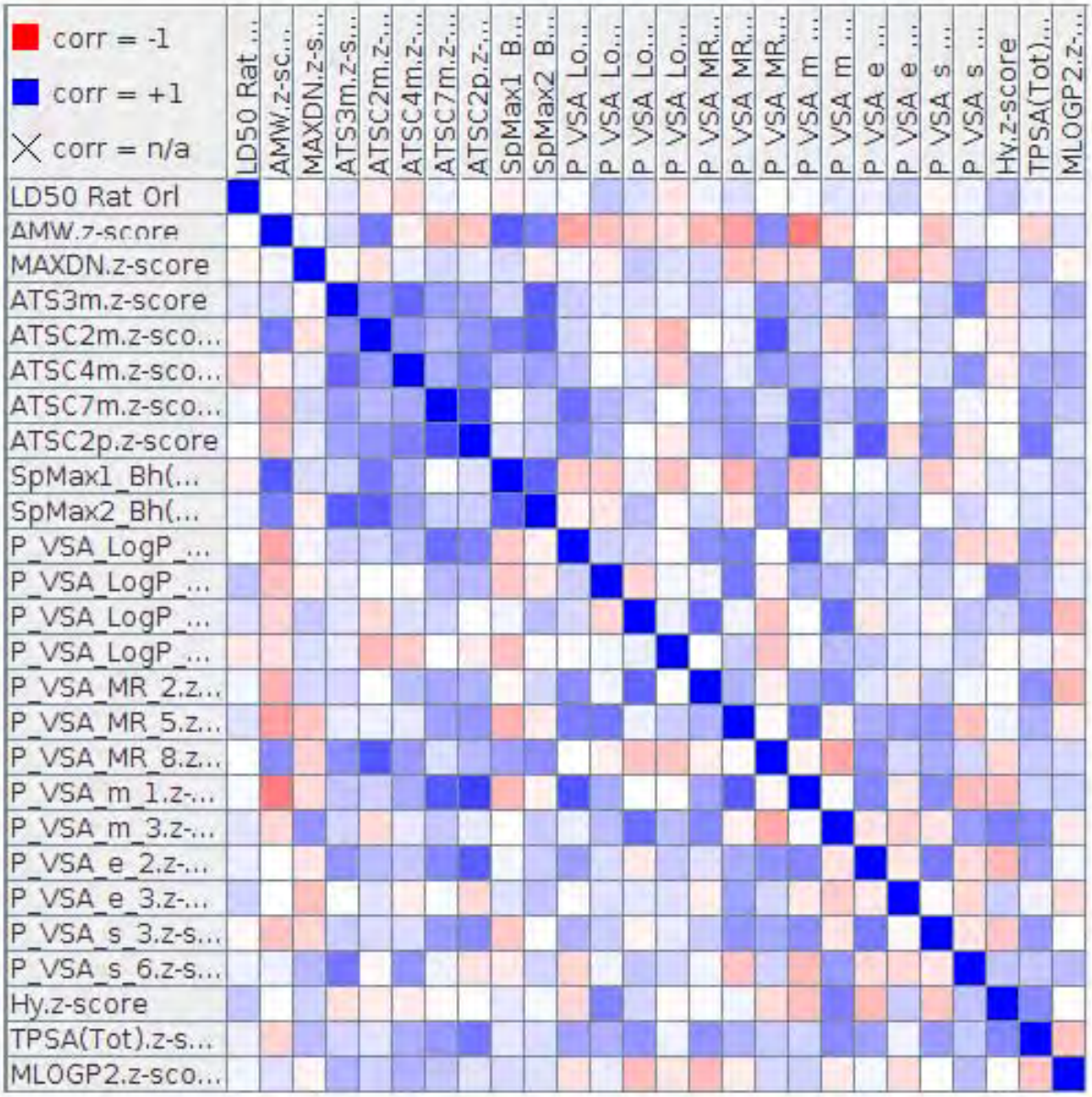


Figura 14. Matriz de correlaciones de los descriptores depués de ser filtrados en KNIME

Tabla 5. Abreviaturas de los descriptores moleculares seleccionados con KNIME

Abreviatura	Descriptor
AMW	Peso Molecular Promedio
MAXDN	Variación electrotopológica negativa máxima
ATS3m	Autocorrelación Broto-Moreau de retraso 3 (función logarítmica) ponderada por masa
ATSC2m	Autocorrelación centrada de Broto-Moreau de retraso 2 ponderado por masa
ATSC4m	Autocorrelación centrada de Broto-Moreau de retraso 4 ponderado por masa
ATSC7m	Autocorrelación centrada de Broto-Moreau de retraso 7 ponderado por masa
ATSC2p	Autocorrelación centrada de Broto-Moreau de retraso 2 ponderada por polarizabilidad
SpMax1_Bh(m)	Valor propio más grande n. 1 de la matriz de Burden ponderada por masa
SpMax2_Bh(m)	Valor propio más grande n. 2 de la matriz de Burden ponderada por masa
P_VSA_LogP_1	P_VSA-like en LogP, bin 1
P_VSA_LogP_2	P_VSA-like en LogP, bin 2
P_VSA_LogP_4	P_VSA-like en LogP, bin 4
P_VSA_LogP_5	P_VSA-like en LogP, bin 5
P_VSA_MR_2	P_VSA-like en Refractividad Molar, bin 2
P_VSA_MR_5	P_VSA-like en Refractividad Molar, bin 5
P_VSA_MR_8	P_VSA-like en Refractividad Molar, bin 8
P_VSA_m_1	P_VSA-like en masa, bin 1
P_VSA_m_3	P_VSA-like en masa, bin 3
P_VSA_e_2	P_VSA-like en Sanderson electronegatividad, bin 2
P_VSA_e_3	P_VSA-like en Sanderson electronegatividad, bin 3
P_VSA_s_3	P_VSA-like en estado I, bin 3
P_VSA_s_6	P_VSA-like en estado I, bin 6
Hy	Factor Hidrofóbico
TPSA(Tot)	Área de superficie polar topológica utilizando contribuciones polares N, O, S, P
MLOGP2	Coeficiente de partición de octanol-agua de Moriguchi al cuadrado (logP^2)

La distribución en los descriptores de los 137 pesticidas después de ser filtrados se puede aprecia a través de los histogramas y diagramas de caja en las **Figuras 15 y 16**.

