UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE QUÍMICA

Modelo predictivo de la toxicidad oral aguda en rata de pesticidas de la clase 2-trifluorometil benzimidazoles

TESIS

QUE PARA OBTENER EL TÍTULO DE

QUÍMICA DE ALIMENTOS

PRESENTA

GABRIELA GÓMEZ JIMÉNEZ

MÉXICO, Ciudad de México
2018

JURADO ASIGNADO:

PRESIDENTE: José Luis Medina Franco

VOCAL: Antonio Elías Kuri Pineda

SECRETARIO: Karina Martínez Mayorga

1er. SUPLENTE: Abraham Madariaga Mazón

2° SUPLENTE: Sol Castrejón Carrillo

SITIO DONDE SE DESARROLLÓ EL TEMA:

INSTITUTO DE QUÍMICA, UNAM

ASESOR DEL TEMA:

Dra. Karina Martínez Mayorga

SUPERVISOR TÉCNICO:

Dr. Abraham Madariaga Mazón

SUSTENTANTE:

Gabriela Gómez Jiménez

Índice

Índice de tablas

Índice de figuras

Dedicatoria

A mis padres Javier y Silvia por impulsarme, colmarme del amor y valores que forman parte de mí, por enseñarme a ser perseverante y luchar por lo que quiero.

A mis hermanos Gerardo, Carolina y a mis maravillosas amigas Shey, Lau, Alicia y Yaz que siempre han estado ahí para sostenerme, acompañarme y apoyarme en este camino. A mi amigo Alex por tomarse el tiempo de ayudarme con los puntos, comas y reglas gramaticales.

A mis tutores Karina, Abraham y a mis compañeros por ayudarme a concluir este paso, por todo el apoyo, la paciencia, confianza y amistad que me brindaron.

A todas las personas que fueron parte de este proceso lleno de aventuras y retos, gracias, gracias, gracias.

Agradecimientos

A la máxima casa de estudios, la Universidad Nacional Autónoma de México, a mi querida Facultad de Química y al Instituto de Química, UNAM, al Grupo QUIBIC – Servicios QSAR del Instituto de Química, y a Senosiain Laboratorios, por el financiamiento otorgado. A los desarrolladores de T.E.S.T., OSIRIS DataWarrior ChemAxon, Statistic, TIBCO Spotfire y QSARINS, por proporcionar licencias académicas de acceso.

Partes de este trabajo fueron presentadas en el Simposio interno del Instituto de Química 2018, Ciudad de México, 13 de junio de 2018 y en el congreso 22EuroQSAR, Thessaloniki, Grecia, septiembre 16-20, 2018.

1. Introducción

La predicción de valores de toxicidad utilizando métodos computacionales (modelos *in silico*), goza de gran interés. Además de la optimización de recursos, estos métodos ofrecen una alternativa rápida, económica y veraz, cuando se comparan con ensayos en animales (Gozalbes et al. 2014). Esto ha llevado a que las predicciones de diversos valores de toxicidad estén siendo impulsadas por agencias regulatorias, tanto en México como a nivel internacional. En el ámbito regulatorio, las predicciones de toxicidad se realizan para las impurezas derivadas de los procesos de fabricación, principalmente aquellos de la industria agroquímica, farmacéutica, cosmética y alimentaria.

En el sector agroalimentario, cada año se introducen nuevos pesticidas en el mercado. Antes de que estos nuevos productos puedan estar disponibles a la sociedad, las autoridades sanitarias evalúan rigurosamente estos compuestos para garantizar que cumplan con los estándares actuales de salud, medio ambiente y seguridad sanitaria (OECD.Org – OECD, 2018). Para la aplicación de estas alternativas (modelos *in silico*) en regulación, la fiabilidad de las predicciones debe estar lo suficientemente fundamentada y documentada, con el fin de tomar decisiones seguras.

1.1 Objetivo

Desarrollar modelos predictivos de toxicidad oral aguda en rata para compuestos usados como pesticidas en la industria agroalimentaria.

1.2 Hipótesis

A través de modelos QSAR (*Quantitative Structure-Activity Relationships*) validados, se podrá calcular la toxicidad de compuestos estructuralmente similares a los usados en la construcción del modelo. Por ejemplo, para impurezas generadas en la fabricación de pesticidas, usados en la industria agroalimentaria. Dichas predicciones podrán ser empleadas con fines regulatorios.

1.3 Justificación

La información química y biológica ha crecido exponencialmente en los últimos años. Esto ha conducido al surgimiento de metodologías para su almacenamiento, manejo y análisis. Dichas metodologías permiten el reconocimiento de patrones de la información. Por ejemplo, entre estructuras químicas y propiedades biológicas, como la toxicidad. Esto justifica el uso de métodos computacionales para la predicción de toxicidades.

La predicción de toxicidades utilizando métodos QSAR es reconocida por organismos regulatorios a nivel internacional. En México, a partir de 2014, la Comisión Federal para la Protección de Riesgos Sanitarios (COFEPRIS) acepta dichas predicciones para las impurezas de pesticidas registrados por equivalencia. (COFEPRIS, 2018). Esta iniciativa es impulsada por la disminución de ensayos con animales y ahorro de recursos.

2. Marco teórico

En este trabajo se utilizaron métodos de cómputo como limpieza de bases de datos, cálculo de descriptores moleculares y generación de modelos estadísticos por el método de regresión lineal múltiple. En las primeras secciones de este apartado se describen conceptos básicos de la quimioinformática y toxinformática para una mejor comprensión del tema. La información complementaria se presenta en la sección de apéndices.

2.1 Química computacional, aplicaciones y ventajas

El crecimiento actual de información ha impulsado considerablemente el empleo y análisis de datos mediante herramientas computacionales. En el caso de la química, el uso de las computadoras y de análisis informáticos, no ha sido la excepción. La química computacional es una rama de la química que aborda problemas mediante la generación de modelos y simulaciones, haciendo uso de las computadoras (Clementi 1980). La química computacional tiene aplicaciones en múltiples ramas de la ciencia, con especial énfasis en la manipulación de información estructural química. La ventaja que tienen estos métodos computacionales radica en la enorme cantidad de información molecular que puede procesarse y generarse gracias a ellos. La información obtenida con estos estudios es particularmente útil para complementar y entender datos obtenidos experimentalmente (IRAIS 2014).

2.2 Representaciones químicas y descriptores moleculares

Un paso fundamental para generar y analizar compuestos químicos es convertir la estructura molecular en representaciones canónicas relevantes. Esta representación permite extraer información sobre una estructura dada desde una base de datos. Existen algoritmos informáticos bien establecidos para la

generación de estas representaciones, como las cadenas canónicas SMILES (Leach and Gillet 2007).

La manipulación y el análisis de la información estructural química son posibles mediante el uso de descriptores moleculares. Éstos son valores numéricos que resultan de un procesamiento matemático o experimento y contienen información química y física. Esta información es utilizada para la generación de modelos predictivos. Las moléculas pueden ser representadas con descriptores moleculares en una, dos o tres dimensiones (1D, 2D, 3D). Las llamadas "huellas digitales moleculares" o "*molecular fingerprints*" son generalmente representaciones características de las estructuras químicas en una o dos dimensiones (Leach and Gillet 2007). El nivel de complejidad de la representación estructural dependerá de los objetivos planteados en cada estudio en particular. La selección de la representación de las estructuras es esencial para estudiarlas adecuadamente.

La dimensionalidad de un conjunto de datos es la cantidad de variables que se utilizan para describir cada objeto. El Análisis de Componentes Principales (PCA por sus siglas en inglés) es un método comúnmente utilizado para reducir la dimensionalidad de un conjunto de datos. PCA proporciona un nuevo conjunto de variables que tienen algunas propiedades especiales; a menudo se encuentra que gran parte de la variación en el conjunto de datos puede explicarse por un pequeño número de estos componentes principales. Los componentes principales también son convenientes para la visualización y el análisis gráfico de datos (Leach and Gillet 2007).

La química informática y el modelado molecular han tenido una gran aplicación en etapas tempranas del desarrollo de fármacos, para la selección y optimización de nuevos compuestos con propiedades terapéuticas (Mignani et al. 2018). Recientemente en la rama de alimentos el uso de estas metodologías ha tenido un gran interés en explorar los posibles beneficios secundarios de los saborizantes y toxicología predictiva (Martínez-Mayorga et al. 2011).

2.3 Toxicología y Toxicología computacional

La toxicología es la investigación científica de los efectos nocivos causados por los venenos. Es una ciencia interdisciplinaria que abarca diversos aspectos de áreas como medicina, química, farmacología, biología, biología molecular, ecología, etc. Los resultados de las investigaciones toxicológicas tienen un impacto considerable en las decisiones políticas y económicas. Esto abarca desde el registro de nuevos productos farmacéuticos y productos químicos industriales hasta la salud pública y los aspectos ambientales de grandes proyectos industriales y de infraestructura (Helma et al. 2000).

El objetivo principal en toxicología es estimar si un agente (generalmente una sustancia química o una mezcla de compuestos) causa daño a un objetivo biológico. Puede ser una macromolécula biológica, una estructura celular, un órgano, un organismo, una población o incluso un ecosistema completo. Investigaciones más específicas intentan elucidar los mecanismos biológicos y químicos responsables del daño tóxico (Helma et al. 2000).

La toxicología computacional es una subdisciplina de la toxicología, que tiene como objetivo utilizar las matemáticas, la estadística, el modelado químico y las herramientas informáticas para predecir los efectos tóxicos de sustancias químicas en la salud humana y/o el medio ambiente, y además comprender mejor los mecanismos por los que un producto químico induce daño (Rusyn and Daston 2010).

Estas predicciones se logran gracias a la disponibilidad de datos, por ejemplo, valores experimentales reportados en la literatura. A través de ellos y mediante la construcción de modelos matemáticos es posible encontrar relaciones entre las estructuras químicas con propiedades fisicoquímicas o actividades biológicas. La importancia de los descriptores moleculares radica en el conocimiento y comprensión de patrones identificados a partir de ellos, por lo tanto, son un elemento muy importante en la generación de modelos predictivos.

Estos modelos permiten predecir: toxicidad en humanos (sistemática y a nivel local), distribución ambiental (persistencia, bioacumulación) y ecotoxicidad (efectos en otros organismos) (Gozalbes et al. 2014). Dicha correlación puede usarse para predecir propiedades o actividades biológicas de nuevas moléculas, basándose en la premisa de que estructuras similares tendrán propiedades similares. (Maggiora et al. 2014). Estos métodos son llamados QSAR (en español relación cuantitativa estructura actividad) o QSTR (*Quinatitative Structure-Toxicity Relationships*, relación cuantitativa estructura toxicidad). Los primeros intentos de QSAR fueron estadísticos por naturaleza, basados en la premisa de que la toxicidad podría correlacionarse con ciertas características moleculares de los agentes químicos que causan ese tipo particular de toxicidad (Rusyn and Daston 2010).

Existe la necesidad, de desarrollar métodos de detección más rápidos basados en una comprensión mecanicista de la toxicidad. La EPA (del inglés *Environmental Protection Agency*) de EE. UU. y otras organizaciones internacionales como REACH (del inglés *Registration, Evaluation, Authorisation and Restriction of Chemicals*), OCDE (del inglés *Organisation for Economic Co-operation and Development*), entre otras, contemplan estos programas de detección como el primer paso para priorizar agentes para pruebas *in vivo* y/o *in vitro* (Judson et al. 2009). Para el proceso general de evaluación de riesgos, los datos *in vitro* y las predicciones computacionales son potencialmente útiles. La gestión, el análisis y la interpretación de los nuevos datos ahora disponibles para la evaluación de la seguridad toxicológica proporcionan una gran cantidad de información sobre el posible modo de acción de los productos químicos en evaluación, así como el valor de los ensayos individuales en apoyo de predicciones para nuevos productos químicos.

La evolución de los modelos QSAR (Tropsha and Golbraikh 2007, Nandi et al. 2006) condujo a su aceptación en la predicción de evaluaciones toxicológicas para fines regulatorios.

Es importante destacar que la ciencia de la toxicología computacional impacta tanto la investigación básica como la toma de decisiones regulatorias y la protección de la salud ambiental (Kavlock, Austin, and Tice 2009). De hecho, estudios recientes demostraron que el poder predictivo de los modelos QSAR para la toxicidad *in vivo* mejora cuando los resultados de las pruebas *in vitro* (es decir, los descriptores biológicos) se combinan con los descriptores químicos tradicionales (Zhu et al. 2008).

Las predicciones de evaluaciones toxicológicas permiten reducir los costos experimentales, pruebas en animales (Gramatica et al. 2013), proporcionan un medio efectivo y rápido para la selección y clasificación de compuestos (*screening*), proporcionan herramientas importantes para incrementar nuestro conocimiento de los efectos tóxicos a nivel celular y molecular y son puentes esenciales entre los animales de experimentación y los humanos (Repetto, Peso, and Zurita 2000).

2.4 Definición de QSAR

Los modelos QSAR o QSTR tienen bases estadísticas que son capaces de detectar patrones en los datos analizados a través de correlaciones matemáticas robustas que involucran descriptores moleculares. Cuando se desarrollan correctamente y se validan rigurosamente, los modelos QSAR son útiles para el cribado y la priorización de productos químicos o incluso antes de su síntesis en el enfoque de diseño químico seguro (Gramatica et al. 2013).

Los pasos principales en el desarrollo y análisis de modelos QSAR son: (1) Preparación de estructuras y cálculo de descriptores, análisis de datos y configuración del conjunto de datos de entrada (creación de dos grupos, entrenamiento y validación), (2) Selección de descriptores y cálculo de modelos, (3) Exploración de modelos, validación, y selección (Gramatica et al. 2013).

En la tabla 1 se muestran los pasos involucrados en las metodologías KDD (*Knowledge Discovery in Databases*, en español extracción de conocimiento de bases de datos) y QSAR con respecto a los principios de la regulación internacional OCDE

Tabla 1. Pasos involucrados en las metodologías KDD, QSAR y los principios relacionados con la OCDE.

	KDD	QSAR	OCDE
1	Definición del objetivo	Definición de respuesta biológica	Definición de respuesta
	1.2 Creación o selección de un conjunto de datos	Preparación de datos	
	1.3 Limpieza y preprocesamiento de datos		
	1.4 Reducción de datos y proyección		
2	Selección de métodos de minería de datos		
3	Análisis exploratorio y selección de modelo / hipótesis	División de datos / Generación del modelo / Validación interna	Definir dominio de aplicación
4	Minería de datos	Generación del modelo	Algoritmo no ambiguo
5	Evaluación	Validación externa / Evaluación experimental	Medidas apropiadas de bondad de ajuste, robustez y poder predictivo
6	Interpretación / Utilización		Interpretación mecanística de ser posible

(Gomez-Jimenez et al. 2018)

Existen programas especializados en QSAR tanto para el cálculo de los descriptores numéricos, como para las técnicas estadísticas para el desarrollo de los modelos matemáticos.

La selección de algoritmos y técnicas estadísticas a usar depende del tipo de datos y el tipo de predicción que se desea extraer (activo/inactivo, DL_{50}).

2.5 Metodología QSAR

A continuación se describen de forma más detallada las etapas de desarrollo de los modelos QSAR/QSTR (Gozalbes et al. 2014).

La creación de una base de datos, consiste en las respuestas experimentales a modelar y el correspondiente conjunto de descriptores de estructura molecular (que se calculan utilizando un software apropiado, inequívocamente representados) para cada compuesto químico. Esta información estructural y biológica se divide en dos grupos: “entrenamiento” para desarrollo del modelo y “validación o prueba” para verificar su poder predictivo.

El conjunto de entrenamiento se utiliza como entrada para la generación del modelo QSAR mediante análisis estadístico. Este conjunto de datos se analiza, es decir, se verifica la distribución de los datos en el espacio químico y experimental, destacando posibles valores atípicos y/o grupos particulares.

Se utilizan diversas herramientas estadísticas para la selección de variables/descriptores. Las herramientas de selección de variables permiten el uso de descriptores adecuados y relevantes para respuestas particulares, eliminando información no relevante y reduciendo el tiempo de análisis. Estas herramientas se dividen en dos grandes grupos: 1) selección del mejor subconjunto, que evalúa todos los subconjuntos posibles (todas las combinaciones de *p* variables, desde el tamaño 1 hasta *p*, donde *p* es el total del número de variables), estos modelos de subconjunto son evaluados con una función de aptitud a través de la cual se selecciona el mejor modelo; y 2) la selección por el método de inteligencia de enjambre (*swarm intelligence)*, en el cual se usa la métrica estadística para encontrar la mejor solución posible (optimización) para un determinado problema.

En estos métodos los modelos evolucionan mediante descentralización del control y la autoorganización. Dentro de este grupo se encuentra la técnica de optimización de partículas, colonia de hormigas y algoritmo genético, por mencionar algunas (Bonabeau et al. 1999).

La técnica de algoritmo genético está basada en la teoría de la evolución donde se intenta replicar el comportamiento biológico de la selección natural y la genética. El algoritmo genético comienza con una población inicial de modelos (simula cromosomas), compuestos a su vez por variables/descriptores (simula genes), estos modelos representan posibles soluciones, los modelos son evaluados por una función de aptitud y de acuerdo con la calificación de aptitud se puede evaluar sí el modelo es eficaz para resolver el problema planteado. Una vez seleccionados los mejores modelos comienza el cruce y recombinación de variables (combinación y mutación). Una ventaja de esta técnica es que posee la habilidad de manejar muchas variables simultáneamente, lo cual permite evaluar muchos modelos, sin la necesidad de explorarlos todos, lo que representa un ahorro de tiempo y recursos (Haupt and Haupt 1998).