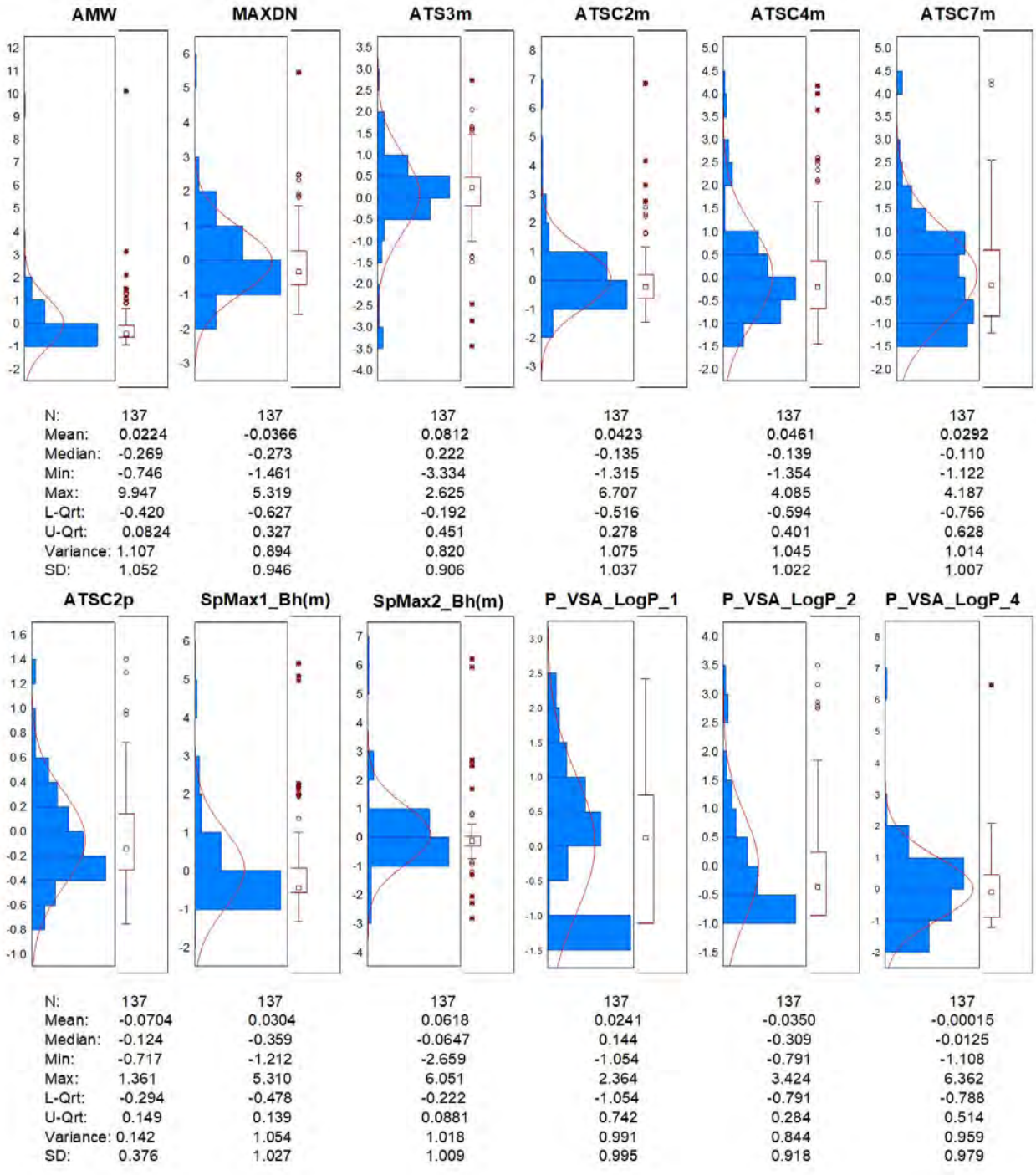


Figura 15. Histogramas y diagramas de caja de los descriptores moleculares: AMW, MAXDN, ATS3m, ATSC2m, ATSC4m, ATSC7m, ATSC2p, SpMax1_Bh(m), SpMax2_Bh(m), P_VSA_LogP_1, P_VSA_LogP_2, P_VSA_LogP_4

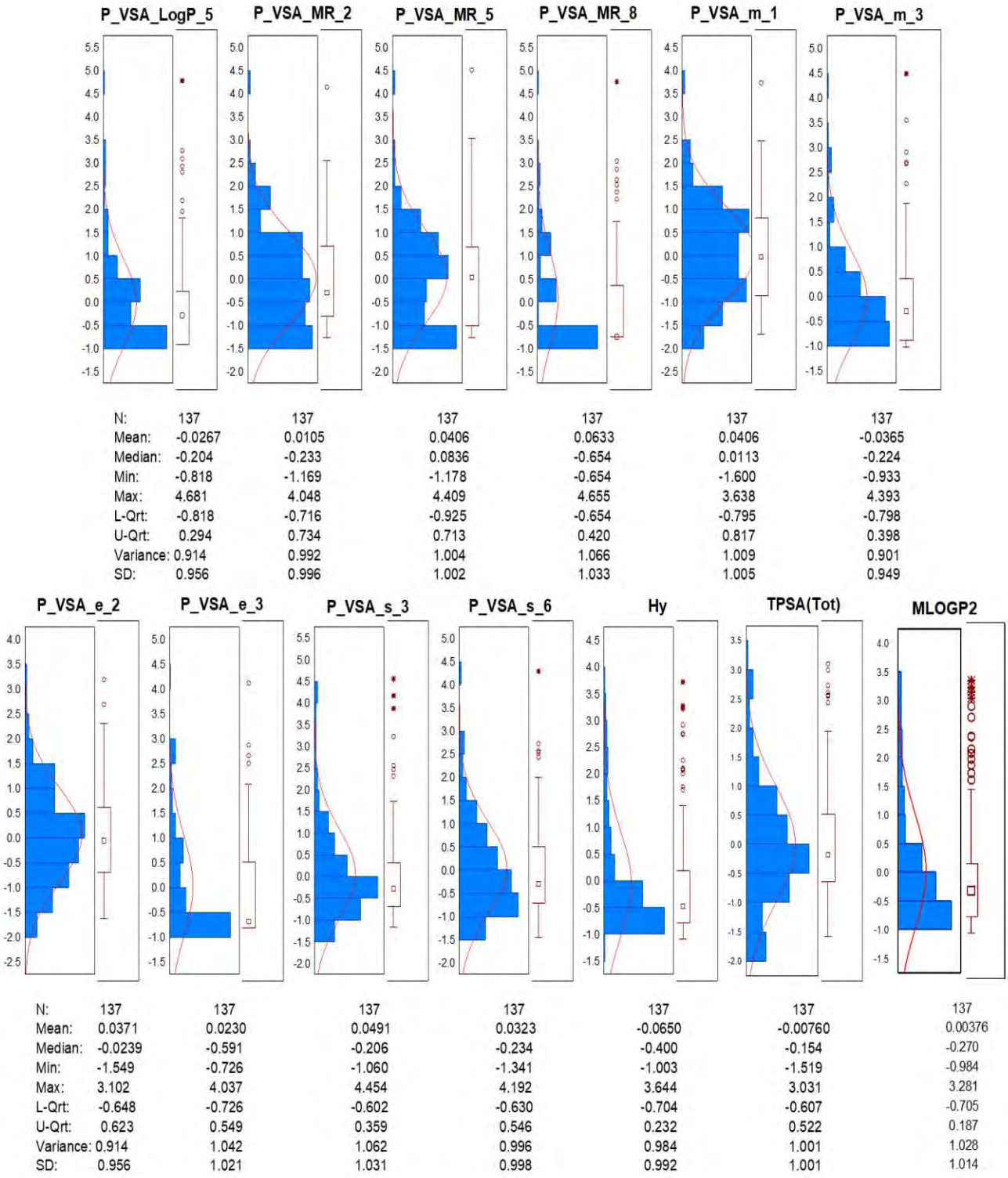


Figura 16. Histogramas y diagramas de caja de los descriptores moleculares: P_VSA_LogP_5, P_VSA_MR_2, P_VSA_MR_5, P_VSA_MR_8, P_VSA_m_1, P_VSA_m_3, P_VSA_e_2, P_VSA_e_3, P_VSA_s_3, P_VSA_s_6, Hy, TPSA(Tot), MLOGP2

En los histogramas de frecuencias se puede observar que la mayor parte de los descriptores presentan una distribución gaussiana con excepción de los descriptores ATSC7m, P_VSA_LogP_1, P_VSA_MR_5, P_VSA_MR_8 y P_VSA_m_1 los cuales tienen un comportamiento asimétrico, no obstante, al ser analizados en los diagramas de caja y bigotes se puede observar que son los que tienen menor número de puntos atípicos. Este comportamiento se derivar de la naturaleza de los valores de los descriptores y a que la amplitud entre los intervalos de cada clase es pequeña, haciendo que sea más difícil obtener distribuciones gaussianas. Los puntos atípicos encontrados en mayor proporción en las distribuciones normales no afectan al modelo predictivo por tener una desviación estándar menor a 2, por lo que ser utilizados para generar el modelo QSAR.

El último paso para la obtención del modelo predictivo, después de haber seleccionado los pesticidas y descriptores, fue la partición de los pesticidas en dos grupos: el grupo de entrenamiento con el 90% de la información y el grupo de prueba con el 10% restante de información (Dobbin y Simon 2011). La partición de los datos en los dos grupos y la obtención del modelo se hicieron en el programa WEKA que tiene la capacidad de generar varios modelos predictivos tanto numéricos como nominales.

Para la separación de grupos se usó un filtro de re-muestreo debido a que este paso permite hacer una separación aleatoria sin repetir datos, teniendo al final 123 pesticidas para el grupo de entrenamiento y 14 para el grupo de prueba (**Tabla 6**). Con el grupo de entrenamiento se aplicó el clasificador red neuronal artificial para generar el modelo predictivo. Este tipo algoritmos es ampliamente utilizado para la generación de modelos predictivos, ya que identifica un error, lo propaga por toda la red hasta que ubica el error e intenta corregir todo el error en la red neuronal (el modelo final se puede ver en la **Figura 17**). Sin embargo, es importante mencionar que no se utilizó un método lineal convencional debido a que estos son sencillos e interpretables, siendo estos los que se analizaron en la primera etapa de evaluación de la base de datos.

Tabla 6. Distribución de datos para generación de modelo

Pesticidas sin datos atípicos	137
Datos grupo de entrenamiento	Datos grupo de prueba
123	14

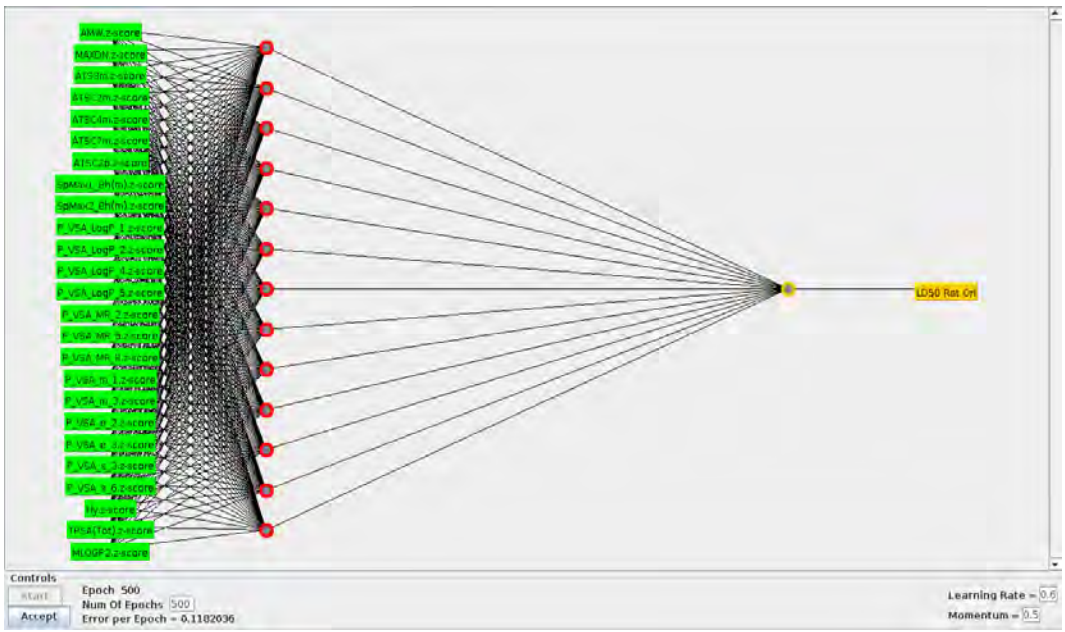


Figura 17. Modelo de red neuronal artificial usado con el que se obtuvo el modelo de QSAR en WEKA

Una vez generado el modelo predictivo, este fue evaluado con el grupo de prueba. El modelo QSAR que se obtuvo consta de 13 nodos o núcleos con un peso sigmoide por cada descriptor que representa numéricamente las interacciones entre estos (el modelo completo con el valor de cada nodo se encuentra en el **Anexo A**). El modelo tuvo una evaluación con el grupo de entrenamiento (q^2) de 0.8253 y una evaluación con el grupo de prueba (r^2) de 0.7714 (ver **Tabla 7** y **8**), Estos valores están dentro de los límites aceptables para un modelo predictivo ($r^2 > 0.6$ y $q^2 > 0.5$) según establece la Agencia de Protección Ambiental de los Estados Unidos de América (EPA) e implementado en el desarrollo de su software *Herramienta Computacional para la Estimación de Toxicidad* (T.E.S.T. por sus siglas en inglés) (T. Martin 2013).

Tabla 7. Evaluación del modelo predictivo usando el grupo de entrenamiento (q^2) con 123 pesticidas y los errores en la bondad de los descriptores con el modelo predictivo.

Evaluación interna con el grupo de entrenamiento	
Coeficiente de Correlación (r^2)	0.8253
Error medio absoluto	710.9713
Error cuadrático medio	959.7207
Error relativo absoluto	50.9747 %
Raíz cuadrada del error cuadrático medio	57.2203 %
Número total de estancias	123

Tabla 8. Evaluación del modelo predictivo usando el grupo de prueba (r^2) con 14 pesticidas y los errores en la bondad de los descriptores con el modelo predictivo.