Existen técnicas estadísticas que se aplican al grupo de compuestos de entrenamiento para generación de modelos. De los distintos métodos matemáticos utilizados para derivar modelos QSAR, la técnica más utilizada es la regresión lineal simple y múltiple. El tipo más simple de ecuación de regresión lineal tiene la siguiente forma:

$$y = mx + c$$

En esta ecuación, (y) se denomina variable dependiente, siendo (x) la variable independiente. En QSAR o QSTR (y) corresponde a la propiedad que se desea modelar, como la actividad biológica y (x) sería un descriptor molecular como “log P” o una característica estructural, etc. El objetivo de la regresión lineal es encontrar valores para el coeficiente (m) y la constante (c) que minimicen la suma de las diferencias entre los valores predichos por la ecuación y las observaciones experimentales (Leach and Gillet 2007).

Toda vez que se genera un modelo, se debe tener en cuenta diferentes aspectos para considerarlo como aceptable. En primer lugar, un modelo debe tener una alta capacidad de reproducir los datos utilizados para calcularlo (conjunto de entrenamiento), es decir, un buen ajuste de valores experimentales con predichos (R^2, bondad de ajuste). Luego, los modelos QSAR desarrollados están sujetos a diferentes pruebas de validación (evaluadas mediante una función de aptitud) para verificar la confiabilidad de los modelos de correlación desarrollados, por lo tanto, se comprueba que el modelo tiene una alta capacidad para predecir porciones del conjunto de datos de entrenamiento, usando técnicas conocidas colectivamente como validación cruzada (CV por sus siglas en inglés) o validación interna (Q^2_{LOO}, Q^2_{LMO}, $Y\text{-}_{Scramble}$). Si el modelo pasa esta verificación, el modelo puede ser definido como robusto y estable, lo cual cumple con las directrices de la OCDE (Gramatica 2007).

Una vez que el modelo es validado de manera interna y excluida la correlación casual, se procede a realizar la validación externa, es decir, se comprueba su capacidad para predecir nuevos compuestos (aquellos que no participaron en la generación del modelo). Esto se hace aplicando la ecuación del modelo, obtenida en el conjunto de entrenamiento, a uno o más conjuntos de datos de prueba, es decir, a los compuestos que nunca se han utilizado en el cálculo del modelo, y midiendo los resultados por medio de diferentes criterios, tales como: $RMSE_{EXT}$ (raíz cuadrada del error medio del conjunto externo) , Q^2_{F1}, Q^2_{F2}, Q^2_{F3} (coeficientes de ajuste de los valores experimentales contra los calculados para el conjunto externo, las discrepancias de cada formula se pueden apreciar en la sección de Apéndices, Validación externa), CCC (coeficiente de correlación de concordancia) y el método de Golbraikh y Tropsha (Gramatica 2007).

Los modelos QSAR se pueden aplicar para predecir variables respuesta de moléculas que carecen de una evaluación experimental, siempre y cuando estén contenidas en el dominio de aplicación estructural y los estadísticos (validación interna y externa) de la predicción sean aceptables.

2.6 Regulación OCDE

Las directrices de la OCDE para la evaluación de sustancias químicas son aceptadas internacionalmente para pruebas de seguridad y evaluación de aditivos alimentarios, pesticidas, medicamentos, cosméticos, o compuestos industriales, e incluso para ayudar en la toma de decisiones en respuesta a diferentes situaciones de emergencia (Demchuk et al. 2011, Ruiz et al. 2012). Estas pautas se actualizan regularmente con la asistencia de miles de expertos de los países miembros de la OCDE (Gomez-Jimenez et al. 2018).

Las reglas existentes para determinar si un modelo QSAR/QSTR es apropiado para uso regulatorio se mencionan a continuación (Gozalbes et al. 2014):

Reglas de Setúbal (nov 2004):

i) Los modelos deben orientarse a parámetros toxicológicos ("*endpoints*") bien definidos y de clara importancia regulatoria,
ii) Deben tomar la forma de un algoritmo inequívoco,
iii) Su dominio de aplicación (AD por sus siglas en inglés) debe estar claramente definido y justificado,
iv) Deben cumplir con las medidas reconocidas científicamente para demostrar la bondad de su ajuste, robustez y capacidad de predicción,
v) De preferencia deben aportar una posible interpretación sobre los mecanismos de acción toxicológico de los compuestos estudiados.

Los modelos QSAR se basan en la agrupación categórica de productos químicos que surgen de los análisis de tendencia. A partir de los resultados existentes, puede predecirse información (ej. toxicidad) para un producto químico no probado, considerando la similitud, por ejemplo, actividad, propiedad o estructura, con los productos químicos que constituyen el análisis en cuestión. Dado que el AD se deriva del modelo y de los descriptores que conforman la ecuación, los modelos tienen limitaciones inherentes. Las incertidumbres de las predicciones de objetivos de toxicidad pueden disminuir con la capacidad de interpretación y el poder predictivo validado de los modelos QSAR. Claramente, es necesario realizar

mejoras continuas en los modelos QSAR, particularmente para aquellos utilizados con fines regulatorios. La falta de datos confiables plantea uno de los mayores desafíos para el desarrollo de QSAR. El libre acceso a bases de datos privadas por parte de la comunidad científica facilitaría la validación y la mejora adicional de los modelos de predicción (Cherkasov et al. 2014).

En este sentido, resultaría deseable la publicación de datos en un esquema armonizado, que pudiera proporcionar toda la información experimental necesaria. Un formato estructurado podría procesarse automáticamente disminuyendo la necesidad de un curado manual.

Se espera que el rápido crecimiento de las bases de datos disponibles al público comience a abordar este problema, y permitirá mejoras en el desarrollo de modelos QSAR.

El modelado QSAR es, por naturaleza, un área multidisciplinaria. Es por ello que, resulta fundamental promover mejores prácticas e implementar el enfoque de comparar los modelos QSAR relacionados para obtener nuevos conocimientos sobre los mecanismos comunes para diferentes objetivos.

2.7 El papel de los pesticidas en México y su regulación

La modernización de la agricultura, la urbanización y el crecimiento acelerado de la población son factores que contribuyen al aumento de la demanda de alimentos, lo cual implica el uso de pesticidas, con el fin de incrementar la producción agrícola. Sin embargo, algunos de estos compuestos son altamente persistentes, tóxicos y bioacumulables. Por lo tanto, el riesgo de usar dichos compuestos debe ser estudiado, tanto a nivel ambiental como en seres humanos y otros organismos vivos (Cabrera-García Héctor 2010).

En México, la autoridad sanitaria encargada de garantizar la eficacia y seguridad del uso y venta de pesticidas es COFEPRIS, que trabaja en coordinación con la Secretaria de Agricultura, Ganadería, Desarrollo Rural, Pesca y Alimentación (SAGARPA) y la Secretaría del Medio Ambiente y Recursos Naturales

(SEMARNAT). Estás dependencias nos brindan los protocolos que marcan la pauta de los ensayos y estudios que deben realizarse para registrar y usar pesticidas en México (COFEPRIS,2018.).

En materia de regulación en México, el 13 de febrero de 2014 se expidió el decreto por el que se reforman, adicionan y derogan diversas disposiciones del Reglamento en Materia de Registros, Autorizaciones de Importación y Exportación y Certificados de Exportación de Plaguicidas, Nutrientes Vegetales y Sustancias y Materiales Tóxicos o Peligrosos (DOF: 13/02/2014). Donde, se sugiere el uso de modelos QSAR para calcular la toxicidad de las impurezas (subproductos de fabricación o almacenamiento de un pesticida), ya que exhiben características físicas y/o químicas similares al ingrediente activo de los pesticidas técnicos registrados como equivalentes a moléculas previamente registradas (COFEPRIS, 2018).

Un pesticida equivalente es aquel material técnico o técnico concentrado que presenta similitud a un perfil de referencia en sus impurezas y/o perfil toxicológico, generados por distintos fabricantes. Un perfil de referencia es una especificación completa (identidad y composición, propiedades fisicoquímicas, información confidencial, perfil toxicológico completo, perfil ecotoxicológico) (COFEPRIS, 2018).

La equivalencia de un producto se establece cuando cumple con los criterios químicos y toxicológicos (COFEPRIS, 2018):

1. Toda impureza diferente a la reportada en el perfil se considera como nueva.
2. Cuando haya impurezas nuevas cuya concentración sea ≥ 0.1%, se solicitará justificar la no relevancia de estas impurezas.
3. Las impurezas relevantes son aquellos subproductos de fabricación o almacenamiento del pesticida, las cuales, comparado con el ingrediente activo son toxicológicamente significativos para la salud.

4. Si la impureza es menor a 1 g/Kg (0.1%), no requiere justificación sobre relevancia.

La toxicidad de la impureza se puede comprobar utilizando estudios *in vivo* o estudios cuantitativos de estructura-actividad (QSAR), para demostrar que la toxicidad de la impureza es igual o menor que la toxicidad del principio activo al comparar la DL_{50} u otro parámetro toxicológico. Es esencial que los resultados de los modelos se adjunten al anexo J del Manual de FAO y los principios de la OCDE (FAO, OCDE, 2018). Además se debe presentar en el reporte la información de respaldo y la memoria de cálculo correspondiente, con el fin de que los datos sean transparentes y reproducibles.

Es importante contar con métodos de evaluación y regulación que aseguren la protección de la salud humana, los animales y el medio ambiente, así como minimizar, en la medida de lo posible, los niveles de riesgo para el hombre, los animales y el medio ambiente como consecuencia de la comercialización y el uso de plaguicidas agrícolas (García Hernández et al., 2018).

Los métodos de computacionales QSAR permiten la predicción de la toxicidad de nuevos compuestos a bajo costo, de manera rápida, segura y sin el uso adicional de animales. Esto representa una alternativa amigable con el medio ambiente que permite seguir impulsando el desarrollo y crecimiento de la industria agroquímica en México.

3 Metodología

3.1 Paquetería de cómputo

Este trabajo se llevó a cabo con una estación de trabajo DELL con sistema operativo Windows 7 Professional. La paquetería de cómputo especializada que se utilizó fue la siguiente:

-Molecular Operating Environment (MOE) Versión 2010 (http://www.chemcomp.com/MOE-Molecular_Operating_Environment.htm)

- Microsoft Excel 2016

- Dragon7 molecular descriptors

- QSAR-INSUBRIA (http://www.qsar.it/)

- Data Warrior (http://www.openmolecules.org/datawarrior/)

3.2 Diagramas de flujo generales

El procedimiento se dividió en tres etapas: i) obtención y curación de las estructuras químicas de las bases de datos de pesticidas; ii) cálculo de descriptores moleculares y iii) desarrollo de modelo QSAR para predicción de toxicidad oral aguda en rata. En la figura 1 se presenta el diagrama de flujo general y en la figura 2 se presenta un diagrama de flujo más detallado.

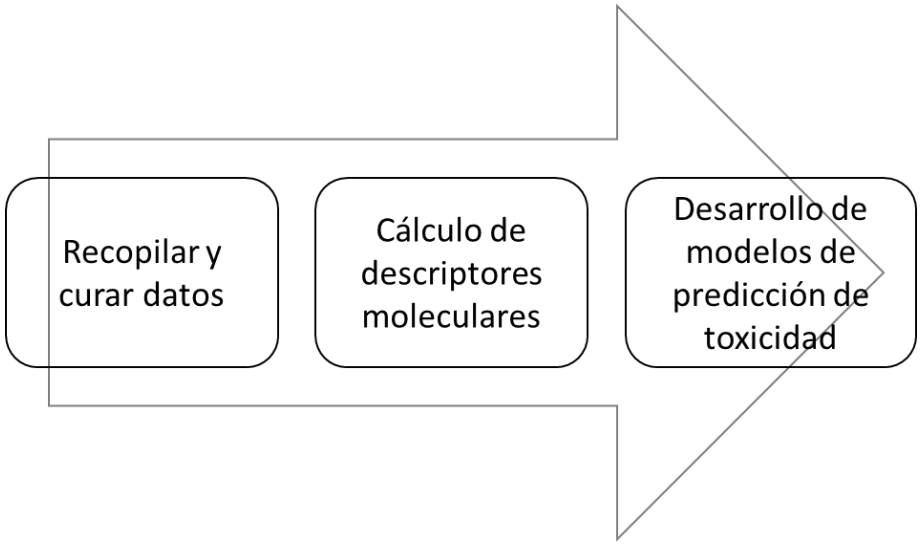


Figura 1 Diagrama de flujo general.

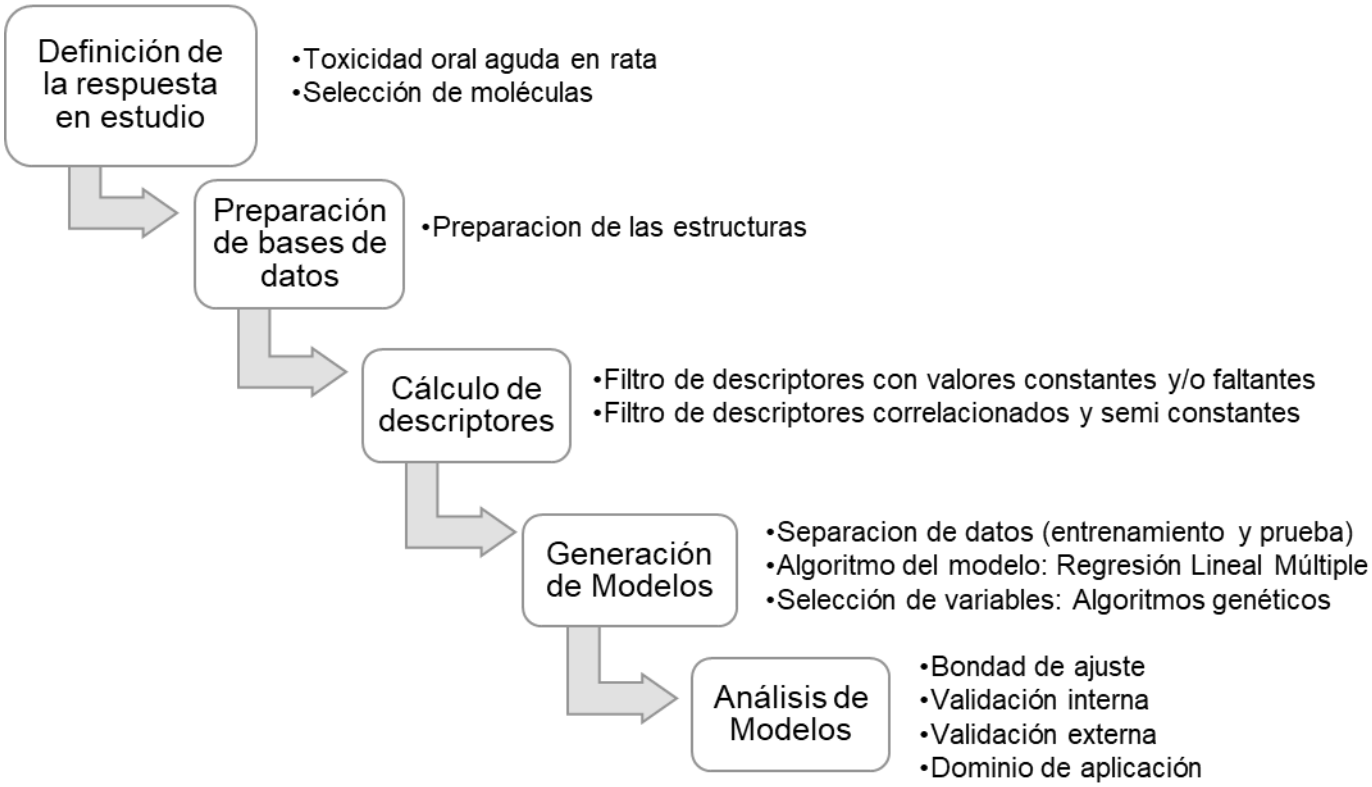


Figura 2 Diagrama de flujo general para el desarrollo de modelos QSAR utilizado en este trabajo.

3.3 Descripción de la metodología

3.3.1 Definición de la respuesta y recopilación de información

Análisis de datos experimentales de toxicidad oral aguda en rata reportada como DL_{50}: mg/kg y DL_{50}: -log10(mol/kg). Recopilación de las estructuras químicas de dos bases de datos: a) PESTIMEP (Chávez-Gómez, 2018) la cual contiene pesticidas evaluados en varios ensayos toxicológicos y b) T.E.S.T. desarrollado por EPA, que contiene compuestos estructuralmente diversos.

3.3.2 Creación de base de datos

Curado y preparación de base de datos: Limpieza y optimización de las estructuras químicas de cada base de datos utilizando MOE. Aplicando la herramienta *wash*, se eliminaron los duplicados y contraiones, se neutralizaron los estados de protonación y se realizó el cálculo de cargas parciales con un campo de fuerza MMFFP94x. Por último, se minimizó la energía de las moléculas restantes con un gradiente RMSD de 0.01.

3.3.3 Selección de las estructuras

Selección de un subgrupo de moléculas (121) de la base de datos T.E.S.T. con base en similitud estructural, para la cual se calculó su *fingerprint* (descriptor de estructura) con el programa Data Warrior y se hizo un análisis de similitud por medio del índice de Tanimoto (información complementaria en la sección de apéndices) con un valor de corte de 0.8.

Se usó la base PESTIMEP completa (145 moléculas).

3.3.4 Cálculo de descriptores

Cálculo de descriptores moleculares en una (1D), dos (2D) y tres (3D) dimensiones para cada compuesto de cada base de datos, utilizando el programa Dragon7.