Evaluación externa con el grupo de prueba	
Coeficiente de Correlación (r^2)	0.7714
Error medio absoluto	868.7011
Error cuadrático medio	1315.3445
Error relativo absoluto	65.6498 %
Raíz cuadrada del error cuadrático medio	65.2691 %
Número total de estancias	14

Sin embargo, es importante reconocer que una de las grandes limitaciones del modelo QSAR se ve reflejado en los porcentajes elevados de error relativo absoluto y la raíz cuadrada del error cuadrático medio que son de 50.97% y 57.22% respectivamente utilizando el grupo de entrenamiento, con el cual se generó el modelo predictivo. Mientras que este error aumenta a 65.65% y 65.23% en la evaluación de la capacidad predictiva del modelo, usando el grupo de prueba. Los errores anteriores reflejan la bondad del ajuste de correlación de cada uno de los 25 descriptores con el valor de DL_{50}. Por otra parte, es importante puntualizar que este modelo creado en WEKA tiene 3562 entradas de datos numéricos. Lo anterior hace que obtener de un modelo predictivo con la capacidad de correlacionar todos los valores de entrada tanto de descriptores moleculares como de información toxicológica es una labor complicada, que a su vez se refleja en los errores de ajuste, así como en los valores de r^2 y q^2.

Por último, se eligió un pesticida ejemplo aleatoriamente cuya toxicidad fuera conocida y se encuentra en los intervalos de predicción del modelo predictivo (115 a 3000 mg/kg determinado en la **Figura 5**), en este caso se tomó el pesticida Propiconazole con un valor de 1517 mg/kg en la evaluación de DL_{50} rata oral. Para obtener la toxicidad predicha de este pesticida se limpió la estructura, se optimizó utilizando el campo de fuerza MMFF y se conservó su quiralidad de este compuesto. Posterior a este proceso, se calcularon los 25 descriptores del modelo predictivo (**Tabla 5**), con los descriptores y los pesos sigmoides del modelo se realizó el tratamiento matemático correspondiente para cada uno de los 13 nodos del modelo.

El tratamiento matemático realizado para cada nodo consistió en aplicar la función lineal y la función tangente hiperbólica a lo largo de la neurona y sumar cada uno de los valores obtenidos de estas funciones. Antes de realizar el tratamiento matemático en los nodos, primero se verificó que estos aportaran información en la determinación toxicológica. Lo anterior se realizó al sumar todos los pesos multiplicados por los valores de los descriptores en los 13 nodos, la cual debe superar el valor del umbral establecido por el programa WEKA (**Anexo A**) en cada nodo para activarse y poder ser usado en la determinación de DL_{50}. Posterior a la

verificación, en la cual todos los nodos superaron su valor de umbral, se aplicó la función lineal, así como la función tangente hiperbólica a todos los descriptores del nodo tomando en cuenta la siguiente restricción para la selección de la función: sí el valor de la función lineal es mayor a 500 mg/kg se elige la función tangente hiperbólica, debido a que valores mayores a 500 mg/kg en uno o más nodos puede hacer que la determinación salga del intervalo de predicción del modelo (115 – 3000 mg/kg). En caso de obtener predicciones inferiores o superiores al intervalo de predicción, aun después de aplicar la restricción, es necesario realizar una determinación experimental. El resultado de la predicción del modelo arrogó un valor de DL_{50} 1221.52 ± 151.96 mg/kg (**Tabla 9**), este resultado da un error de 19.48 % en la predicción (**Tabla 10**), el cual se encuentra en el intervalo esperado de error por los valores de validación con el grupo de entrenamiento y grupo de prueba (q^2 y r^2).

Tabla 9. Valores de DL_{50} orl rat (mg/Kg) para cada nodo del modelo predictivo

Nodos para la predicción del pesticida Propiconazole	DL_{50} orl rat (mg/kg)
Nodo 1	134.03
Nodo 2	7.95
Nodo 3	130.36
Nodo 4	15.07
Nodo 5	329.85
Nodo 6	490.29
Nodo 7	5.99
Nodo 8	83.25
Nodo 9	3.73
Nodo 10	3.40
Nodo 11	2.02
Nodo 12	-5.07
Nodo 13	20.64
Total	1221.52 ± 151.96

Tabla 10. Error entre el valor de DL_{50} experimental y predicho

Error en la toxicidad predicha	
DL_{50}(mg/kg) orl rat experimental	1517
DL_{50}(mg/kg) orl rat predicción	1221.52 ± 151.96
% de Error de la predicción	19.48

Capítulo V

Conclusiones

5.1 Conclusiones

En el presente trabajo se generó un repositorio digital de pesticidas reportados en el libro “CRC Handbook of Pesticides”. PESTIMEP, como se denominó a esta base de datos, contiene información de 158 pesticidas con determinaciones toxicológicas (DL_{50} y CL_{50}) en diferentes especies animales.

PESTIMEP representa una herramienta que permite analizar correlaciones de toxicidad interespecie, debido a la gran diversidad de evaluaciones toxicológicas en vías de administración y organismos evaluados. El análisis de estas correlaciones sugiere que aquellas que presentan valores altos de r^2 carecen de información que describa de una mejor manera la tendencia de cambio en la toxicidad de una especie a otra; siendo aquellas con mayor número de datos las que mejor describen este cambio. No obstante, estudios adicionales son necesarios para ayudar a describir de una mejor manera la toxicidad interespecie.

Mediante el uso de redes neuronales artificiales se generó un modelo predictivo de toxicidad, con parámetros de validación interna y externa (r^2 y de q^2, respectivamente) superiores a los establecidos por la EPA, lo que permite utilizar este modelo en la predicción de toxicidad aguda de pesticidas con un nivel de confianza aceptable. En este sentido, el modelo generado se aplicó para predecir la toxicidad de un pesticida ejemplo, la cual tuvo un error de predicción del 19.48%.

Para finalizar, el uso de modelos QSAR en el área de evaluaciones toxicológicas representa una alternativa regulatoria rápida y de bajo costo. La construcción de los modelos predictivos QSAR se hace a partir del conocimiento de estructuras químicas y de información experimental biológica, fisicoquímica y ambiental, por lo que pueden ser utilizados en la industria farmacéutica, alimenticia, textil, química e incluso en la predicción de condiciones climatológicas específicas de una región.

Referencias

- Avdeef, Alex, David A. Barrett, P. Nicholas Shaw, Roger D. Knaggs, y Stanley S. Davis. 1996. “Octanol-, chloroform-, and propylene glycol dipelargonat-water partitioning of morphine-6-glucuronide and other related opiates”. *Journal of Medicinal Chemistry* 39 (22): 4377–81.

- Balakin, Konstantin V., Sean Ekins, Andrey Bugrim, Van A. Ivanenkov, Dmitry Korolev, Yuri V. Nikolsky, Andrey A. Ivashchenko, Nikolay P. Savchuk, y Tatiana Nikolskaya. 2004. “Quantitative structure-metabolism relationship modeling of metabolic N-dealkylation reaction rates”. *Drug Metabolism and Disposition* 32 (10): 1111–20.

- Benigni, Romualdo, y Ann M. Richard. 1996. “QSARs of mutagens and carcinogens: Two case studies illustrating problems in the construction of models for noncongeneric chemicals”. *Mutation Research - Genetic Toxicology*.

- Bolton, Evan E., Yanli Wang, Paul A. Thiessen, y Stephen H. Bryant. 2008. “Chapter 12 PubChem: Integrated Platform of Small Molecules and Biological Activities”. *Annual Reports in Computational Chemistry*.