3.3.5 Generación de modelos

i. Introducción del conjunto de datos al programa QSARINS.

ii. Aplicación del tratamiento de pre-reducción de descriptores, mediante el cual se eliminan aquellos descriptores que cumplan con alguno de los siguientes filtros: (1) pruebas de valores idénticos (variables constantes (>80%)); (2) correlaciones en pares (de acuerdo con un valor de corte de: >98%) (información complementaria en la sección de apéndices).

iii. Visualización y análisis preliminar de los datos curados, mediante herramientas disponibles en el programa QSARINS (perfil de las variables y de los compuestos). (Véase sección de resultados)

iv. Creación del conjunto de entrenamiento y prueba.

v. Normalización de las variables sobre el intervalo de valores de los compuestos químicos correspondientes.

vi. Generación de modelos de baja dimensión (2 D) por el método de todos los subconjuntos.

vii. Selección de variables por Algoritmos Genéticos (AG). Configuración del método para realizar 2000 iteraciones a cada modelo de baja dimensión, el AG evolucionó en los 200 mejores modelos (tamaño de la población inicial) con una tasa de mutación del 40%.

viii. Evaluación de los modelos generados de acuerdo con las siguientes funciones de aptitud*:* R^2 y Q^2_{Loo}.

ix. Eliminación de compuestos atípicos del conjunto de entrenamiento.

x. Selección del mejor modelo con base en los criterios estadísticos de ajuste y validación interna.

3.3.6 Validación de modelos

i. Verificación de la capacidad de predicción del modelo seleccionado, evaluando los compuestos químicos del conjunto de prueba con los criterios estadísticos siguientes: R^2_{EXT}, Q^2_{F1}, Q^2_{F2}, Q^2_{F3} y *CCC*.

ii. Eliminación de compuestos atípicos del conjunto de prueba.

iii. Análisis del dominio de aplicación y compuestos atípicos del modelo seleccionado.

4 Resultados y Discusión

4.1 Definición de la respuesta de estudio y curado de la base de datos

Se estudiaron dos bases de datos para la generación de modelos predictivos de toxicidad oral aguada en rata. La evaluación toxicológica se empleó en términos de dosis letal 50 (DL_{50}), que es la medida de la toxicidad de un material que conduce a la muerte de la mitad de la población de muestra a través de la exposición por ingestión expresada en miligramos del material por kilogramo del peso corporal del animal de prueba (Ruiz et al. 2012). Los valores de DL_{50} se convirtieron a unidades -log10(mol/kg). En la siguiente tabla se muestran las bases de datos estudiadas.

Tabla 2 Bases de datos utilizadas para la generación de modelos predictivos

Base de Datos	**Número de moléculas**	**Contenido**	**Moléculas seleccionadas**
T.E.S.T.	7413	Moléculas diversas	121
PESTIMEP	146	Pesticidas evaluados en múltiples ensayos toxicológicos	145

Las bases de datos fueron analizadas y curadas con el programa de computo MOE. Se eliminaron estructuras repetidas y se prepararon las estructuras para el cálculo de descriptores, la optimización de la geometría completa se realizó mediante el método del Campo de Fuerza Molecular Merck (MMFF94). En total fueron eliminadas 107 estructuras repetidas de la base de datos T.E.S.T. y una de la base de datos PESTIMEP.

Un análisis de similitud estructural por medio del índice de Tanimoto (valor de corte de 0.8) demostró que la base de datos PESTIMET presenta una amplia diversidad estructural.

La selección de moléculas para la generación de los modelos se realizó mediante un análisis de similitud (índice de Tanimoto) de la base de datos T.E.S.T. Se encontraron 121 moléculas similares las cuales pertenecen a la clase química de los 2-trifluorometil benzimidazoles, en la figura 3 se muestra el núcleo base de esta clase química.

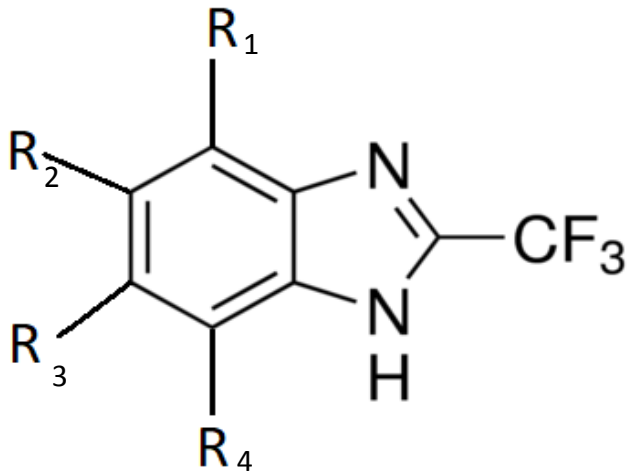


Figura 3 Núcleo base de los pesticidas seleccionados de la base de datos T.E.S.T.

4. 2 Cálculo de descriptores

4.2.1 Base de datos T.E.S.T.

Los archivos preparados y seleccionados (121) fueron utilizados como entrada para el cálculo de descriptores en Dragon7, en formato SDF. Se calcularon un total de 5270 descriptores moleculares de diferentes tipos (1D, 2D y 3D). Se excluyeron los descriptores constantes y aquellos que tuvieran al menos un valor faltante (1700 constantes y 1979 descriptores con valores faltantes), en total se

excluyeron 3679 y fueron exportados 1597 descriptores. En la tabla 3, se muestra una fracción de los datos exportados.

Tabla 3 Compuestos químicos seleccionados de la base de datos T.E.S.T. con sus respectivos descriptores moleculares.

	ID	NAME	Tox -log 10 (mol/kg)	MW	AMW	Se	Me	Mp	Mi
1	369		4.8000	231.1500	11.5570	22.2650	1.1130	0.6630	1.1880
2	479		4.5100	220.5900	12.2550	19.7160	1.0950	0.7200	1.1780
3	672		4.5200	422.8200	23.4900	20.0800	1.1160	0.8980	1.1550
4	682		4.9800	265.5900	13.2790	22.5890	1.1290	0.7060	1.1850
5	715		4.9900	255.0300	14.1680	20.0400	1.1130	0.7680	1.1750
6	716		4.2200	303.5000	14.4520	23.2470	1.1070	0.7830	1.1670
7	795		3.9760	265.0400	14.7240	19.6220	1.0900	0.7480	1.1720
8	797		5.4000	310.0400	15.5020	22.4940	1.1250	0.7310	1.1800
9	886		5.0800	412.8100	22.9340	20.4980	1.1390	0.9180	1.1570
10	892		3.4700	345.0300	15.6830	25.7860	1.1720	0.7330	1.1890

4.2.2 Base de datos PESTIMEP

Para las moléculas presentes en la base de datos PESTIMEP se calcularon un total de 5270 descriptores moleculares de diferentes tipos (1D, 2D y 3D) mediante el paquete Dragon7. Se excluyeron los descriptores constantes y aquellos que tuvieran al menos un valor faltante (1659 constantes y 2605 descriptores con valores faltantes), en total se excluyeron 4048 y fueron exportados 1186 descriptores. En la tabla 4, se muestra una fracción de los datos exportados.

Tabla 4 Compuestos químicos seleccionados de la base de datos PESTIMEP con sus respectivos descriptores moleculares.

	No.	NAME	Tox -log10(mol/kg)	orl rat (mg/kg)	MW	AMW	Se	Sp	Si
1	1		1.7919	7533.3335	466.5100	9.9257	47.8618	35.4091	52.0310
2	2		1.9129	3960.0000	324.0500	14.0891	23.7892	20.8863	24.7720
3	3		2.3854	750.0000	182.1800	9.5884	19.6109	13.1250	21.6386
4	4		2.4037	1065.0000	269.8000	7.1000	37.9161	24.3866	43.0131
5	5		1.8996	3400.0000	269.8000	7.1000	37.9161	24.3866	43.0131
6	6		2.4272	1352.5000	361.6700	11.6668	34.0181	21.7556	35.5844
7	7		1.2804	8400.0000	160.2000	6.9652	23.6035	13.1819	26.7008
8	8		3.6561	21.0000	95.1400	6.7957	13.9126	8.9149	16.0346
9	9		2.0542	2850.0000	322.9100	11.1348	28.7235	24.2104	32.5004
10	10		1.9339	3900.0000	334.9000	7.6114	44.3125	28.4604	49.6840

4.3 Generación de modelos QSAR (T.E.S.T.)

4.3.1 Base de datos T.E.S.T.

La base de datos con los descriptores moleculares fue importada al programa QSARINS, donde los descriptores semi-constantes en más del 80% (al menos el 20% de los compuestos debe tener valores diferentes de cero o de los valores de otras sustancias químicas) y descriptores correlacionados en más de 98% (mediante comparación por pares) fueron excluidos en el paso de pre-reducción, se excluyó el 76.58% de los descriptores (1223 descriptores), en este paso se eliminó información redundante o sin relevancia, debido a que no todos los descriptores tienen una aportación significativa en el cálculo del modelo. Se continuó la selección de variables por el método de algoritmos genéticos (374 descriptores restantes) para el desarrollo del modelo QSAR. Para generar el

modelo se aplicó el método de regresión lineal múltiple (por el método de Mínimos Cuadrados Ordinario MCO).

4.3.2 División de datos

Se separaron los datos en dos grupos (*splitting*), el primer grupo consistió en 101 moléculas para el desarrollo de los modelos (grupo de entrenamiento) y el segundo grupo consistió en 20 moléculas para evaluar el poder predictivo de los modelos (grupo de validación o prueba). En la figura 4 se muestra la distribución de la toxicidad expresada en dosis letal media (DL_{50}, -log10) que presentan las moléculas en estudio para cada grupo de la base de datos T.E.S.T.

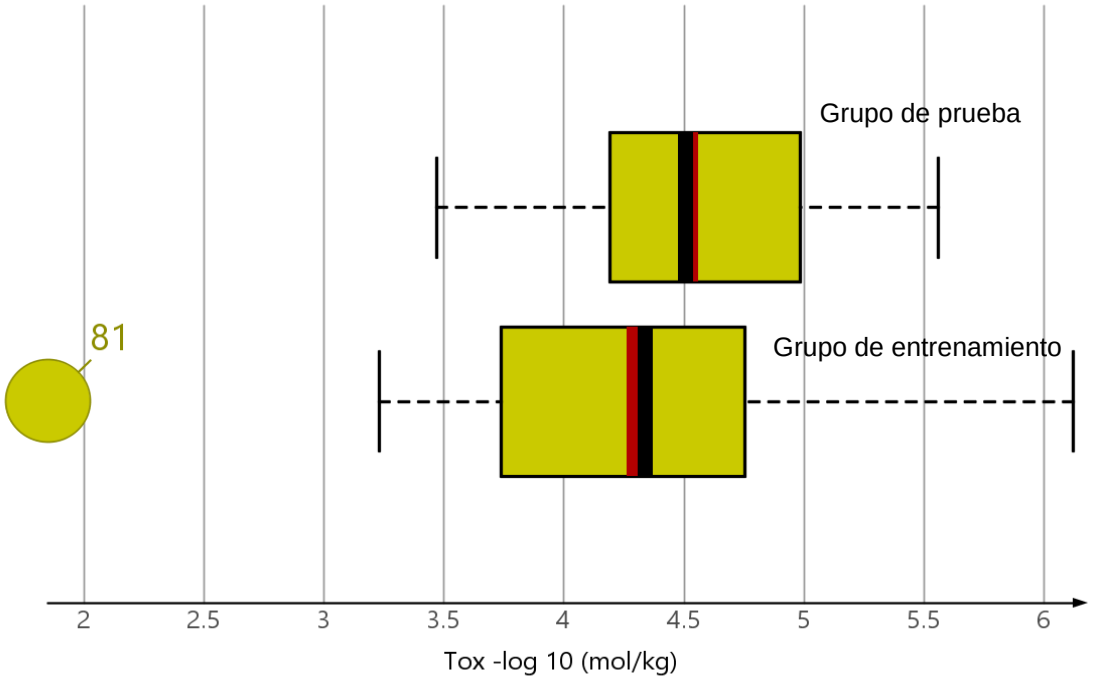


Figura 4 Intervalo de la toxicidad oral aguda en rata para las moléculas seleccionadas de la base de datos T.E.S.T.

Se comenzó el desarrollo del modelo con el grupo de entrenamiento (*full model*) donde los 101 compuestos presentes contribuyen con sus características para la selección de variables de los Modelos – QSAR de regresión lineal múltiple (MCO).

4.3.3 Selección de Variables

Se creó una población inicial de modelos conformada por todas las combinaciones posibles de dos descriptores, generando una correlación lineal con la respuesta (modelos de baja dimensión). Se continuó la selección de variables con el método de algoritmos genéticos, utilizando como base la población de los 200 mejores modelos de baja dimensión, 2000 iteraciones para la evolución y una taza de mutación del 40%. La optimización de los modelos se llevó a cabo modificando el conjunto de datos hasta encontrar una relación entre los estadísticos R^2 y Q^2 satisfactorios. En la tabla 5 se muestra la evolución de las configuraciones para la selección de variables con sus respectivos resultados.

Tabla 5 Resultados de la selección de variables y sus respectivos criterios estadísticos.

Mejor modelo	***Full Model 1***	***Full Model 2***	***Full Model 3***	***Full Model 4***
Número de descriptores	7	7	10	9
Algoritmo explícito	pDL_{50}= 2.37 - 1.13ATSC6p + 1.03MATS4p + 0.36MATS7i + 1GATS4i + 0.52 JGI5 - 0.28CATSD_05_AL + 1 MLOGP	pDL_{50}= 3.56 - 2.07ATSC4p + 0.58 ATSC3i - 0.69MATS3m + 0.59MATS8m - 0.63GATS4m + 0.77 JGI5 + 1.36 MLOGP	pDL_{50}= 3.25 - 1.16ATSC4p - 0.43 MATS3m + 0.65MATS8e - 0.46MATS7p - 0.27GATS4m + 0.87 JGI5 + 0.58 P_VSA_i_4 - 0.31 CATS2D_03_DL + 1.14 MLOGP + 0.64 LLS_01	pDL_{50}= 4.86 + 0.85ATSC3m - 0.72ATSC6m - 0.81 MATS5e + 0.73 MATS8e + 0.32 P_VSA_MR_5 - 0.52 P_VSA_e_5 - 0.77 SaaaC - 1.5T(O..Cl) + 0.24B04[N-Cl]
R^2	0.593	0.607	0.789	0.816
Q^2_{loo}	0.540	0.553	0.737	0.777
Número de compuestos fuera del AD	3	3	2	3
ID de los Compuestos fuera del AD	8, 66, 81	8, 11, 81	61, 62	45, 62, 66
Moléculas totales para construcción del modelo	101	101	99 (8, 81 excluidos)	99 (8, 81 excluidos)

4.3.4 Dominio de Aplicación (AD)

Se analizó el dominio de aplicación de los mejores modelos en cada corrida mediante el enfoque de apalancamiento, (Sahigara et al. 2012) donde se usaron umbrales fijos para definir los valores atípicos estructurales y el enfoque de residuos estandarizados de respuesta, donde se consideran valores atípicos a las moléculas que presentan un residuo estandarizado mayor a 2.5 (ver apéndices). El mejor modelo se obtuvo en la corrida 4 *(Full Model* 4) donde se excluyeron del conjunto de entrenamiento las moléculas 8 y 81. En la figura 5 se muestra de

manera gráfica el perfil de la toxicidad oral aguda en rata de los compuestos químicos en estudio.

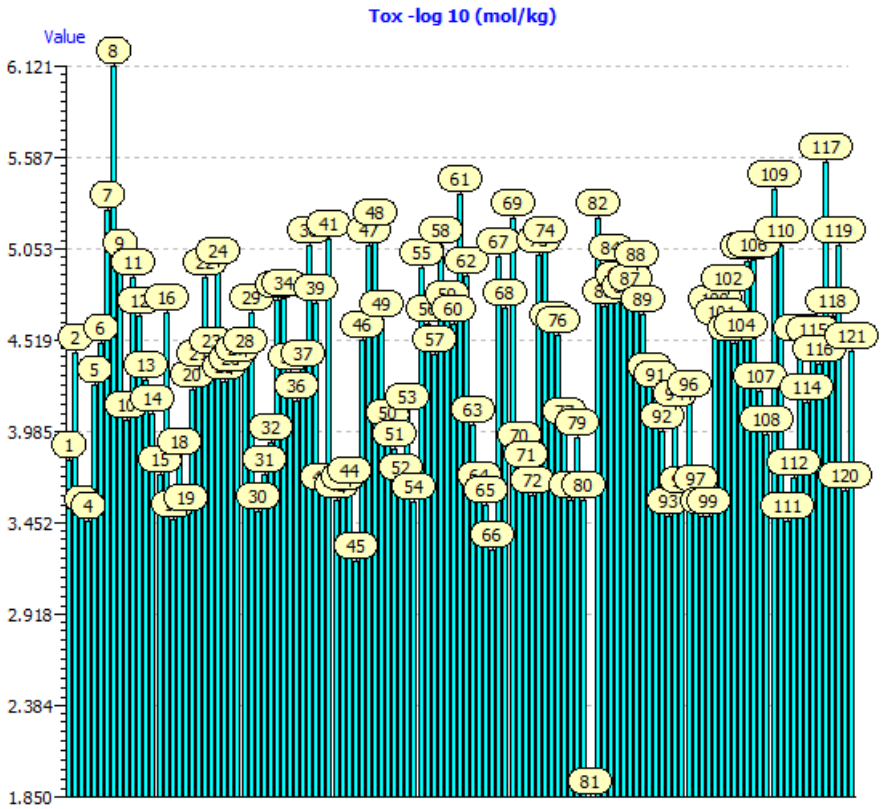
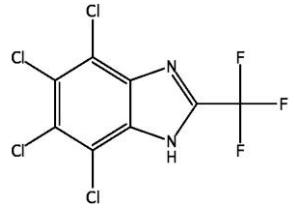
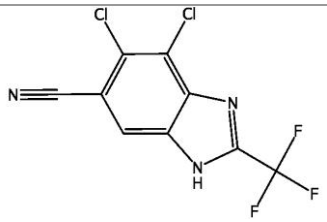


Figura 5 Distribución de los valores de la toxicidad oral aguda en rata para las moléculas seleccionadas de la base de datos T.E.S.T.