- Bouckaert, Remco R, Eibe Frank, Mark Hall, Richard Kirkby, Peter Reutemann, Alex Seewald, y David Scuse. 2016. “WEKA Manual for Version 3-8-1”. *University of Waikato*, 341.

- Cassano, Antonio, Alberto Manganaro, Todd Martin, Douglas Young, Nadège Piclin, Marco Pintore, Davide Bigoni, y Emilio Benfenati. 2010. “CAESAR models for developmental toxicity”. *Chemistry Central Journal* 4 (1): S4.

- Charton, Marvin. 1969. “Organic and Biological Chemistry: The Nature of the ortho Effect. II. Composition of the Taft Steric Parameters”. *Journal of the American Chemical Society*.

- “Chemical Substances - CAS REGISTRY”. s/f. Consultado el 29 de marzo de 2018. http://support.cas.org/content/chemical-substances.

- “Chemicalize - Instant Cheminformatics Solutions”. s/f. Consultado el 15 de abril de 2018. https://chemicalize.com/.

- “ChemSpider | Search and share chemistry”. s/f. Consultado el 15 de abril de 2018. http://www.chemspider.com/.

- Cherkasov, Artem, Eugene N Muratov, Denis Fourches, Alexandre Varnek, Igor I Baskin, Mark Cronin, John C Dearden, et al. 2014. “Perspective QSAR Modeling : Where have you been ? Where are you going to ? QSAR Modeling : Where have you been ? Where are you going to ?” *Journal of Medicinal Chemistry*.

- Clark, Robert D., Wenkel Liang, Adam C. Lee, Michael S. Lawless, Robert Fraczkiewicz, y Marvin Waldman. 2014. “Using beta binomials to estimate classification uncertainty for ensemble models”. *Journal of Cheminformatics* 6 (1).

- Contrera, Joseph F., Edwin J. Matthews, y R. Daniel Benz. 2003. “Predicting the carcinogenic potential of pharmaceuticals in rodents using molecular structural similarity and E-state indices”. *Regulatory Toxicology and Pharmacology* 38 (3): 243–59.

- Cramer, Richard D., David E. Patterson, y Jeffrey D. Bunce. 1988. “Comparative Molecular Field Analysis (CoMFA). 1. Effect of Shape on Binding of Steroids to Carrier Proteins”. *Journal of the American Chemical Society* 110 (18): 5959–67.

- David, Anderson, Sweeney Dennis, y Williams Thomas. 2011. *Statistics for Business and Economics*. Mason, Ohio: South-Western Cengage Learning.

- Dobbin, Kevin K., y Richard M. Simon. 2011. “Optimally splitting cases for training and testing high dimensional classifiers”. *BMC Medical Genomics* 4 (1). BioMed Central Ltd: 31.

- Dowdy, Shirley, Stanley Weardon, y Daniel Chilko. 2005. *Statistics for Research*. *Wiley*.

- Durant, G. J., J. C. Emmett, C. R. Ganellin, P. D. Miles, M. E. Parsons, H. D. Prain, y G. R. White. 1977. “Cyanoguanidine-Thiourea Equivalence in the Development of the Histamine H2-Receptor Antagonist, Cimetidine”. *Journal of Medicinal Chemistry* 20 (7): 901–6.

- “Environmental Information Sheets (EIS) - Voluntary Initiative”. s/f. Consultado el 29 de marzo de 2018. https://voluntaryinitiative.org.uk/resources/eis/.

- Eriksson, Lennart, Joanna Jaworska, Andrew P. Worth, Mark T.D. Cronin, Robert M. McDowell, y Paola Gramatica. 2003. “Methods for reliability and uncertainty assessment and for applicability evaluations of classification- and regression-based QSARs”. *Environmental Health Perspectives*.

- Fieser, Louis F., Martin G. Ettlinger, y George Fawaz. 1948. “Naphthoquinone Antimalarials. XV. Distribution between Organic Solvents and Aqueous Buffers”. *Journal of the American Chemical Society* 70 (10): 3228–32.

- Fujita, Toshio, Junkichi Iwasa, y Corwin Hansch. 1964. “A New Substituent Constant, ir, Derived from Partition Coefficients”. *Journal of the American Chemical Society* 86 (23): 5175–80.

- Golbraikh, Alexander, Min Shen, Zhiyan Xiao, Yun De Xiao, Kuo Hsiung Lee, y Alexander Tropsha. 2003. “Rational selection of training and test sets for the development of validated QSAR models”. *Journal of Computer-Aided Molecular Design* 17 (2–4): 241–53.

- Gold, Lois Swirsky, Thomas H. Slone, Bruce N. Ames, y Neela B. Manley. 2001. “Pesticide Residues in Food and Cancer Risk: A Critical Analysis”. *Handbook of Pesticide Toxicology*, 799–843.

- Goyer, R.A;, y T.W. Clarkson. 2001. *Toxic effects of metals. Casarett and Doull’s toxicology: The Basic Science of Poisons.*

- Gozalbes, R, Julián Ortiz, y Fito López. 2014. “Métodos computacionales en toxicología predictiva : aplicación a la reducción de ensayos con animales en el contexto de la legislación comunitaria REACH”. *Revista de Toxicología* 31 (2): 157–67.

- Hacker, Miles, Kenneth Bachmann, y William Messer. 2009. “Pharmacology Principles and Practice”, 109–11.

- Hall, Lowell H., y Lemont B. Kier. 2001. “Issues in representation of molecular structure: The development of molecular connectivity”. *Journal of Molecular Graphics and Modelling*.

- Hansch, Corwin. 1969. “A Quantitative Approach to Biochemical Structure-Activity Relationships”. *Accounts of Chemical Research* 2 (8): 232–39.

- Hansch, Corwin, Peyton P. Maloney, Toshio Fujita, y Robert M. Muir. 1962. “Correlation of biological activity of phenoxyacetic acids with Hammett substituent constants and partition coefficients”. *Nature* 194 (4824): 178–80.

- Jiawei, Han, y Micheline Kamber. 2001. *Data mining: concepts and techniques*. *San Francisco, CA, itd: Morgan Kaufmann*. Vol. 5.

- Jonh H., Duffus, y Worth Howard. 2007. *Fundamental Toxicology*. Editado por John Duffus y Howard Worth. Cambridge: Royal Society of Chemistry.

- Kaivosaari, Sanna, Moshe Finel, y Mikko Koskinen. 2011. “N-glucuronidation of drugs and other xenobiotics by human and animal UDP-glucuronosyltransferases”. *Xenobiotica*.

- Kauzmann, W. 1959. “Some Factors in the Interpretation of Protein Denaturation”. *Advances in Protein Chemistry* 14 (C): 1–63.

- Kiyoi, Takao, Julia M. Adam, John K. Clark, Keneth Davies, Anna-Marie Easson, Darren Edwards, Helen Feilden, et al. 2011. “Discovery of potent and orally bioavailable heterocycle-based cannabinoid CB1 receptor agonists”. *Bioorganic & Medicinal Chemistry Letters* 21 (6): 1748–53.

- Klopman, Gilles. 1984. “Artificial Intelligence Approach to Structure-Activity Studies. Computer Automated Structure Evaluation of Biological Activity of Organic Molecules”. *Journal of the American Chemical Society* 106 (24): 7315–21.

- Klopman, Gilles, y Shaomeng Wang. 1991. “A computer automated structure evaluation (CASE) approach to calculation of partition coefficient”. *Journal of Computational Chemistry* 12 (8): 1025–32.

- “KNIME Analytics Platform | KNIME”. s/f. Consultado el 30 de marzo de 2018. https://www.knime.com/knime-analytics-platform.

- Kubat, Miroslav. 1999. “Neural networks: a comprehensive foundation.” *The Knowledge Engineering Review*.

- Kutter, Eberhard, y Corwin Hansch. 1969. “Steric Parameters in Drug Design. Monoamine Oxidase Inhibitors and Antihistamines”. *Journal of Medicinal Chemistry* 12 (4): 647–52.

- Leeuwen, K. van, T. W. Schultz, T. Henry, B. Diderich, y G. D. Veith. 2009. “Using chemical categories to fill data gaps in hazard assessment”. *SAR and QSAR in Environmental Research* 20 (3–4): 207–20.

- Lewis, Kathy, y Andy Green. s/f. “Chemistry International -- Newsmagazine for IUPAC”. Consultado el 29 de marzo de 2018. https://www.iupac.org/publications/ci/2011/3303/ic.html.

- Martin, T. 2013. “User’s Guide for T.E.S.T. (version 4.2)”. *United States Environmental Protection Agency*, núm. EPA/600/R-16/058: 63.

- Martin, T. M., D. M. Young, C. R. Lilavois, y M. G. Barron. 2015. “Comparison of global and mode of action-based models for aquatic toxicity”. *SAR and QSAR in Environmental Research* 26 (3): 245–62.

- Martin, Todd M., Christopher M. Grulke, Douglas M. Young, Christine L. Russom, Nina Y. Wang, Crystal R. Jackson, y Mace G. Barron. 2013. “Prediction of aquatic toxicity mode of action using linear discriminant and random forest models”. *Journal of Chemical Information and Modeling* 53 (9): 2229–39.

- Martin, Todd M., Paul Harten, Raghuraman Venkatapathy, Shashikala Das, y Douglas M. Young. 2008. “A hierarchical clustering methodology for the estimation of toxicity”. *Toxicology Mechanisms and Methods* 18 (2–3): 251–66.

- Martin, Todd M., y Douglas M. Young. 2001. “Prediction of the acute toxicity (96-h LC50) of organic compounds to the fathead minnow (Pimephales promelas) using a group contribution method.” *Chemical Research in Toxicology* 14 (10): 1378–85.

- Michael, Evans, y Swartz Timothy. 2005. *Approximating Integrals via Monte Carlo and Deterministic Methods*. New York: Oxford University Press.

- Milan, Chiara, Onofrio Schifanella, Alessandra Roncaglioni, y Emilio Benfenati. 2011. “Comparison and possible use of in Silico tools for carcinogenicity within REACH legislation”. *Journal of Environmental Science and Health - Part C Environmental Carcinogenesis and Ecotoxicology Reviews* 29 (4): 300–323.

- Milne, George W. A. 1995. *CRC handbook of pesticides*. CRC Press.

- Miners, John O., Kathleen M. Knights, J. Brian Houston, y Peter I. Mackenzie. 2006. “In vitro-in vivo correlation for drugs and other compounds eliminated by glucuronidation in humans: Pitfalls and promises”. *Biochemical Pharmacology* 71 (11): 1531–39.

- “Molecular descriptors calculation - Dragon - Talete srl”. s/f. Consultado el 15 de abril de 2018. http://www.talete.mi.it/products/dragon_description.htm.

- Montgomery, Douglas C, Elizabeth A Peck, y G Geoffrey Vining. 2001. *Introduction to Linear Regression Analysis*. *Technometrics*. Vol. 49.

- Nisbet, Robert, John Elder IV, y Gary Miner. 2009. *Handbook of Statistical Analysis and Data Mining Applications*.

- Paolo, G. 2003. *Applied data mining: statistical methods for business and industry*. Wiley.

- Polanski, Jaroslaw. 2009. “Receptor dependent multidimensional QSAR for modeling drug-receptor interactions.” *Current medicinal chemistry* 16 (25): 3243–57.

- Romesburg, H. Charles. 1984. *Cluster Analysis for Researchers*. Belmont, CA: Lifetime Learning Publications.

- Rumelhart, D. E., J. L. McClelland, y The PDP Research Group. 1988. “Parallel distributed processing, explanations in the micro structure of cognition, 1: Foundations”. *A Bradford Book*, 576.

- Sarangapani, Jagannathan. 2006. *Neural Network Control of Nonlinear Discrete-Time Systems*. *Taylor & Francis Group*.

- Sundriyal, Sandeep, Smriti Khanna, Rikta Saha, y Prasad V. Bharatam. 2008. "Metformin and glitazones: Does similarity in biomolecular mechanism originate from tautomerism in these drugs?" *Journal of Physical Organic Chemistry* 21 (1): 30–33.

- Taft, R. 1956. "Separation of Polar, Steric, and Resonance Effects in Reactivity". En *In Steric Effects in Organic Chemistry*, p 556. New York: Wiley.

- Tetko, Igor V., Yurii Sushko, Sergii Novotarskyi, Luc Patiny, Ivan Kondratov, Alexander E. Petrenko, Larisa Charochkina, y Abdullah M. Asiri. 2014. "How accurately can we predict the melting points of drug-like compounds?" *Journal of Chemical Information and Modeling* 54 (12): 3320–29.

- "The PubChem Project". s/f. Consultado el 15 de abril de 2018. https://pubchem.ncbi.nlm.nih.gov/.

- Todeschini, Roberto, y Viviana Consonni. 2010. *Molecular Descriptors for Chemoinformatics*. *Molecular Descriptors for Chemoinformatics*. Vol. 2.

- Topliss, J G, y R P Edwards. 1979. "Chance factors in studies of quantitative structure-activity relationships." *Journal of medicinal chemistry* 22 (10): 1238–44.

- Topliss, John G., y Robert J. Costello. 1972. "Chance Correlations in Structure-Activity Studies Using Multiple Regression Analysis". *Journal of Medicinal Chemistry* 15 (10): 1066–68.

- US EPA, ORD,OSIM. s/f. "Species Sensitivity Distributions". Consultado el 29 de marzo de 2018. https://www.epa.gov/exposure-assessment-models/species-sensitivity-distributions.

- Veldstra, H. 1953. "The Relation of Chemical Structure to Bio-Logical Activity in Growth Substances". *Annual Review of Plant Physiology* 4 (1). Annual Reviews: 151–98.

- Waldman, Marvin, Robert Fraczkiewicz, y Robert D. Clark. 2015. "Tales from the war on error: The art and science of curating QSAR data". *Journal of Computer-Aided Molecular Design* 29 (9). Springer International Publishing: 897–910.

- Weininger, David. 1988. "SMILES, a Chemical Language and Information System: 1: Introduction to Methodology and Encoding Rules". *Journal of Chemical Information and Computer Sciences* 28 (1): 31–36.