En la Tabla 6 se muestran la estructura de las moléculas excluidas para este modelo.

Tabla 6 Compuestos químicos excluidos del conjunto de entrenamiento con su respectivo ID, toxicidad y residuo estandarizado.

ID	Estructura	Toxicidad reportada -log10(mol/kg)	Toxicidad calculada -log10(mol/kg)	Residuo	Residuo *std*
8		6.12	4.86	-1.25	-3.12
81		1.85	4.13	2.28	5.60

Residuo std= residuo estandarizado

Las moléculas 8 y 81 del conjunto de entrenamiento tienen un residuo estandarizado mayor a 2.5 (5.60 y 3.12 respectivamente). Esto significa que las moléculas 8 y 81 no siguen la tendencia que el resto de las moléculas, predicción hecha por la ecuación del modelo para esas dos moléculas no es acertada, es decir la toxicidad calculada se aleja del valor real en más de 2.5 unidades de desviación estándar (umbral definido), por lo que se consideran valores atípicos de respuesta. En la figura 6 se muestra gráficamente el residuo estandarizado de todas las moléculas del conjunto de entrenamiento.

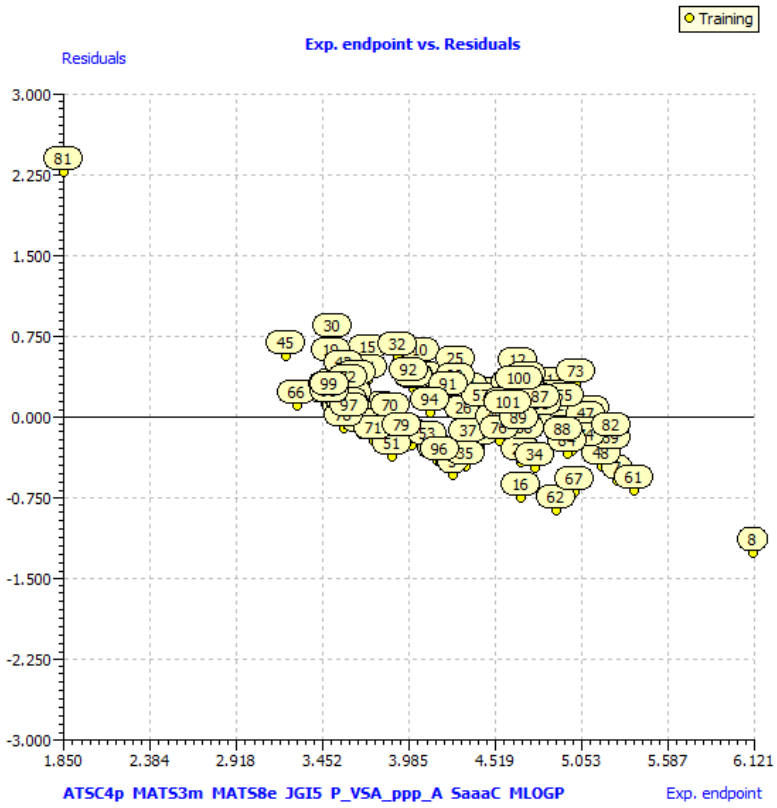


Figura 6 Residuos estandarizados del cálculo de toxicidad de las moléculas del conjunto de entrenamiento.

4.3.5 Modelo seleccionado

Al eliminar las moléculas 8 y 81 del conjunto de entrenamiento el espacio químico definido se redujo, pero el modelo mejoró sustancialmente, esto se ve reflejado en los criterios de ajuste (R^2 0.816) y validación interna (Q^2_{loo} 0.777). En la figura 7 se muestra el gráfico de dispersión de la toxicidad experimental contra la calculada por la ecuación del modelo.

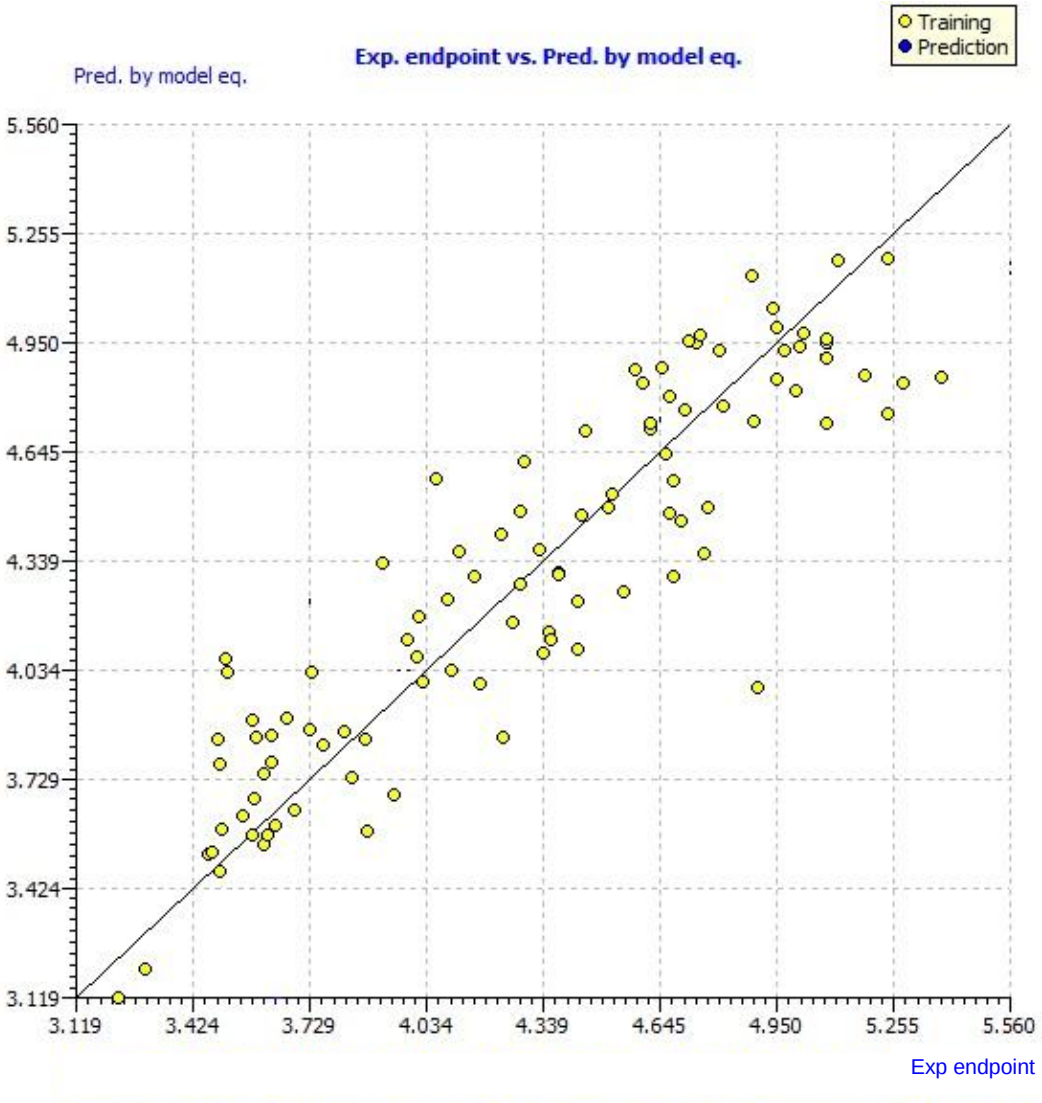


Figura 7 Gráfico de dispersión de la toxicidad experimental contra la calculada por la ecuación del modelo para el grupo de entrenamiento.

Los valores R^2 y Q^2_{LOO} más altos (de ajuste y validación cruzada) se encontraron con nueve descriptores. Con la ecuación generada se representa la mejor descripción para la actividad de los 2-trifluorometil benzimidazoles. En la figura 8 se muestra de forma gráfica este resultado.

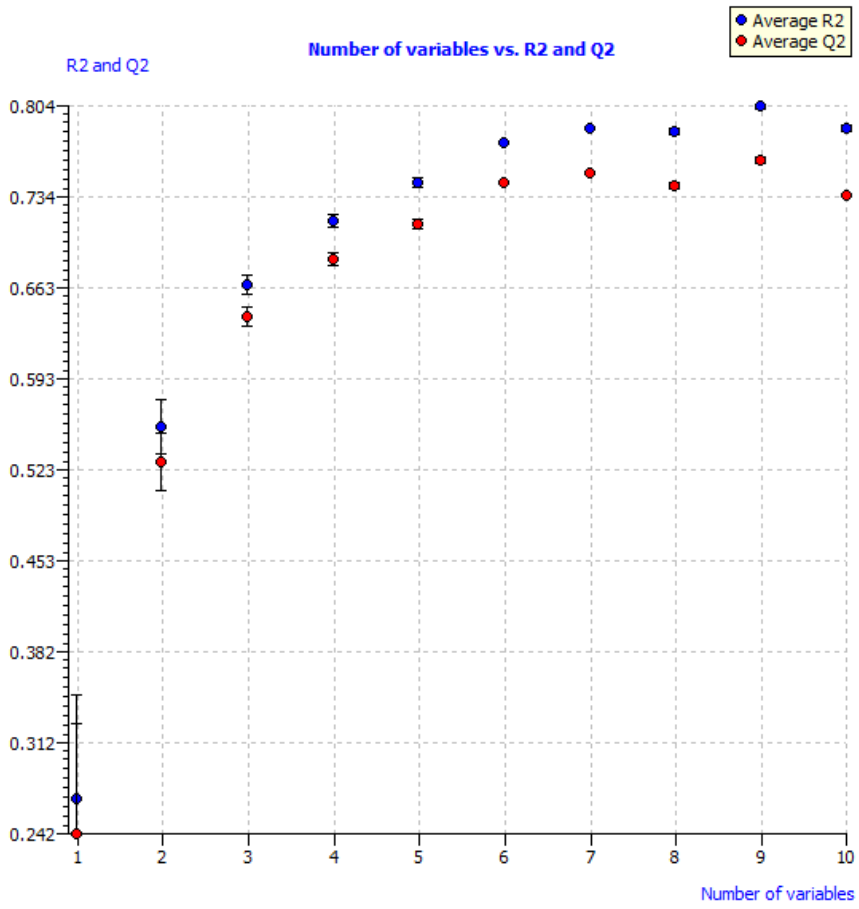


Figura 8 Gráfico de los valores medios (con su desviación estándar) de R^2 y Q^2_{LOO} en relación con el número de variables de modelado

Los valores de R^2 aumentan cuando se agregan descriptores al modelo, mientras que los valores de Q^2_{LOO} sólo aumentan hasta que se añaden descriptores útiles. Así, los descriptores innecesarios que se añaden al modelo (variables 8 y 10 en la Figura 6), hacen que Q^2_{LOO} disminuya, mostrando que el poder predictivo del modelo disminuye.

De esta manera, el mejor modelo tuvo los siguientes estadísticos de bondad de ajuste: R^2: 0.816; CCC tr: 0.8988 RMSEtr: 0.241

tr: del inglés training, conjunto de entrenamiento.

El algoritmo explícito para el mejor modelo es:

1. Modelo QSAR Dragon-Descriptor de la toxicidad por pDL_{50} oral en rata para moléculas de la clase 2-trifluorometil benzimidazol.

MCO - Modelo de regresión lineal múltiple desarrollado en un conjunto de entrenamiento de 101 compuestos químicos.

Ecuación completa del modelo:

pDL_{50}= 4.86 + 0.85 ATSC3m - 0.72 ATSC6m - 0.81 MATS5e + 0.73 MATS8e + 0.32 P_VSA_MR_5 - 0.52 P_VSA_e_5 - 0.77 SaaaC - 1.5 T(O..Cl) + 0.24B04[N-Cl]

Descriptores en el modelo:

- ATSC3m y ATSC6m: describen cómo se distribuye la masa a lo largo de la estructura topológica, es una autocorrelación espacial en un gráfico molecular definido. Autocorrelación centrada de Broto-Moreau del retraso 3 y 6 ponderado por masa (Autocorrelation Descriptors[1,13,12,11], 2018).
- MATS5e y MATS8e: miden la autocorrelación espacial basada en las ubicaciones y los valores de la electronegatividad de Sanderson (electronegatividad como una función de la densidad electrónica) del átomo. Autocorrelación de Moran del retraso 5 y 8 ponderado por la electronegatividad de Sanderson (Mauri, Consonni, & Todeschini, 2017).
- P_VSA_MR_5 y P_VSA_e_5: definen la cantidad de superficie de van der Waals (VSA) que tiene una propiedad P en un cierto rango. En Refractividad Molar (MR), bin 5 y electronegatividad de Sanderson (e), bin 5 (TALETE p_vsa_like_descriptors, 2018).
- SaaaC: índices del estado electrotopológico del átomo, el E-State de un átomo viene dado por su estado intrínseco más la suma de las perturbaciones en ese átomo por todos los demás átomos de la molécula. Suma de estados electrónicos aaaC (E-state Indices, 2018).
- T(O..Cl) y B04[N-Cl]: hace referencia a las distancias topológicas en la molécula. Suma de distancias topológicas entre O - Cl y presencia / ausencia de N - Cl a la distancia topológica 4 (TALETE,2018).

Como regla empírica, el conjunto de datos debe ser aproximadamente 5 veces mayor que el número de descriptores seleccionados para obtener buenos resultados (Tropsha and Golbraikh 2007).

Relación Compuestos químicos / Descriptores:

Modelo: 101 compuestos químicos / 9 descriptores = 11.22

El espacio de respuesta y descriptores se muestra en la siguiente tabla:

Tabla 7 Intervalo de valores de las variables involucradas en la ecuación del modelo.

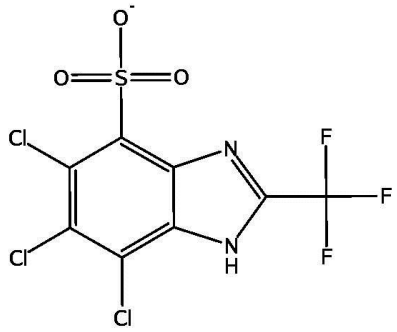
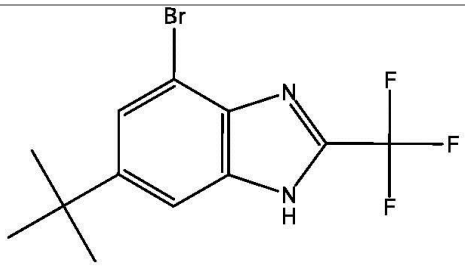
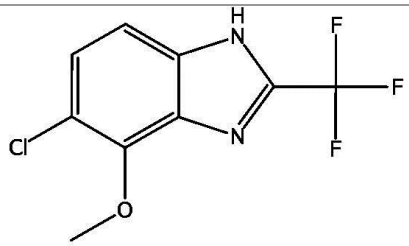
Variable	**Intervalo de valores**
pDL_{50} (log10 mol/kg)	3.229 - 5.56.
ATSC3m	1.31 - 146.94
ATSC6m	1.93 - 59.84
MATS5e	-0.61 - 0.09
MATS8e	-2.06 - 2.32
P_VSA_MR_5	29 - 64
P_VSA_e_5	0 - 101.49
SaaaC	-2 - 1
T(O..Cl)	0 - 45
B04[N-Cl]	0 - 1

Límites de aplicabilidad:

Como se indicó anteriormente, el dominio de aplicación estructural del modelo se evaluó mediante el enfoque de apalancamiento, proporcionando un valor de sombreado (HAT) de corte $h^* = 0.303$. Por otro lado, el dominio de aplicación de la respuesta se evaluó por el residuo estándar, con un valor de corte de 2.5 unidades de desviación estándar.

En la Tabla 8 se muestran las estructuras de los compuestos que están fuera del dominio de aplicación para el modelo seleccionado.

Tabla 8 Compuestos del grupo de entrenamiento fuera del dominio de aplicación del modelo desarrollado.

ID	Estructura	Nombre	Evaluación para AD
	Compuesto atípico estructural		**HAT (h*=0.3030)**
45		Ácido 5,6,7-Tricloro-2- (trifluorometil) -1H-bencimidazol-4-sulfónico	0.391
66		4-Bromo-6- (2-metil-2-propanil) -2- (trifluorometil) -1H-benzimidazol	0.595
	Compuesto atípico de respuesta		**Residuo estándar**
62		5-cloro-4-metoxi-2- (trifluorometil) -1H-benzimidazol	-3.699

4.3.6 Validación Interna

La estabilidad y robustez del modelo seleccionado se evaluó mediante los parámetros estadísticos Q^2_{LOO} y Q^2_{LMO}, donde se obtuvieron valores cercanos a la R^2 inicial del modelo (diferencia menor a 0.1). Este resultado indica que el modelo se mantiene con estadísticos en promedio similares cuando uno o más

compuestos del grupo de entrenamiento se excluyen en la construcción del modelo. Por otro lado, mediante la técnica *Y-scrambling* se demostró que el modelo no es el resultado de una correlación de azar, en donde se logró descartar la correlación entre las respuestas mezcladas y los descriptores, ya que los valores obtenidos disminuyeron drásticamente comparados con la R^2 inicial y los demás criterios de validación interna. Los resultados se muestran en la siguiente tabla:

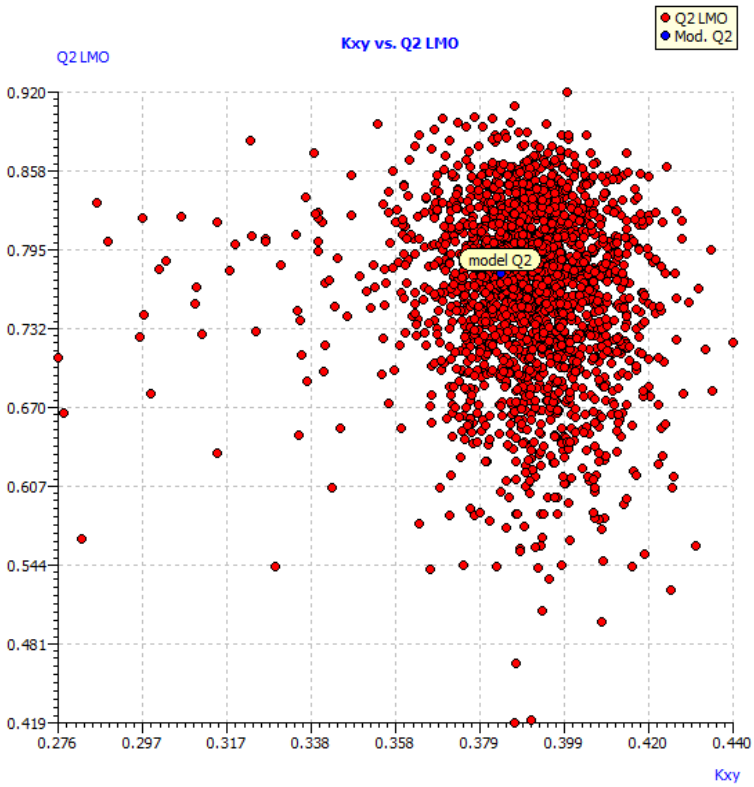
Tabla 9 Criterios estadísticos obtenidos en la validación interna

Técnica	**Criterios Estadísticos**		
Validación cruzada por leave-one-out	Q^2_{LOO}: 0.777	CCCcv: 0.878	RMSEcv: 0.265
Validación cruzada por leave-many-out	Q^2_{LMO}: 0.764		
Y-scrambling	R^2Yscr: 0.091	Q^2Yscr: -0.138	$RMSE_{Yscr}$: 0.535

RMSE: root-mean-square error (Raíz cuadrada del error medio)
Cv: croos-validation (Validación Cruzada)
Yscr: Y scrambling

A continuación, en la figura 9 se muestran las gráficas de Q^2_{LMO} y *Y-scrambling,* donde en el eje de las ordenadas se encuentra el valor promedio de Q^2_{LMO} (Figura A) y la correlación de R^2 y Q^2 para los modelos originales y los generados por *Y-scrambling* (Figura B), mientras que en el eje de las abscisas se encuentra Kxy que es la correlación entre los descriptores (x) del modelo y la toxicidad (y):

A)



B)

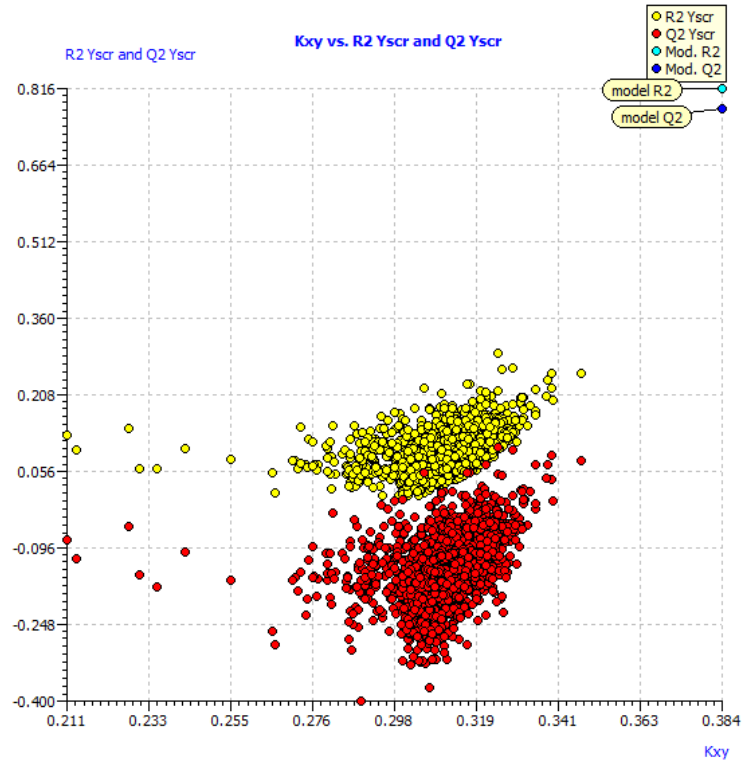


Figura 9 Validación interna del modelo **A)** Gráfico de dispersión de modelos LMO y **B)** *Y-Scrambling* comparado con el modelo de predicción.

4.3.7 Validación Externa

Una vez seleccionado el modelo con los mejores criterios de bondad de ajuste y validación interna, se incorporó el grupo de datos de validación externa para verificar el poder predictivo del modelo seleccionado.

Esta evaluación se realizó a través de un grupo de prueba que inicialmente estaba compuesto por 20 moléculas estructuralmente similares al grupo de entrenamiento (Figura 10) (grupo de prueba 1). Estas moléculas se sometieron a un análisis de dominio de aplicación para comprobar que estuvieran dentro del espacio químico del modelo seleccionado. Se encontraron 3 moléculas cuyos residuos estandarizados fueron mayores a 2.5 unidades de desviación estándar, por lo tanto, se clasificaron como valores atípicos de respuesta y fueron excluidas de la segunda evaluación del poder predictivo del modelo seleccionado (grupo de prueba 2). Los resultados se muestran a continuación.

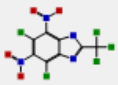
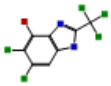
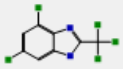
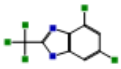
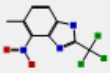
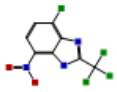
ID	NAME	Tox -log 10 (mol/kg)	ATSC3m	ATSC6m	MATS5e	MATS8e
114		3.4700	9.4130	5.1520	-0.3990	-1.9570
115		3.7170	14.0470	12.1450	-0.1050	0.1200
116		4.5100	9.7050	18.2290	-0.2090	-1.8660
117		4.1590	11.0940	17.5030	0.0540	-1.6500
118		4.5000	5.9820	17.3110	-0.1690	-1.1980
119		4.3800	2.8230	18.7810	-0.3020	-0.1010

Figura 10 Fracción de moléculas seleccionadas para la validación externa con su respectiva toxicidad y descriptores.

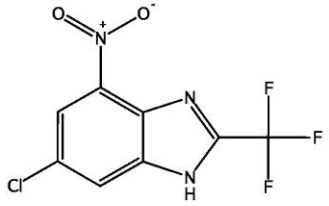
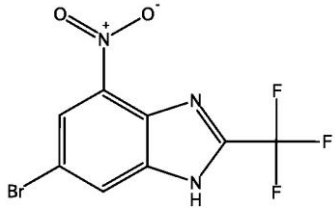
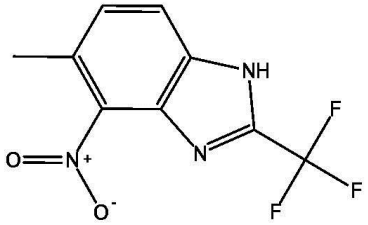
A continuación se calculan y comparan tres criterios para la evaluación del poder predictivo del modelo R^2_{ext}, Q^2_{F1} y CCC. Con lo que se demuestra que el modelo presenta un poder predictivo aceptable por la regulación OCDE. Los resultados y los criterios de la OCDE se muestran en la tabla 10.

Tabla 10 Criterios estadísticos aceptados por la OCDE y resultados obtenidos en la validación externa.

Validación externa	**Umbral de Valores OCDE**	**Grupo 1**	**Grupo 2**
Ecuación de modelo	Ecuación explicita	pDL_{50}= 4.86 + 0.85ATSC3m - 0.72ATSC6m - 0.81 MATS5e + 0.73MATS8e + 0.32P_VSA_MR_5 - 0.52P_VSA_e_5 - 0.77 SaaaC - 1.5T(O..Cl) + 0.24B04[N-Cl]	
R^2	>0.6	0.816	
Q^2_{loo}	>0.5	0.777	
Compuestos del grupo de prueba fuera del AD (ID)	Dominio de Aplicación definido	3 (105, 109, 115)	0
R^2_{ext}	>0.5	0.379	0.644
Q^2_{F1}	>0.6	0.303	0.647
CCC	-	0.561	0.789

Las moléculas 105, 109 y 115 fueron excluidas del grupo de prueba. En la tabla 11 se muestran las estructuras de estas moléculas que se encuentran fuera del dominio de aplicación.

Tabla 11 Compuestos químicos excluidos del grupo de prueba 1 con su respectivo ID, toxicidad y residuo estandarizado.

ID	Estructura	Toxicidad reportada -log10(mol/kg)	Toxicidad calculada -log10(mol/kg)	Residuo estandarizado
105		4.98	3.95	-4.13
109		5.4	4.14	-5.12
115		4.5	3.83	-2.70

Esto podría sugerir que el grupo nitro presente en la estructura (electro-atractor) incrementa la reactividad de la molécula debido a la formación de estructuras resonantes, y por lo tanto su toxicidad se ve modificada; o que las moléculas con un grupo nitro como sustituyente podrían causar toxicidad por otra vía. Dado que no se cuenta con descriptores que representen estas estructuras resonantes o que provean información suficiente para modelar la toxicidad de estas moléculas, la predicción para este tipo de estructuras es deficiente. Los resultados de la predicción usando la ecuación del modelo para el grupo de prueba 2 se presentan en la Figura 11 y la Tabla 12. Para el caso del gráfico C de la figura 11 la línea vertical corresponde al valor umbral HAT del dominio estructural (h *) mientras que

las líneas horizontales discontinuas son el umbral definido de desviación estándar para los valores Y-atípicos. Estos tienen un error cercano a cero que representa la capacidad de predicción de la ecuación QSAR final.

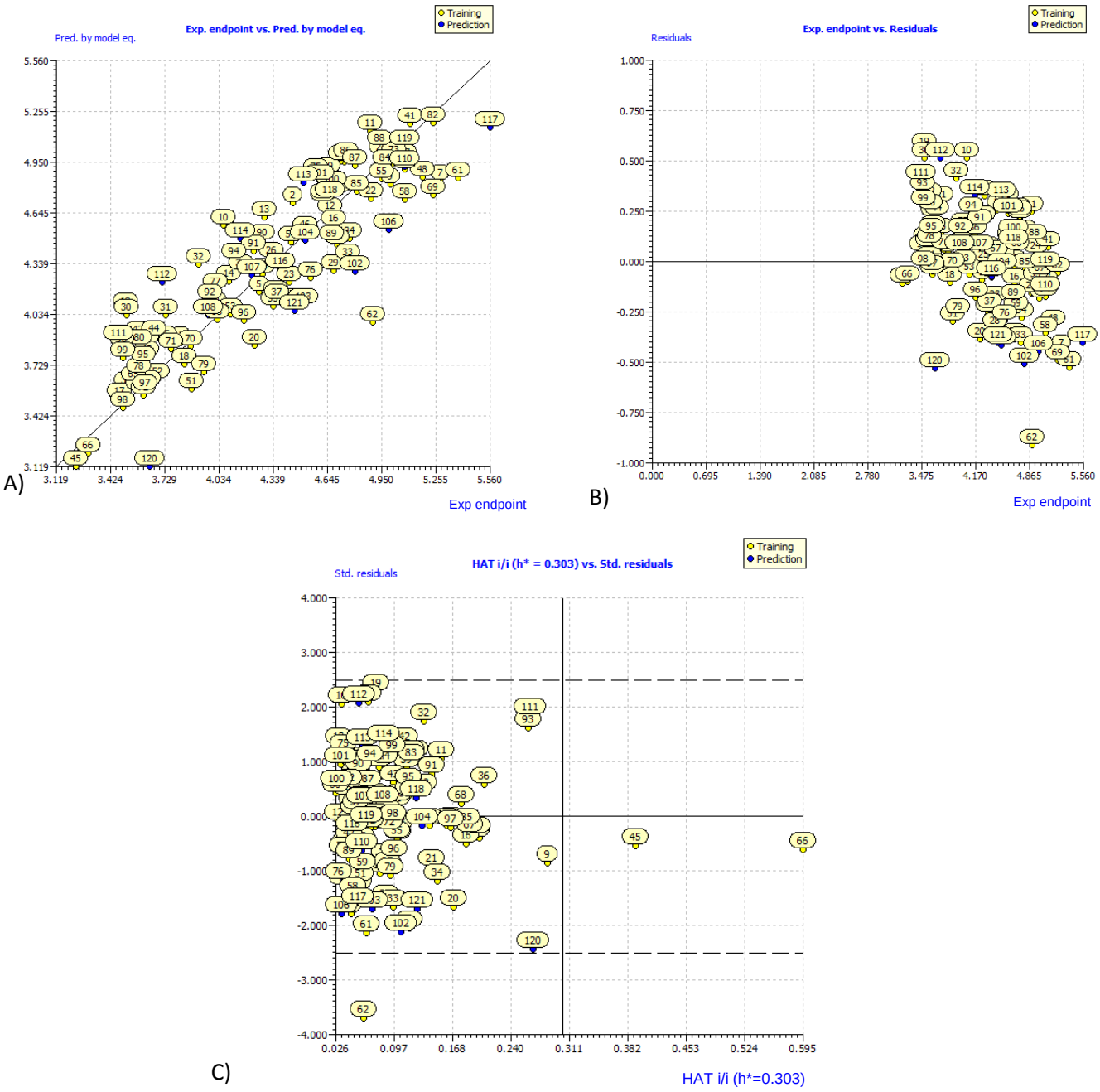


Figura 11 Gráficos de Validación externa y domino de aplicación del modelo con los datos extraídos de la base de datos T.E.S.T. **A)** Gráfico de dispersión de los datos experimentales frente a los predichos. **B)** Gráfico de la dispersión los residuos. **C)** Diagrama de Williams (AD del modelo).

Tabla 12 Grupo de moléculas consideradas para la validación externa con sus respectivas predicciones de toxicidad por la ecuación del modelo.

ID	Valor Experimental	Valor Calculado	Residuo de predicción	HAT i/i (h*=0.3030)	Residuo estándar
102	4.8	4.2935	-0.5065	0.1061	-2.108
103	4.51	4.0943	-0.4157	0.0707	-1.6967
104	4.52	4.4821	-0.0379	0.1317	-0.1599
106	4.99	4.5445	-0.4455	0.0343	-1.7836
107	4.22	4.2751	0.0551	0.0595	0.2236
108	3.976	4.0319	0.0559	0.0837	0.23
110	5.08	4.9268	-0.1532	0.0585	-0.621
111	3.47	3.8738	0.4038	0.2636	1.8513
112	3.717	4.2313	0.5143	0.0551	2.0815
113	4.51	4.8269	0.3169	0.0596	1.286
114	4.159	4.4905	0.3315	0.0851	1.3638
116	4.38	4.3082	-0.0718	0.0465	-0.2892
117	5.56	5.1597	-0.4003	0.0535	-1.6191
118	4.66	4.7393	0.0793	0.1252	0.3334
119	5.08	5.0497	-0.0303	0.0645	-0.1233
120	3.647	3.1188	-0.5282	0.2669	-2.4274
121	4.46	4.0597	-0.4003	0.126	-1.6847

Residuo de predicción: Valor experimental-Valor calculado

El diagrama de dispersión de las respuestas experimentales frente a las predichas detecta con facilidad las tendencias sistemáticas o el agrupamiento de datos y, en su caso, los valores extremos en los datos (ver gráfico A de la figura 11). En este caso, los datos están uniformemente distribuidos a lo largo del intervalo de respuesta, no se aprecian agrupados en una sola región y los únicos valores extremos son los ya identificados anteriormente para el grupo de entrenamiento que presentan un HAT mayor al umbral definido (moléculas 45, 62 y 66). Cabe destacar que la molécula 120 del grupo de prueba se aprecia desfasada de la tendencia ideal, de forma similar que en las moléculas 105, 109 y 115 presenta un grupo electro-atractor (sulfonamida). Se sugiere que el carácter ácido y la estabilidad de la deslocalización de los electrones de este grupo genera estructuras resonantes que aunado a la presencia de halógenos en la estructura modifican su toxicidad, aun así la molécula 120 cae dentro de los límites de

aplicación y su valor calculado por la ecuación del modelo no presenta una desviación estándar mayor a 2.5 unidades, su estructura se muestra en la figura 12.

La gráfica de los residuos (figura 11 B) permite evaluar las desviaciones de la coincidencia ideal entre datos experimentales y predichos y detectar tendencias anómalas. Nuevamente las moléculas 62 y 120 resultaron con un residuo negativo, lo que indica que la toxicidad calculada por la ecuación del modelo es menor a la experimental, es decir, subestima la actividad tóxica de estas moléculas.

Por otro lado, el gráfico de Williams detecta los valores atípicos de la respuesta (*Y-outliers*) y los de estructura (*X-outliers*). Consiste en trazar los residuos estandarizados en el eje “y” y los valores de apalancamiento de la diagonal de la matriz de “hat” en el eje “x” (figura 11 C). Con respecto a los residuos, todos los productos químicos caen en el umbral definido, excepto el compuesto 62 que se encuentra por debajo (2.5 unidades de desviación estándar) concluyendo que esta molécula no está bien predicha y por lo tanto es considerada como un valor atípico. Los valores de apalancamiento representan el grado de influencia que la estructura que cada producto químico tiene sobre el modelo. Un compuesto de alto apalancamiento en el conjunto de predicción se detecta lejos del dominio de aplicación de los compuestos de entrenamiento, por lo que podría conducir a datos predichos no fiables, siendo el resultado de una extrapolación sustancial del modelo. Por lo tanto, la información estructural de las moléculas incluidas en el conjunto de entrenamiento podría no ser suficiente para una predicción confiable de aquellos que se encuentran fuera del AD. No obstante, en este caso no se encontró ninguna molécula con un valor de apalancamiento mayor al umbral definido.