- Weininger, David, Arthur Weininger, y Joseph L. Weininger. 1989. “SMILES. 2. Algorithm for Generation of Unique SMILES Notation”. *Journal of Chemical Information and Computer Sciences* 29 (2): 97–101.

- Witten, Ian .H., y Eibe Frank. 2005. *Data Mining: Practical machine learning tools and techniques*. *Machine Learning*.

- World Health Organization. 2010. “The Who Recommended Classification of Pesticides By Hazard and Guidelines To Classification 2009”. *World Health Organization*, 1–60.

- Zhu, Hao, Alexander Tropsha, Denis Fourches, Alexandre Varnek, Ester Papa, Paola Gramatical, Tomas Öberg, Phuong Dao, Artem Cherkasov, y Igor V. Tetko. 2008. “Combinatorial QSAR modeling of chemical toxicants tested against Tetrahymena pyriformis”. *Journal of Chemical Information and Modeling* 48 (4): 766–84.

Anexos

Anexo A – Modelo predictivo de red neuronal artificial

NODO 1	
Nombre de entrada	Peso Sináptico
AMW	0.13299
MAXDN	0.602612
ATS3m	-0.16644
ATSC2m	0.550069
ATSC4m	1.441353
ATSC7m	-1.04527
ATSC2p	0.983956
SpMax1_Bh(m)	0.313888
SpMax2_Bh(m)	-0.36237
P_VSA_LogP_1	0.954834
P_VSA_LogP_2	-0.46241
P_VSA_LogP_4	0.099667
P_VSA_LogP_5	1.207515
P_VSA_MR_2	-1.67725
P_VSA_MR_5	-0.48458
P_VSA_MR_8	0.8904
P_VSA_m_1	-0.24795
P_VSA_m_3	0.22109
P_VSA_e_2	0.013368
P_VSA_e_3	0.86958
P_VSA_s_3	-0.99338
P_VSA_s_6	0.716479
Hy	0.93623
TPSA(Tot)	1.443913
MLOGP2	-0.36646
Umbral	-0.71913

NODO 2	
Nombre de entrada	Peso Sináptico
AMW	2.842182766
MAXDN	1.230686986
ATS3m	3.001946693
ATSC2m	-4.14070747
ATSC4m	-1.478464772
ATSC7m	-0.3837408
ATSC2p	-1.394388039
SpMax1_Bh(m)	0.644307235
SpMax2_Bh(m)	0.872707576
P_VSA_LogP_1	2.384314603
P_VSA_LogP_2	2.336312406
P_VSA_LogP_4	9.443333907
P_VSA_LogP_5	-8.145506673
P_VSA_MR_2	3.584234958
P_VSA_MR_5	5.158170965
P_VSA_MR_8	-6.516766387
P_VSA_m_1	-0.708040925
P_VSA_m_3	5.149013558
P_VSA_e_2	2.188399397
P_VSA_e_3	6.231678863
P_VSA_s_3	-0.758931337
P_VSA_s_6	2.369622358
Hy	-3.089875828
TPSA(Tot)	-15.8475938
MLOGP2	1.594566496
Umbral	-4.658551671

NODO 3	
Nombre de entrada	Peso Sináptico
AMW	3.73698528
MAXDN	0.325169849
ATS3m	0.555466497
ATSC2m	-0.691665515
ATSC4m	0.732801087
ATSC7m	0.438394456
ATSC2p	-3.459770349
SpMax1_Bh(m)	1.389687625
SpMax2_Bh(m)	0.738563825
P_VSA_LogP_1	3.562502998
P_VSA_LogP_2	2.123968143
P_VSA_LogP_4	0.832486616
P_VSA_LogP_5	-1.560565287
P_VSA_MR_2	4.423563173
P_VSA_MR_5	2.463116262
P_VSA_MR_8	-1.807631684
P_VSA_m_1	-1.105023464
P_VSA_m_3	2.661775945
P_VSA_e_2	-0.232022221
P_VSA_e_3	2.217211195
P_VSA_s_3	2.429434959
P_VSA_s_6	0.762129675
Hy	2.17239619
TPSA(Tot)	-5.390568504
MLOGP2	-0.172108936
Umbral	-3.872027336

NODO 4	
Nombre de entrada	Peso Sináptico
AMW	3.093485
MAXDN	0.658945
ATS3m	0.225473
ATSC2m	-0.38369
ATSC4m	0.434025
ATSC7m	1.962264
ATSC2p	-1.47264
SpMax1_Bh(m)	1.37113
SpMax2_Bh(m)	1.229482
P_VSA_LogP_1	1.378414
P_VSA_LogP_2	1.34183
P_VSA_LogP_4	2.390911
P_VSA_LogP_5	0.749444
P_VSA_MR_2	0.306082
P_VSA_MR_5	0.63608
P_VSA_MR_8	-2.27798
P_VSA_m_1	0.670885
P_VSA_m_3	-0.15361
P_VSA_e_2	1.228928
P_VSA_e_3	2.037664
P_VSA_s_3	3.922578
P_VSA_s_6	0.734707
Hy	-0.74028
TPSA(Tot)	-2.83733
MLOGP2	4.040902
Umbral	-3.27013

NODO 5	
Nombre de entrada	Peso Sináptico
AMW	1.673153
MAXDN	0.096789
ATS3m	-0.641
ATSC2m	4.147407
ATSC4m	1.445677
ATSC7m	-2.97515
ATSC2p	1.538935
SpMax1_Bh(m)	0.808458
SpMax2_Bh(m)	0.145596
P_VSA_LogP_1	2.939633
P_VSA_LogP_2	0.047509
P_VSA_LogP_4	-1.95112
P_VSA_LogP_5	4.755614
P_VSA_MR_2	-0.83503
P_VSA_MR_5	-0.16814
P_VSA_MR_8	3.092915
P_VSA_m_1	0.457712
P_VSA_m_3	-0.0476
P_VSA_e_2	-1.49718
P_VSA_e_3	0.101325
P_VSA_s_3	4.046748
P_VSA_s_6	1.730833
Hy	0.430852
TPSA(Tot)	1.101374
MLOGP2	-0.06163
Umbral	-2.39449

NODO 6	
Nombre de entrada	Peso Sináptico
AMW	1.523955
MAXDN	1.204148
ATS3m	-0.42662
ATSC2m	1.958858
ATSC4m	0.396824
ATSC7m	0.323248
ATSC2p	0.978829
SpMax1_Bh(m)	1.386356
SpMax2_Bh(m)	0.316346
P_VSA_LogP_1	-0.44363
P_VSA_LogP_2	1.811012
P_VSA_LogP_4	0.624708
P_VSA_LogP_5	0.628712
P_VSA_MR_2	0.705612
P_VSA_MR_5	0.033104
P_VSA_MR_8	1.784256
P_VSA_m_1	0.482432
P_VSA_m_3	0.95349
P_VSA_e_2	-0.32952
P_VSA_e_3	0.52351
P_VSA_s_3	0.531354
P_VSA_s_6	0.933466
Hy	0.280451
TPSA(Tot)	0.612741
MLOGP2	1.74841
Umbral	-1.43076