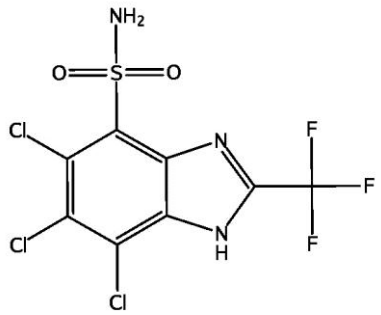


Figura 12 Estructura química de molécula con ID 120

4.3 Generación de modelos QSAR (PESTIMEP)

4.3 1 Datos PESTIMEP

Una vez ingresada la base de datos PESTIMEP al programa QSARINS fue tratada de la misma forma que la base de datos de T.E.S.T. En el paso de pre-reducción se excluyó el 57.76% de los descriptores (685 descriptores). Se continuó la selección de variables con 501 descriptores por el método de Algoritmos Genéticos para generar el modelo de regresión lineal múltiple (método de Mínimos Cuadrados Ordinario).

4.3.2. División de datos

Tomando en cuenta la gran diversidad estructural de las moléculas, como primer paso se clasificaron todas en el conjunto de entrenamiento (*full model*), donde los 145 compuestos contribuyen con sus características para la selección de variables. Esto permitiría contar con variables que explicaran mejor el comportamiento de todos los datos. Se evaluaron diferentes configuraciones para *full model* pero no se obtuvieron modelos satisfactorios por este método por lo que no se generó un grupo de moléculas para validar el modelo.

En la figura 13 se muestra la distribución de la toxicidad que presentan todas las moléculas en estudio.

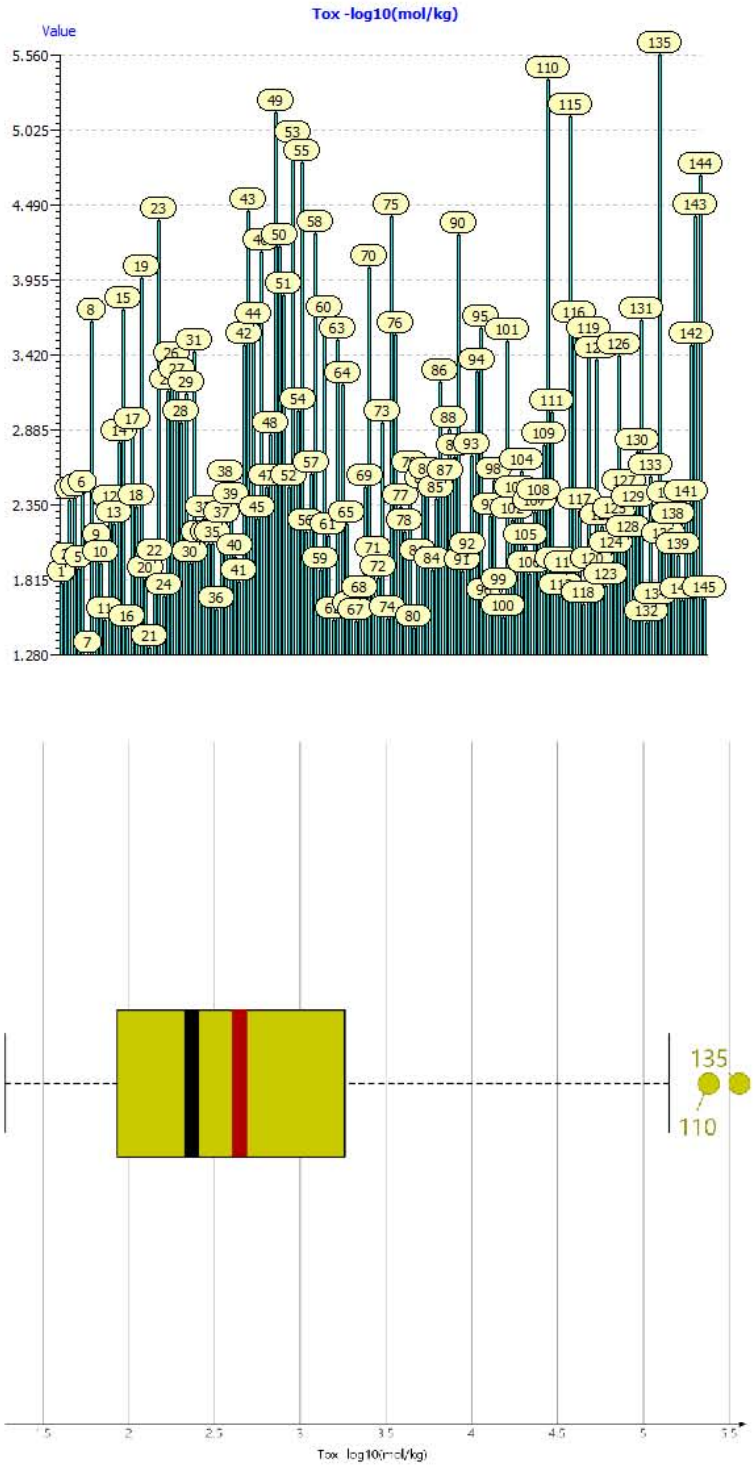


Figura 13 Intervalo de la toxicidad oral aguda en rata para los pesticidas de la base de datos PESTIMEP

4.3.3 Selección de Variables

Se comenzó explorando todos los modelos de baja dimensión (2D) y se continuó la selección de variables con el método de algoritmos genéticos, en la tabla 13 se muestran las mejores configuraciones de *full model* para la selección de variables con sus respectivos resultados.

Tabla 13 Resultados de la selección de variables y sus respectivos criterios estadísticos

Mejor modelo	***Full model 1***	***Full model 2***	***Full model 3***
Número de descriptores	7	8	7
Algoritmo explicito	pDL_{50}= 1.928 +1.74 ATSC3m + 0.77 GATS4p + 1.08 P_VSA_MR_5 -1.22 NsssN +1.88 CATS2D_04_AA - 0.94 CATS2D_08_LL -1 F04[C-N]	pDL_{50}= 0.86 + 2.49 ATSC3m + 2.27 GATS2m + 1.37 GATS4m – 1.60 C-002 - 1.72 NsssN + 1.69 CATS2D_04_AA - 1.55 CATS2D_02_AN - 0.67 CATS2D_08_LL	pDL_{50}= 2.42 +1.98ATSC3m - 0.88 GATS7m + 1.47 P_VSA_MR_5 - 1.20 H-050 -1.81 NsssN + 2.01 CATS2D_04_AL - 1.18 CATS2D_08_LL
R^2	0.4021	0.4563	0.4206
Q^2_{loo}	0.3246	0.3909	0.3294
Número de compuestos fuera del AD	6	7	5
Compuestos fuera del AD	11, 16, 63, 101, 113, 132	20, 27, 49, 53, 72, 110, 135	2, 17, 20, 135, 145
Moléculas totales para construcción del modelo	145	129 (1, 9, 13, 28, 31, 50, 63, 65, 78, 81, 94, 126, 131, 138 excluidos)	122 (7, 16, 62, 68, 100)

4.3.4 Dominio de Aplicación (AD)

El AD de los mejores modelos se analizó con los enfoques ya mencionados (apalancamiento y residuos estandarizados). Al excluir compuestos químicos que se encontraban fuera del dominio de aplicación, el espacio químico del modelo se reducía, pero la bondad de ajuste (R^2) y la validación interna (Q^2_{loo}) no mejoraban. Esta situación se atribuye a la diversidad no solo estructural, sino también de modos de acción y clases de pesticidas de la base de datos. El mejor modelo se obtuvo en la corrida 2, donde se excluyeron del conjunto de entrenamiento los compuestos con al menos un elemento metálico en su estructura, compuestos inorgánicos (excepto el monóxido de carbono) y compuestos de alto peso molecular con el fin de reducir la variabilidad de los datos. En la figura 14 se muestra el gráfico del peso molecular de los compuestos en estudio, los compuestos marcados presentan un alto peso molecular. En la figura 15 se muestran las estructuras de los compuestos químicos excluidos para este modelo.

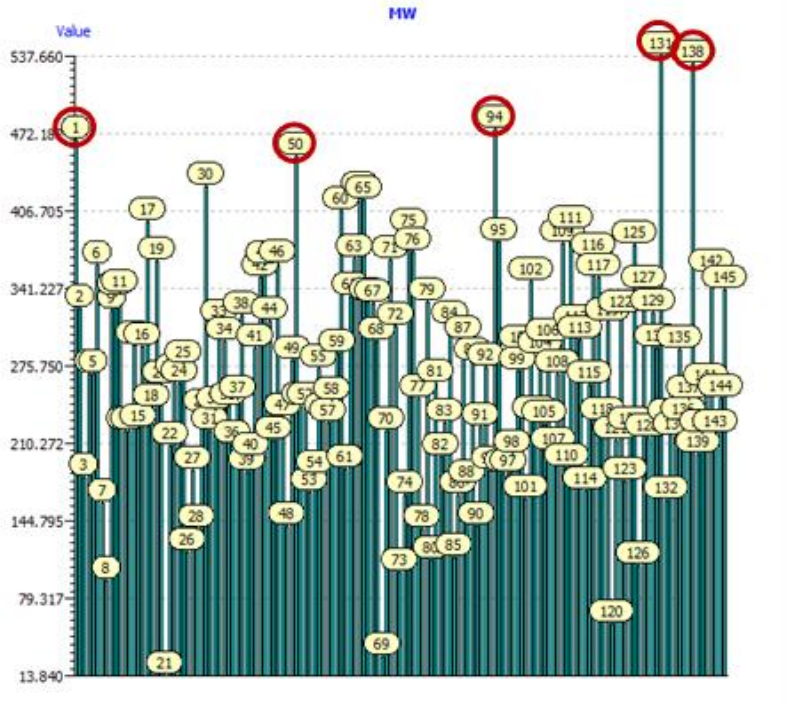
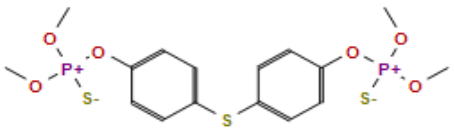
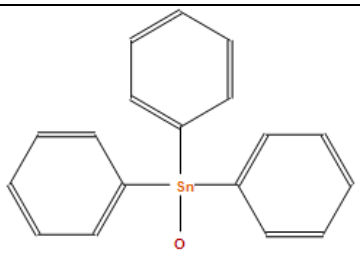
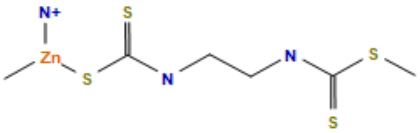
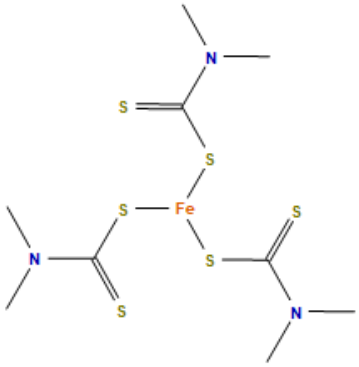
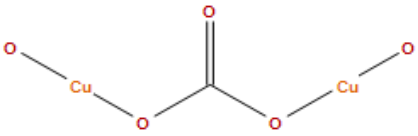
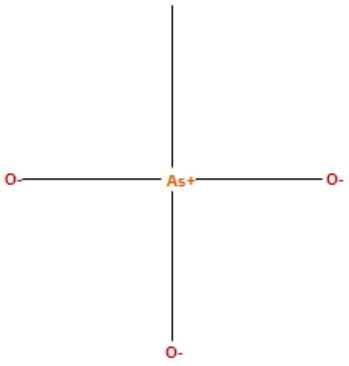
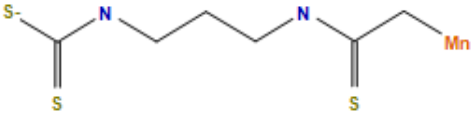


Figura 14 Intervalo de peso molecular para los pesticidas en estudio expresados en g/mol para la base de datos PESTIMEP.

ID	Estructura	ID	Estructura
1		63	
9		65	
13		78	
21		81	

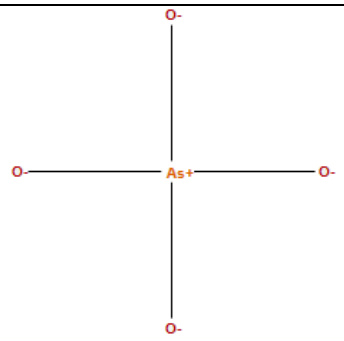
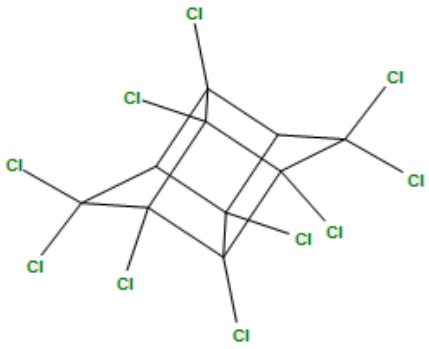
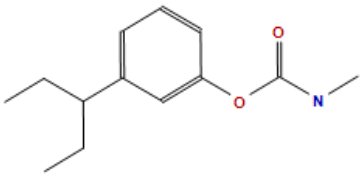
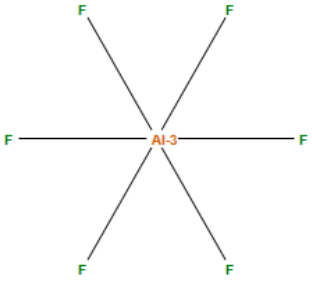
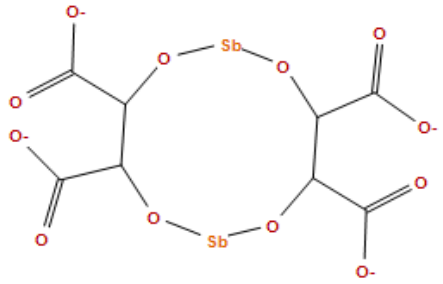
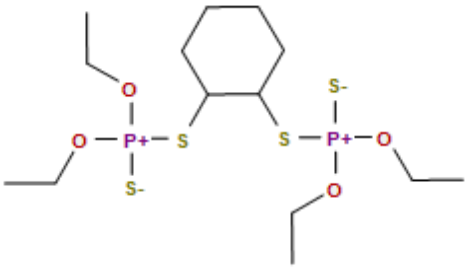
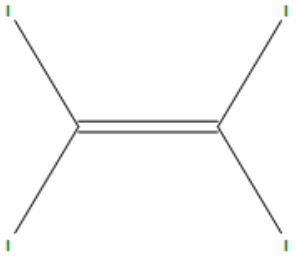
ID	Estructura	ID	Estructura
28		94	
31		126	
48		131	
50		138	

Figura 15 Compuestos químicos excluidos del conjunto de datos de PESTIMEP con su respectivo ID.

4.3.5 Modelo Seleccionado

No se encontró ninguna ecuación que describa adecuadamente la información en la base de datos PETIMEP. El mejor modelo presenta un criterio de ajuste igual a R^2 0.456 y validación interna Q^2_{loo} 0.390.

Los resultados de la generación del modelo se representan en la Figura 16 y la Tabla 14.

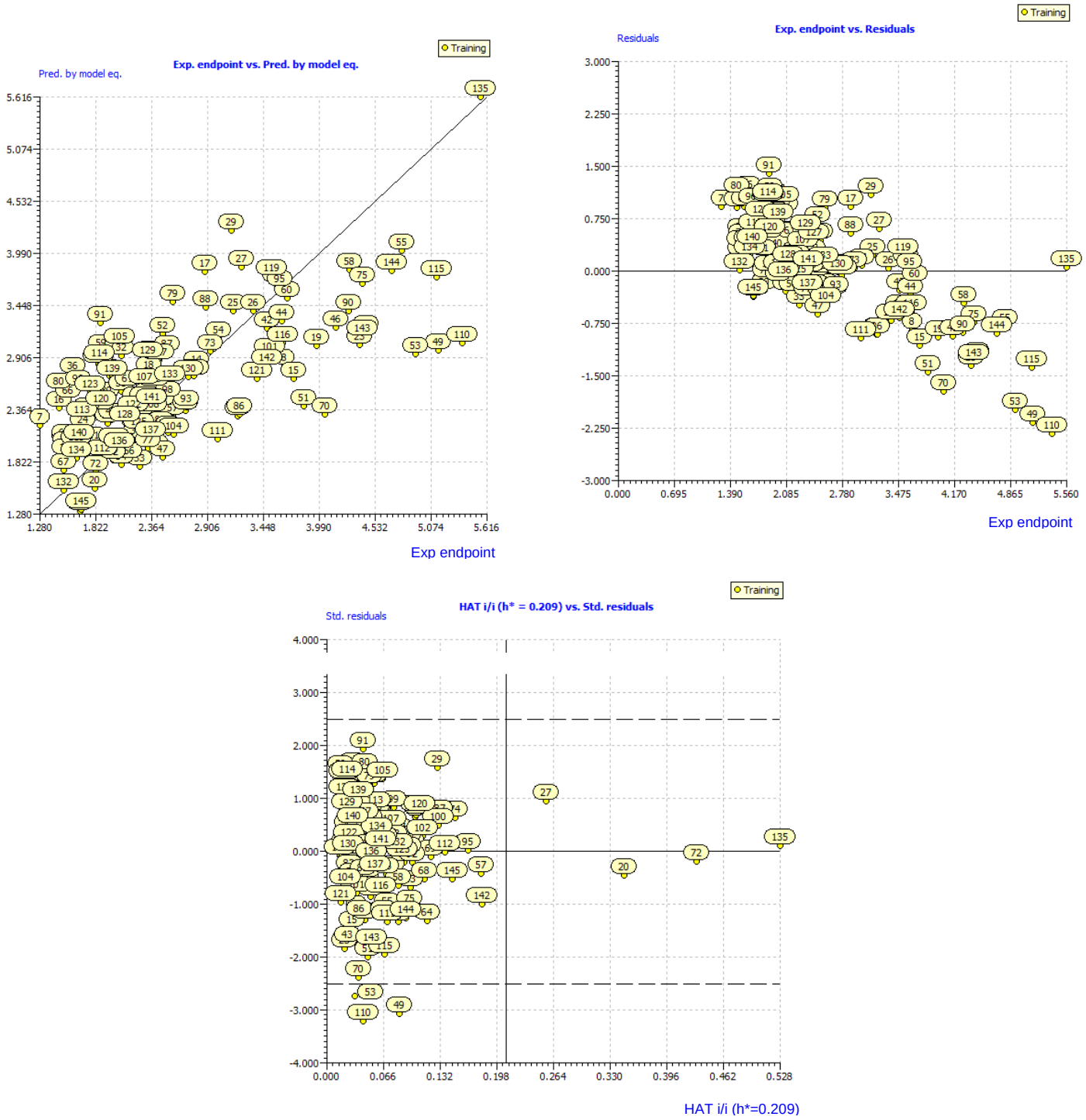


Figura 16 Gráficos de Validación externa y domino de aplicación del modelo de los datos extraídos de la base de datos PESTIMEP. **A)** Gráfico de dispersión de los datos experimentales frente a los predichos. **B)** Gráfico de la dispersión los residuos. **C)** Diagrama de Williams (AD del modelo).

.

Tabla 14 Resultados estadísticos para el mejor modelo de la base de datos PESTIMEP.

Variable	Coeff.	Std. coeff.	Std. err.	(+/-) Co. int. 95%	p-value
Intercept	0.8564		0.3632	0.7192	0.0199
ATSC3m	2.4932	0.3054	0.5844	1.157	0
GATS2m	2.2773	0.274	0.6279	1.2432	0.0004
GATS4m	1.3711	0.2501	0.3805	0.7533	0.0004
C-002	-1.6016	-0.2154	0.6106	1.209	0.0098
NsssN	-1.7243	-0.2673	0.4638	0.9184	0.0003
CATS2D_04_AA	1.6904	0.2226	0.5531	1.0951	0.0027
CATS2D_02_AN	-1.5516	-0.2424	0.4521	0.8951	0.0008
CATS2D_08_LL	-0.6718	-0.1563	0.3222	0.6379	0.0391
Bondad de ajuste					
R^2: 0.4563	RMSE tr:	0.7127			
Validación Interna					
Q^2_{loo}: 0.3909	RMSE cv:	0.7544			

El diagrama de dispersión de las respuestas experimentales frente a las predichas (figura 16 A) muestra que los datos no están uniformemente distribuidos a lo largo del intervalo de respuesta, se aprecian agrupados sobre todo en una sola región (poco tóxicos). Aunque para este modelo pocos compuestos se muestran fuera de los límites de aplicación (7 compuestos: 20, 27, 49, 53, 72, 110, 135) su valor de toxicidad calculado por la ecuación del modelo no es confiable. La tabla 15 muestra los resultados de la toxicidad calculada frente a la experimental, su residuo estándar y el valor de HAT. De estas 7 moléculas, 5 están subestimadas por la ecuación del modelo (el valor de la toxicidad pDL_{50} es menor que el valor experimental) lo que nos lleva a no usar este modelo ya que no es predictivo.

Tabla 15 Estadísticos para el mejor modelo de la base de datos PESTIMEP

ID	Tox. Experimental	Tox. Calculada	HAT i/i (h*=0.2093)	Residuo Estándar
20	1.8213	1.5557	0.3473	-0.4447
27	3.2404	3.8496	0.2565	0.956
49	5.1483	2.9858	0.0854	-3.0599
53	4.9275	2.9495	0.0335	-2.7227
72	1.8281	1.7251	0.4311	-0.1847
110	5.3818	3.061	0.0423	-3.2093
135	5.5596	5.6159	0.5281	0.1109

En la gráfica de los residuos (ver gráfico B de la figura 16) se aprecia una gran cantidad de moléculas que presentan un residuo negativo lo que indica que la toxicidad calculada por la ecuación del modelo es menor a la experimental, es decir, se subestima la toxicidad de estas moléculas. Este resultado es indeseable, ya que representa un riesgo al momento de tomar decisiones basadas en las predicciones hechas por este modelo. El gráfico de Williams detecta 4 compuestos fuera del dominio de aplicación por la técnica de apalancamiento (20, 27, 72, 135) y 3 con residuos estandarizados mayores a 2.5 unidades (49, 53, 110). Se encontraron moléculas con diferentes modos de acción y clases químicas con un solo representante en la base de datos, por lo que no se cuenta con suficiente información acerca de la relación estructura-toxicidad. Esto se ve reflejado en los estadísticos obtenidos de las diferentes configuraciones. No se obtuvo ningún modelo satisfactorio por el método de Regresión Lineal Múltiple (MCO). Esto se atribuye a la gran diversidad estructural, de modos de acción, tipos de pesticidas, y clases químicas presentes en la base de datos PESTIMEP.

5 Conclusiones.

Se desarrolló un modelo QSAR validado para la predicción de toxicidad oral aguda en rata en un grupo de moléculas usadas como pesticidas en el sector agroalimentario. La base de datos utilizada fue la del programa T.E.S.T., publicada por la US-EPA. El modelo QSAR final presenta estadísticos que cumple con los requerimientos de la OCDE para modelos validados: R^2 = 0.816; Q^2_{LOO} = 0.777; Q^2_{F1} = 0.647.

La ecuación del modelo es:
pDL_{50}= 4.86 + 0.85 ATSC3m - 0.72 ATSC6m - 0.81 MATS5e + 0.73 MATS8e + 0.32 P_VSA_MR_5 - 0.52 P_VSA_e_5 - 0.77 SaaaC - 1.5 T(O..Cl) + 0.24B04 [N-Cl].

Por otro lado, para el conjunto de pesticidas presentes en la base de PESTIMEP, no se logró desarrollar un modelo QSAR satisfactorio, por el método de regresión lineal múltiple (MCO). Para genera modelos QSAR con los compuestos de la base de datos PESTIMEP se sugiere:

1. Enriquecer el espacio químico (estructuralmente, de modos de acción y tipos de pesticidas).
2. Agregar moléculas que compensen la distribución a lo largo del intervalo de respuesta (pDL_{50}).
3. Usar métodos no lineales para encontrar correlación con la toxicidad oral aguda en rata.

Los resultados obtenidos de los grupos de datos utilizados, T.E.S.T. y PESTIMEP, subrayan la importancia de los datos utilizados para la generación de modelos predictivos. Además de las sugerencias mencionadas arriba para la base de datos PESTIMEP, este trabajo condujo a las siguientes perspectivas:

1. Generar modelos locales utilizando subgrupos de moléculas de la base de datos T.E.S.T.
2. Desarrollar modelos predictivos para moléculas de la clase química 2-trifluorometil benzimidazoles, que presenten en los sustituyentes grupos nitro u otros altamente electro-atractores.
3. Incorporar otras bases de datos utilizadas en toxicología computacional, como la llamada *Pesticide Properties Database* (PPDB), desarrollada por la IUPAC y la *Agricultural & Environmental Research Unit* (AERU) de la Universidad de Hertfordshire, Reino Unido.

6 Referencias

Ambure, Pravin, Rahul Balasaheb Aher, Agnieszka Gajewicz, Tomasz Puzyn, y Kunal Roy. 2015. "'NanoBRIDGES' Software: Open Access Tools to Perform QSAR and Nano-QSAR Modeling." *Chemometrics and Intelligent Laboratory Systems* 147 (Octubre): 1–13..

Autocorrelation Descriptors[1,13,12,11]. Retrieved October 5, 2018, from http://www.rguha.net/writing/notes/desc/node2.html

Bonabeau, Eric., Marco. Dorigo, y Guy. Theraulaz. 1999. *Swarm Intelligence : From Natural to Artificial Isystems*. Oxford University Press.

Cabrera-García Héctor. 2010. "Análisis de Pesticidas Organoclorados En Agua Utilizando Microextracción En Fase Sólida Del Espacio de Cabeza y Cromatografía de Gases Con Espectrometría de Masas(MEFS-CG-EM).Pdf." UNAM.

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*.

Clementi, Enrico. 1980. "Quantum Mechanical Calculations of Molecular Properties and Mulliken's Influence in Their Developments." *The Journal of Physical Chemistry* 84 (17): 2122–34.

Demchuk, Eugene, Patricia Ruiz, Selene Chou, and Bruce A. Fowler. 2011. "SAR/QSAR Methods in Public Health Practice." *Toxicology and Applied Pharmacology* 254 (2): 192–97.

Díaz Caraballo, José N. 2005. “Estadística Con Programación.” http://math.uprag.edu/residuales1.pdf.

E-state Indices. Retrieved October 5, 2018, from http://www.vcclab.org/lab/indexhlp/etstate.html

Evaluación toxicológica de impurezas en pesticidas - Azierta Science to Business. Acceso Octubre 2, 2018 https://azierta.eu/2017/12/11/evaluacion-toxicologica-impurezas-pesticidas/

Fao. ESTUDIO FAO PRODUCCIÓN Y PROTECCIÓN VEGETAL Manual sobre la elaboración y uso de las especificaciones de plaguicidas de la FAO y la OMS Tercera revisión de la primera edición. Acceso Octubre 2, 2018 http://www.who.int/whopes/quality/

García Hernández, Jaqueline, José Belisario LEYVA MORALES, Irma Eugenia MARTÍNEZ RODRÍGUEZ, María Isabel HERNÁNDEZ OCHOA, María Lourdes ALDANA MADRID, Aurora Elizabeth ROJAS GARCÍA, Miguel Betancourt Lozano, et al. 2018. “ESTADO ACTUAL DE LA INVESTIGACIÓN SOBRE PLAGUICIDAS EN MÉXICO.” *Rev. Int. Contam. Ambie* 34: 29–60.

Gomez-Jimenez, Gabriela, Karla Gonzalez-Ponce, Durbis J. Castillo-Pazos, Abraham Madariaga-Mazon, Joaquin Barroso-Flores, Fernando Cortes-Guzman, y Karina Martinez-Mayorga. 2018. “The OECD Principles for (Q)SAR Models in the Context of Knowledge Discovery in Databases (KDD).” *Advances in Protein Chemistry and Structural Biology*, Mayo 2018.

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.

Gramatica, Paola; Chirinco, Nicola; Papa, Ester; Cassani, Stefano; Kovarich, Simona. 2013. “QSARINS v 2.1 Manual.”

Gramatica, Paola. 2007. “Principles of QSAR Models Validation: Internal and External.” *QSAR and Combinatorial Science* 26 (5): 694–701.

Gramatica, Paola, Nicola Chirico, Ester Papa, Stefano Cassani, y Simona Kovarich. 2013. “QSARINS: A New Software for the Development, Analysis, and Validation of QSAR MLR Models.” *Journal of Computational Chemistry* 34 (24): 2121–32.

Gramatica, Paola, y Alessandro Sangion. 2016. “A Historical Excursus on the Statistical Validation Parameters for QSAR Models: A Clarification Concerning Metrics and Terminology.” *Journal of Chemical Information and Modeling* 56 (6): 1127–31

Haupt, Rl, y Se Haupt. 1998. “The Binary Genetic Algorithm.” *Practical Genetic Algorithms, Second …*, 27–50.

Helma, Christoph, Eva Gottmann, Stefan Kramer, y S Krämer. 2000. “Knowledge Discovery and Data Mining in Toxicology.” *Stat Methods Med Res* 9 (00): 329–58.

IRAIS, RAMÍREZ HERNÁNDEZ ARIADNA. 2014. “Desarrollo de Modelos Estadísticos y Moleculares de Saborizantes y Potencial Efecto de Saciedad.” UNAM.

Judson, Richard, Ann Richard, David J Dix, Keith Houck, Matthew Martin, Robert Kavlock, Vicki Dellarco, et al. 2009. “The Toxicity Data Landscape for Environmental Chemicals.” *Environmental Health Perspectives* 117 (5): 685–95.

Kavlock, Robert J., Christopher P. Austin, y Raymond R. Tice. 2009. “Toxicity Testing in the 21st Century: Implications for Human Health Risk Assessment.” *Risk Analysis* 29 (4): 485–87.

Leach, Andrew R., y Valerie J. Gillet. 2007. *An Introduction To Chemoinformatics*. Dordrecht: Springer Netherlands.

Maggiora, Gerald, Martin Vogt, Dagmar Stumpfe, y Jürgen Bajorath. 2014. "Molecular Similarity in Medicinal Chemistry." *Journal of Medicinal Chemistry* 57 (8): 3186–3204.

Martínez-Mayorga, Karina, Terry L. Peppard, Austin B. Yongye, Radleigh Santos, Marc Giulianotti, y Jose L. Medina-Franco. 2011. "Characterization of a Comprehensive Flavor Database." *Journal of Chemometrics*, no. Septiembre 2010: 550–60.

Mauri, A., Consonni, V., & Todeschini, R. (2017). 49 Molecular Descriptors. https://doi.org/10.1007/978-3-319-27282-5_51

Mignani, Serge, João Rodrigues, Helena Tomas, Rachid Jalal, Parvinder Pal Singh, Jean Pierre Majoral, y Ram A. Vishwakarma. 2018. "Present Drug-Likeness Filters in Medicinal Chemistry during the Hit and Lead Optimization Process: How Far Can They Be Simplified?" *Drug Discovery Today* 23 (3): 605–15.

Nandi, Dipankar, Pankaj Tahiliani, Anujith Kumar, y Dilip Chandu. 2006. "The Ubiquitin-Proteasome System." *J. Biosci.* 31 (1): 137–55.

"OECD.Org - OECD." Acceso Junio 20, 2018. http://www.oecd.org/.

Registro Sanitario de Plaguicidas y Nutrientes Vegetales | Comisión Federal para la Protección contra Riesgos Sanitarios | Gobierno | gob.mx. (n.d.). Acceso Octubre 2, 2018, https://www.gob.mx/cofepris/acciones-y-programas/registro-sanitario-de-plaguicidas-y-nutrientes-vegetales

Repetto, Guillermo., Ana del. Peso, y Jorge Luis. Zurita. 2000. *La Aplicación de Procedimientos in Vitro En La Evaluación Toxicológica Alimentaria.* Ediciones Díaz de Santos.

Ruiz, Patricia, Gino Begluitti, Terry Tincher, John Wheeler, y Moiz Mumtaz. 2012. "Prediction of Acute Mammalian Toxicity Using QSAR Methods: A Case Study

of Sulfur Mustard and Its Breakdown Products.” *Molecules* 17 (8): 8982–9001.

Rusyn, Ivan, y George P. Daston. 2010. “Computational Toxicology: Realizing the Promise of the Toxicity Testing in the 21st Century.” *Environmental Health Perspectives* 118 (8): 1047–50.

Sahigara, Faizan, Kamel Mansouri, Davide Ballabio, Andrea Mauri, Viviana Consonni, y Roberto Todeschini. 2012. “Comparison of Different Approaches to Define the Applicability Domain of QSAR Models.” *Molecules* 17 (5): 4791–4810.

TALETE, 7. List of molecular descriptors calculated by Dragon, List of molecular descriptors calculated by Dragon §.

Tropsha, Alexander, y Alexander Golbraikh. 2007. “Predictive QSAR Modeling Workflow, Model Applicability Domains, and Virtual Screening.” *Current Pharmaceutical Design* 13 (34): 3494–3504.

Zhu, Hao, Ivan Rusyn, Ann Richard, y Alexander Tropsha. 2008. “Use of Cell Viability Assay Data Improves the Prediction Accuracy of Conventional Quantitative Structure-Activity Relationship Models of Animal Carcinogenicity.” *Environmental Health Perspectives* 116 (4): 506–13.

7 Apéndices

7.1 Coeficiente de Tanimoto

La similitud molecular utilizando el índice de Tanimoto se calcula como:

$$S_{Tanimoto} = \frac{C}{A + B - C}$$

Donde A es el número de fragmentos estructurales presentes en la molécula A, B es el número de fragmentos estructurales presentes en la molécula B y C es el número de fragmentos estructurales en común. El coeficiente de Tanimoto oscila entre 0 y 1, donde 0 representa disimilitud y 1 representa máxima similitud (IRAIS 2014).

7.2 Metodologías empleadas en el paquete QSARINS

El paquete QSARINS emplea diversas técnicas de selección, evaluación de parámetros y se basa en diferentes métodos para la discriminación y selección tanto de variables, como de modelos generados. A continuación, se describen los métodos y técnicas empleados para la generación de los modelos reportados en esta tesis.

7.3 Pre selección de variables

El paquete QSARINS tiene la opción de realizar una pre reducción de variables para mitigar la redundancia de los descriptores intercorrelacionados, se realiza una pre reducción de descriptores en función de un objetivo de selección que usa solo las variables independientes (X). Los descriptores por descartar se identifican por:

1. Pruebas de valores idénticos (variables constantes). Los descriptores pueden eliminarse si el porcentaje de los compuestos que comparten el mismo valor es demasiado alto (ejemplo: 80%). Esta última opción es útil para los descriptores de contadores (por ejemplo, el número de átomos de carbono, el número de un grupo funcional, etc.).
2. Correlaciones en pares (de acuerdo con un valor de corte definido por el usuario, ejemplo: 95%). Se calcula la correlación entre todas las parejas de descriptores y, si una pareja se encuentra altamente correlacionada, el descriptor con mayor correlación con el resto de los descriptores es eliminado.