NODO 7	
Nombre de entrada	Peso Sináptico
AMW	3.134209
MAXDN	-0.20933
ATS3m	-0.26769
ATSC2m	-1.25009
ATSC4m	-0.94429
ATSC7m	-1.55605
ATSC2p	-2.12875
SpMax1_Bh(m)	0.066935
SpMax2_Bh(m)	1.623051
P_VSA_LogP_1	0.845754
P_VSA_LogP_2	0.988217
P_VSA_LogP_4	1.689909
P_VSA_LogP_5	-0.56857
P_VSA_MR_2	2.939005
P_VSA_MR_5	0.681874
P_VSA_MR_8	-2.9079
P_VSA_m_1	1.042199
P_VSA_m_3	0.392441
P_VSA_e_2	2.589182
P_VSA_e_3	1.347503
P_VSA_s_3	2.647023
P_VSA_s_6	-0.14782
Hy	-2.53341
TPSA(Tot)	-5.36757
MLOGP2	6.116762
Umbral	-3.63678

NODO 8	
Nombre de entrada	Peso Sináptico
AMW	1.024904
MAXDN	0.803379
ATS3m	0.237861
ATSC2m	4.671026
ATSC4m	0.015226
ATSC7m	-0.7183
ATSC2p	2.315428
SpMax1_Bh(m)	2.353749
SpMax2_Bh(m)	0.592735
P_VSA_LogP_1	1.482526
P_VSA_LogP_2	3.501901
P_VSA_LogP_4	-0.33452
P_VSA_LogP_5	-0.4059
P_VSA_MR_2	2.461996
P_VSA_MR_5	-0.04475
P_VSA_MR_8	2.814987
P_VSA_m_1	0.964008
P_VSA_m_3	-0.64789
P_VSA_e_2	-3.19659
P_VSA_e_3	-0.04969
P_VSA_s_3	4.536324
P_VSA_s_6	-0.59402
Hy	0.192271
TPSA(Tot)	-0.81793
MLOGP2	0.420151
Umbral	-2.04775

NODO 9	
Nombre de entrada	Peso Sináptico
AMW	7.714255
MAXDN	1.514126
ATS3m	0.758283
ATSC2m	-4.35233
ATSC4m	-2.94261
ATSC7m	2.459578
ATSC2p	-8.47451
SpMax1_Bh(m)	2.60025
SpMax2_Bh(m)	3.0051
P_VSA_LogP_1	7.95715
P_VSA_LogP_2	2.259967
P_VSA_LogP_4	4.508417
P_VSA_LogP_5	-2.53161
P_VSA_MR_2	-3.27139
P_VSA_MR_5	-4.05417
P_VSA_MR_8	-5.71326
P_VSA_m_1	-6.06444
P_VSA_m_3	-1.78314
P_VSA_e_2	1.660377
P_VSA_e_3	-0.47808
P_VSA_s_3	2.671448
P_VSA_s_6	5.909789
Hy	-1.42181
TPSA(Tot)	-1.44606
MLOGP2	0.726783
Umbral	-7.02757

NODO 10	
Nombre de entrada	Peso Sináptico
AMW	6.511793
MAXDN	0.163457
ATS3m	0.100286
ATSC2m	1.852568
ATSC4m	-5.4341
ATSC7m	2.317066
ATSC2p	-3.91823
SpMax1_Bh(m)	0.553136
SpMax2_Bh(m)	-1.17539
P_VSA_LogP_1	-1.20926
P_VSA_LogP_2	0.095606
P_VSA_LogP_4	3.837634
P_VSA_LogP_5	-5.4572
P_VSA_MR_2	-0.69219
P_VSA_MR_5	3.222683
P_VSA_MR_8	4.109843
P_VSA_m_1	-1.74879
P_VSA_m_3	1.863468
P_VSA_e_2	1.378488
P_VSA_e_3	-0.76285
P_VSA_s_3	3.507894
P_VSA_s_6	4.078516
Hy	0.161655
TPSA(Tot)	0.163661
MLOGP2	-2.97345
Umbral	-3.44523

NODO 11	
Nombre de entrada	Peso Sináptico
AMW	2.3100859
MAXDN	2.0651154
ATS3m	-4.002782
ATSC2m	-4.091718
ATSC4m	1.8975193
ATSC7m	-1.262211
ATSC2p	1.4774514
SpMax1_Bh(m)	-4.101937
SpMax2_Bh(m)	1.9678955
P_VSA_LogP_1	-0.648623
P_VSA_LogP_2	5.9782859
P_VSA_LogP_4	1.2312906
P_VSA_LogP_5	-4.093231
P_VSA_MR_2	7.1671443
P_VSA_MR_5	-1.308946
P_VSA_MR_8	-6.086091
P_VSA_m_1	0.3357003
P_VSA_m_3	0.4302543
P_VSA_e_2	14.630078
P_VSA_e_3	0.2151599
P_VSA_s_3	0.45749
P_VSA_s_6	-1.67061
Hy	3.5036102
TPSA(Tot)	-11.27477
MLOGP2	-0.733083
Umbral	-6.024947

NODO 12	
Nombre de entrada	Peso Sináptico
AMW	4.485276
MAXDN	-0.93386
ATS3m	-1.27249
ATSC2m	-0.00875
ATSC4m	-8.15425
ATSC7m	3.327692
ATSC2p	-5.61483
SpMax1_Bh(m)	-3.45445
SpMax2_Bh(m)	-2.33987
P_VSA_LogP_1	-4.87111
P_VSA_LogP_2	-1.46797
P_VSA_LogP_4	6.660497
P_VSA_LogP_5	-9.20311
P_VSA_MR_2	-1.10084
P_VSA_MR_5	5.765963
P_VSA_MR_8	5.320646
P_VSA_m_1	-1.57222
P_VSA_m_3	3.75034
P_VSA_e_2	-3.20175
P_VSA_e_3	2.530773
P_VSA_s_3	5.089426
P_VSA_s_6	6.617801
Hy	2.922509
TPSA(Tot)	2.718457
MLOGP2	-5.51558
Umbral	-5.38435

NODO 13	
Nombre de entrada	Peso Sináptico
AMW	1.021753
MAXDN	1.145578
ATS3m	-0.4855
ATSC2m	1.541019
ATSC4m	1.52361
ATSC7m	1.401885
ATSC2p	1.18856
SpMax1_Bh(m)	0.447962
SpMax2_Bh(m)	0.150407
P_VSA_LogP_1	0.8537
P_VSA_LogP_2	1.347881
P_VSA_LogP_4	0.174181
P_VSA_LogP_5	0.978295
P_VSA_MR_2	0.934844
P_VSA_MR_5	0.891132
P_VSA_MR_8	1.372829
P_VSA_m_1	1.487163
P_VSA_m_3	0.222372
P_VSA_e_2	0.296363
P_VSA_e_3	0.791792
P_VSA_s_3	0.886036
P_VSA_s_6	0.778176
Hy	-0.20688
TPSA(Tot)	0.315709
MLOGP2	1.031525
Umbral	-1.63325