7.4 Métodos de modelado

En este paso es necesario aplicar un método cuantitativo capaz de encontrar la relación existente entre un número limitado de descriptores estructurales y la respuesta modelada. En QSARINS, el método utilizado es el enfoque de Regresión Lineal Múltiple (MLR) que puede ejemplificarse mediante la siguiente fórmula:

$$yi = b_0 + \sum_{j=1}^{n} bj\ xij + ei$$

donde se calcula una relación lineal entre las respuestas estudiadas (yi) y los valores seleccionados de los descriptores (xij); ei es el error aleatorio (llamado también residual del modelo). De este modo se estima la intercepción (b_0) y los coeficientes (bj). Donde y es el vector de respuestas, b el vector de los coeficientes, y e el vector de los errores. X es la matriz del modelo, donde las columnas son los descriptores. En este programa, para estimar el vector de los coeficientes, se utiliza la técnica de Mínimos Cuadrados Ordinarios (OLS por sus siglas en inglés):

$$\hat{b} = (X^T X)^{-1} X^T y$$

donde $\hat{b}$ es el vector que estima el vector b de los coeficientes, X^T la matriz X transpuesta y -1 es la operación de matriz inversa. OLS minimiza la suma de los cuadrados, de la diferencia entre las respuestas experimentales y las calculadas por el modelo. Para funcionar correctamente, el OLS supone que: (1) existe una relación lineal entre los descriptores y la respuesta, (2) los errores de respuesta son independientes y distribuidos de forma similar, (3) los descriptores no están demasiado correlacionados entre ellos, (4) hay más compuestos que descriptores de modelado (una relación que debe ser siempre superior a 5:1).

Una vez calculados los coeficientes del modelo, es posible obtener el vector de $\hat{y}$, como en la siguiente fórmula:

$$\hat{y} = X\hat{b} = X(X^T X)^{-1} X^T y = Hy$$

donde H es la matriz de apalancamiento que relaciona las respuestas calculadas y las experimentales. Los elementos diagonales de la matriz HAT hii son útiles para determinar la distancia del objeto i desde el centro del espacio químico del modelo, así, para comprobar el AD estructural del modelo (Gramatica et al. 2013).

7.5 Selección de descriptores por Algoritmo Genético

QSARINS está habilitado para seleccionar variables por Algoritmo Genético (Haupt and Haupt 1998) Esta técnica permite explorar una amplia gama de soluciones buscando la optimización del problema, maximizando (o minimizando) una función de aptitud seleccionada (R^2, Q^2). Esto se hace imitando la selección natural, donde las mejores soluciones sustituyen a las menos productivas. En términos biológicos, se diría que los mejores genes de la población sustituyen a los menos adecuados. En nuestro caso, cada descriptor representa un gen, y un conjunto de descriptores representa un cromosoma. La aptitud de un cromosoma está relacionada con el correspondiente desempeño del modelo. (Gramatica et al. 2013)

El AG comienza definiendo un cromosoma o una matriz de variables que se optimizarán. Si el cromosoma tiene variables N_{var} (un problema de optimización dimensional N_{var}) dado por $p1, p2, ..., pN_{var,}$ entonces el cromosoma se escribe como un vector de fila de elementos N_{var}.

$$\text{cromosoma} = [p1, p2, p3, ..., pNvar]$$

Cada cromosoma tiene una calificación al evaluar la función de aptitud.

El valor cuantificado del gen o variable se encuentra matemáticamente multiplicando el vector que contiene los bits por un vector que contiene los niveles de cuantificación:

$$q_n = gen \times Q^T$$

Donde
gen = $[b_1\ b_2 . . . b_{Ngen}]$
N_{gen} = número de bits en un gen
b_n = bit binario = 1 o 0
Q = Vector de cuantificación = $[2^{-1}\ 2^{-2} . . . 2^{Ngen}]$
Q^T = transpuesta de Q

La población es el grupo de cromosomas con el que comienza el AG. La población tiene cromosomas N_{pop} y es una matriz $N_{pop} \times N_{bits}$ llena de aleatorias y ceros generados usando

$$pop = round\ (rand(N_{pop}, N_{bits}))$$

donde la función (N_{pop}, N_{bits}) genera una matriz $N_{pop} \times N_{bits}$ de números aleatorios uniformes entre cero y uno.

La supervivencia del más apto se traduce en descartar los cromosomas con la menor aptitud. En primer lugar, la aptitud de N_{pop} y los cromosomas asociados se clasifican de menor aptitud a mayor aptitud. Luego, solo los mejores se

seleccionan para continuar, mientras que el resto se eliminan. La tasa de selección, Xrate, es la fracción de N_{pop} que sobrevive para el siguiente paso de apareamiento. La cantidad de cromosomas que se conservan en cada generación es:

$$N_{Keep} = X_{rate} N_{pop}$$

La selección natural ocurre en cada generación o iteración del algoritmo. De los cromosomas N_{pop} en una generación, solo el N_{keep} superior sobrevive para el apareamiento, y el inferior N_{pop} - N_{keep} se descarta para dejar espacio a la nueva descendencia (Haupt and Haupt 1998).

Hay dos técnicas: ponderación de rango y ponderación de aptitud.

Las mutaciones aleatorias alteran un cierto porcentaje de los bits en la lista de cromosomas. La mutación es la segunda forma en que el AG explora una superficie de soluciones. Puede introducir rasgos que no están en la población original y evita que el AG converja demasiado rápido antes de muestrear toda la superficie de soluciones (Haupt and Haupt 1998).

En QSARINS, con el fin de evitar un comienzo completamente aleatorio del AG, el mejor conjunto de descriptores extraídos de la evaluación de todo el subconjunto de pequeña dimensión (2D) se utiliza como el núcleo de los cromosomas de la población inicial (Gramatica et al. 2013).

7.6 Validación interna

Durante el cálculo del modelo los puntos clave para verificar son el ajuste y la robustez de este, para lo cual se utiliza el coeficiente de correlación múltiple R^2 y otras técnicas de validación interna (Q^2_{LOO}, Q^2_{LMO}, $Y_{scrambling}$) respectivamente.

El valor del coeficiente de correlación (R^2) es una medida de cuan exactamente los datos experimentales concuerdan con los datos calculados por la ecuación del

modelo. Este criterio tiene un valor entre cero y uno. Si R^2 = 1 quiere decir que el ajuste entre los datos experimentales y calculados es perfecto.

A continuación, se muestra la ecuación de R^2:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(\widehat{yi} - yi)^2}{\sum_{i=1}^{n}(yi - \overline{yi})^2} = 1 - \frac{RSS}{TSS}$$

Donde $\widehat{yi}$ es la respuesta calculada por la ecuación del modelo para el grupo de entrenamiento, $\overline{yi}$ es la respuesta experimental promedio del grupo de entrenamiento, yi la respuesta experimental, RSS es la suma de cuadrados residuales, y TSS es la suma total de cuadrados para n elementos del conjunto de datos (Gramatica and Sangion 2016).

Los métodos de validación cruzada o interna proporcionan una forma para superar algunos de los problemas inherentes al uso del valor R^2. Una de las técnicas de validación cruzada implica la eliminación de algunos de los valores del conjunto de datos de entrenamiento, la derivación de un modelo QSAR utilizando los datos restantes, y luego la aplicación de este modelo para predecir los valores de los datos que se han eliminado (LOO-Leave one out, LMO-Leave Many Out). La forma más simple de validación cruzada es el enfoque de dejar uno fuera (LOO), donde se elimina el valor de un dato iterativamente del conjunto de entrenamiento (Leach and Gillet 2007).

La repetición de este proceso para cada valor en el conjunto de datos conduce a un R^2 de validación cruzada (más comúnmente escrito como Q^2). El valor de Q^2 de validación cruzada es normalmente más bajo que el R^2 original del modelo, pero debe tener un valor comparable a R^2 (diferencia menor a 0.1) lo que significaría que las predicciones internas son buenas y el modelo se considera internamente estable o robusto (Gramatica et al. 2013).

El criterio Q^2_{LMO} excluye iterativamente al azar un cierto porcentaje de datos del conjunto de entrenamiento, construye un modelo con los datos restantes y evalúa los compuestos excluidos. Es una técnica más fuerte que LOO y estudia el

comportamiento del modelo cuando se excluye un mayor número de compuestos (Gramatica et al. 2013).

El modelo en análisis puede considerarse estable si los valores de R^2 y Q^2 calculados en cada iteración de LMO y sus promedios (R^2_{LMO} y Q^2_{LMO}) están próximos a los valores R^2 y Q^2_{LOO} del modelo original (Gramatica et al. 2013).

A continuación, se muestra la ecuación de Q^2_{LOO}:

$$Q^2_{LOO} = 1 - \frac{\sum_{i=1}^{n}(\hat{y}i/i - yi)^2}{\sum_{i=1}^{n}(yi - \overline{yi})^2} = 1 - \frac{PRESS}{TSS}$$

Donde yi es la respuesta experimental, $\hat{y}i/i$ es el valor predicho de la respuesta calculada por LOO (excluyendo el i-ésimo elemento del cómputo del modelo o excluyendo más de un elemento en cada iteración (LMO)), $\overline{yi}$ es el promedio de la respuesta experimental del grupo de entrenamiento, PRESS es la suma de cuadrados de error predictivo, y TSS es la suma total de cuadrados para n elementos del conjunto de datos de entrenamiento (Gramatica and Sangion 2016).

Para demostrar que el modelo no es el resultado de la correlación al azar, se aplica la técnica de $Y\text{-}_{scrambling}$. En esta técnica, las respuestas se mezclan al azar, por lo que no debe existir correlación entre ellas y los descriptores. Como consecuencia, el coeficiente de correlación ($R^2_{YScrambling}$ y $Q^2_{YScrambling}$) modificados deberían disminuir drásticamente. En este caso, si el modelo original en validación es bueno, los valores de R^2 y Q^2 de cada iteración, y sus promedios (R^2_{YS} y Q^2_{YS}), deben ser mucho menores de los valores del modelo original (Gramatica and Sangion 2016).

7.7 Validación Externa

El objetivo de la validación externa es comprobar la capacidad del modelo para predecir nuevos compuestos con exactitud. Esto se hace aplicando la ecuación del modelo, obtenida a través del conjunto de entrenamiento, a uno o más conjuntos de datos de predicción, es decir, a los compuestos excluidos que nunca se han utilizado ni para la selección de variables, ni para el cálculo del modelo, y midiendo los resultados por medio de diferentes criterios, tales como: $RMSE_{EXT}$, R^2_{EXT}, Q^2_{F1}, Q^2_{F2}, Q^2_{F3} y CCC.

Es importante recordar que el conjunto de predicción no influye sobre la presencia o ausencia de un descriptor molecular en el modelo final, ya que nunca participa en el proceso de selección de variables. La primera medida de validación externa será el coeficiente de correlación calculado comparando los datos experimentales contra los predichos del conjunto de prueba: R^2_{EXT} que tiene la siguiente formula:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(\widehat{yiext} - yiext)^2}{\sum_{i=1}^{n}(yiext - \overline{yiext})^2} = 1 - \frac{RSS}{TSS}$$

Donde $\widehat{yiext}$ es la respuesta calculada por la ecuación del modelo para el grupo de prueba, $\overline{yi}$ es la respuesta experimental promedio del grupo de prueba, yi la respuesta experimental, RSS es la suma de los cuadrados residuales, y TSS es la suma total de cuadrados para n elementos del conjunto de datos (Gramatica and Sangion 2016).

El criterio ampliamente usado por diferentes autores en la literatura QSAR y también sugerido en el documento de orientación de la OCDE sobre la validación externa de los modelos QSAR, tiene la siguiente formulación:

$$Q^2_{F1} = 1 - \frac{\sum_{i=1}^{next}(\widehat{yi} - yi)^2}{\sum_{i=1}^{next}(yi - \overline{yTr})^2}$$

La forma es similar a Q^2_{LOO}, pero aquí las sumas están sobre los elementos del conjunto de predicción externo y, en lugar de PRESS (que se calcula utilizando el conjunto de entrenamiento en validación cruzada), se utilizan la suma de las diferencias al cuadrado entre yi los valores experimentales y los calculados por el modelo $\widehat{yi}$. En el denominador, se usa la suma de las diferencias cuadradas de los valores experimentales yi y el promedio del conjunto de entrenamiento $\overline{yTr}$, en lugar del TSS simple (que se calcula utilizando los valores del conjunto de entrenamiento). El uso del promedio de los valores del conjunto de entrenamiento, en lugar del conjunto de predicción, es una forma de mantener un registro de la "distancia" entre los dos conjuntos.

Schüürmann et al. propusieron un criterio alternativo:

$$Q^2_{F2} = 1 - \frac{\sum_{i=1}^{next}(\widehat{yi} - yi)^2}{\sum_{i=1}^{next}(yi - \overline{yext})^2}$$

Q^2_{F2} es similar a Q^2_{F1} solo que en el denominador de este criterio se calcula el valor promedio utilizando el conjunto de predicción $(\overline{yext})$ en lugar del de entrenamiento.

El método no toma en cuenta la "distancia" del promedio de los valores del conjunto de entrenamiento, obteniendo independencia de este.

El siguiente criterio Q^2_{F3} propuesto por Consonni et al. tiene la siguiente expresión:

$$Q^2_{F3} = 1 - \frac{\sum_{i=1}^{next}(\widehat{yi} - yi)^2/next}{\sum_{i=1}^{next}(yi - \overline{yTr})^2/nTr}$$

Donde tanto el numerador como el denominador se dividen entre el número de elementos correspondientes del conjunto de prueba $next$ y entrenamiento nTr. La ventaja de Q^2_{F3} radica en ser independiente del tamaño de muestreo.

El criterio *CCC* propuesto por Lin tiene la siguiente expresión:

$$CCC = \frac{2\sum_{i=1}^{n}(xi - \bar{x})(yi - \bar{y})}{\sum_{i=1}^{n}(xi - \bar{x})^2 + \sum_{i=1}^{n}(yi - \bar{y})^2 + n(\bar{x} - \bar{y})^2}$$

Donde x e y corresponden a los valores de abscisas y ordenadas del gráfico que representa los valores de datos experimentales frente a los calculados usando el modelo, n es el número de productos químicos, y $\bar{x}$ e $\bar{y}$ corresponden a los promedios de los valores de abscisas y ordenadas, respectivamente.

Este coeficiente mide tanto la precisión (cuán lejos están las predicciones de la línea de ajuste) como la exactitud (hasta qué punto la línea de regresión se desvía de la línea de pendiente 1 que pasa por el origen, es decir, la línea de concordancia ideal), por consiguiente, cualquier divergencia de la línea de regresión a la línea de concordancia da como resultado un valor de CCC menor que 1.

La raíz cuadrada del error medio en predicción, formulado como:

$$RMSE = \sqrt{\sum_{i=1}^{next} \frac{(\widehat{yi} - yi)^2}{next}}$$

mide las discrepancias entre los valores experimentales y los predichos por el modelo, por lo tanto, yi corresponde a los valores experimentales mientras que $\widehat{yi}$ a los calculados por el modelo para el conjunto de prueba, $next$ es el número de compuestos presentes en el conjunto de prueba. Es importante señalar que este criterio no es útil para evaluar y comparar diferentes modelos porque depende de la escala de medida de las respuestas modeladas (por lo tanto, no es una medida relativa). Sin embargo, RMSE siempre debe calcularse tanto para el conjunto de entrenamiento como para los conjuntos de predicción, como un criterio adicional para la medición absoluta del error. Los datos predichos por cualquier modelo QSAR para el conjunto de compuestos químicos externos, deben, en el máximo

nivel, estar en completo acuerdo con los datos reales de esos compuestos químicos, que nunca fueron utilizados en el desarrollo del modelo.

7.8 Dominio de Aplicación

El dominio de aplicación (AD) se define como "el espacio de respuesta y estructura química en el que el modelo QSAR hace predicciones con una fiabilidad determinada". (Ambure et al. 2015). Los modelos de predicción deben usarse para predecir compuestos químicos estructuralmente similares, ya que, las predicciones son más confiables para compuestos externos que caen dentro de los límites estructurales del modelo generado.

En el paquete QSARINS el dominio de aplicación estructural se define a través del método de apalancamiento (HAT matriz del sombrero) y el dominio de la respuesta mediante residuos estandarizados.

Como se vio antes es posible obtener el vector de $\hat{y}$, donde H es la matriz de apalancamiento (o HAT) que relaciona las respuestas calculadas y las experimentales. Los elementos diagonales de la matriz HAT hii son útiles para determinar la distancia del objeto i desde el centro del espacio químico del modelo, así, para comprobar el AD estructural del modelo (Gramatica et al. 2013).

Destacando aquellos objetos (compuestos químicos) con valor h > h *, definido como:

$$h^* = 3p'/n$$

donde p' es el número de variables del modelo + 1 y n es el número de compuestos en el conjunto de entrenamiento.

El residuo es la diferencia entre el valor estimado por la línea de regresión y el valor experimental. El residuo estandarizado, se obtiene de:

$$ri = \frac{e_i}{\sqrt{s^2\,(1-h_i)}}$$

Donde e_i es el i ésimo residuo, h_i i ésimo elemento diagonal de $X(X^TX)–1X^T$ y s^2 es el cuadrado medio del error. X es la matriz de diseño y X^T la matriz de diseño transpuesta. El análisis de residuales permite cotejar si las suposiciones del modelo de regresión se cumplen. Se puede detectar: a) Si efectivamente la relación entre las variables X e Y es lineal, b) Si hay normalidad de los errores. c)Si hay valores anormales en la distribución de errores (Si se usa Residual estandarizado, cualquier observación con un residual mayor de 2 o menor de -2 es considerado “outlier”) d) Si hay varianza constante (propiedad de Homocedasticidad) y e) Si hay independencia de los errores. (Díaz Caraballo 2005).