UNIVERSIDAD AUTÓNOMA DEL ESTADO DE HIDALGO

INSTITUTO DE CIENCIAS BÁSICAS E INGENIERÍA

MAESTRÍA EN QUÍMICA

TESIS

DETERMINACIÓN DE MODELOS DE TOXICIDAD EMPLEANDO DFT E INTELIGENCIA ARTIFICIAL EN COMPUESTOS ORGANOTIOFOSFORADOS

Para obtener el grado de
Maestro en Química

PRESENTA

Q. Uriel Josafat Rangel Peña

Director
Dr. Julián Cruz Borbolla

Codirectora
Dra. Rosa Luz Camacho Mendoza

Comité tutorial
Dr. José Antonio Alvarado Rodríguez
Dr. Amilcar Meneses Viveros

Mineral de la Reforma, Hgo., México., agosto 2023

UNIVERSIDAD AUTÓNOMA DEL ESTADO DE HIDALGO
Instituto de Ciencias Básicas e Ingeniería
School of Engineering and Basic Sciences
Área Académica de Química
Department of Chemistry

Número de control: ICBI-AAQ/2276/2023
Asunto: Autorización de impresión

Mtra. Ojuky del Rocío Islas Maldonado
Directora de Administración Escolar
Presente.

El Comité Tutorial de la **Tesis de Maestría** titulada **"DETERMINACIÓN DE MODELOS DE TOXICIDAD EMPLEANDO DFT E INTELIGENCIA ARTIFICIAL EN COMPUESTOS ORGANOTIOFOSFORADOS"**, realizado por el sustentante **Uriel Josafat Rangel Peña** con **número de cuenta 351708** perteneciente al programa de **Maestría en Química**, una vez que ha revisado, analizado y evaluado el documento recepcional de acuerdo a lo estipulado en el Artículo 110 del Reglamento de Estudios de Posgrado, tiene a bien extender la presente:

AUTORIZACIÓN DE IMPRESIÓN

Por lo que el sustentante deberá cumplir los requisitos del Reglamento de Estudios de Posgrado y con lo establecido en el proceso de grado vigente.

Atentamente
"Amor, Orden y Progreso"
Mineral de la Reforma, Hidalgo a 31 de agosto de 2023

El Comité Tutorial

Dr. Julián Cruz Borbolla
Director de Tesis

Dra. Rosa Luz Camacho Mendoza

Dr. José Antonio Rodríguez Ávila

Dr. Amilcar Meneses Viveros

Ciudad del Conocimiento
Carretera Pachuca-Tulancingo km 4.5 Colonia
Carboneras, Mineral de la Reforma, Hidalgo,
México. C.P. 42184
Teléfono: +52 (771) 71 720 00 ext. 2200, 2201
Fax 6502
aaq_icbi@uaeh.edu.mx
www.uaeh.edu.mx

Agradecimientos

Al Consejo Nacional de Humanidades Ciencias y Tecnologías por el apoyo otorgado de la beca de maestría número 800093, así como por el proyecto número 1561802,

Resumen

El presente trabajo se utiliza descriptores de reactividad derivados de la teoría de los funcionales de la densidad conceptual y la teoría cuántica de átomos en moléculas para identificar zonas de interés en compuestos organotiofosforados. Además, muestra propiedades específicas correlacionadas con la toxicidad de estas moléculas, ofreciendo información sobre su reactividad. Estos enfoques, junto con el empleo de inteligencia artificial, posibilitaron la selección de variables significativas mediante un análisis minucioso de las propiedades químicas y estructurales que influyen en la toxicidad. En conjunto, estos resultados amplían la comprensión de los compuestos examinados y establecen los cimientos para estrategias más eficaces en el diseño de compuestos menos tóxicos y más seguros.

Se identificaron variables clave, en modelos lineales de regresión y clasificación, relacionadas con la toxicidad de los compuestos. En el modelo de regresión, ciertas variables mostraron una correlación positiva con la disminución de la toxicidad, mientras que otras presentaron una relación negativa. En el modelo de clasificación, diversas variables emergieron como relevantes, subrayando la importancia de la distribución de carga eléctrica y la orientación de los átomos en la reactividad y toxicidad. Asimismo, se identificaron regiones de gran importancia, como enlaces y átomos de oxígeno y fósforo.

Igualmente, se desarrollaron modelos no lineales de regresión basados en descriptores cuánticos. Todos los modelos presentan un rendimiento estadístico aceptable para predecir la toxicidad de los compuestos organotiofosforados. En resumen, este estudio proporciona información esencial sobre la relación entre descriptores de reactividad y toxicidad, marcando un paso significativo hacia la formulación de compuestos menos tóxicos en el futuro.

ÍNDICE

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

1 ANTECEDENTES

1.1 PESTICIDAS

La productividad de los cultivos se ve afectada por una variedad de organismos, tales como hongos, insectos, animales e incluso otras plantas, que son conocidas comúnmente como plagas. Estos organismos causan pérdidas económicas y problemas en las cadenas de suministro de alimentos [1,2]. Para hacer frente a esta problemática, se utilizan los pesticidas, también conocidos como plaguicidas, los cuales son productos químicos persistentes y poseen el potencial para controlar enfermedades, eliminar, inhibir y suprimir plagas. Estos pesticidas se aplican en jardines, áreas agrícolas y zonas urbanas con el fin de prevenir y reducir los daños, buscando minimizar las pérdidas y mejorar la calidad de los cultivos [3].

El uso extensivo de pesticidas ayuda a mejorar la producción de alimentos y solventar crisis alimentarias, permitiendo una disminución de la población que padece hambre. Sin embargo, el uso prolongado de pesticidas también trae afecciones a la salud de la población y organismos en los ecosistemas [4]. Estos se consideran de interés ambiental por su potencial toxicidad y sus propiedades para interactuar con el medio ambiente [5]. La facilidad de acceso hace que los pesticidas sean la forma más común de envenenamiento en muchas regiones rurales, asociándose a una tasa alta de mortalidad [6], por lo que sabiendo que son perjudiciales para la salud, su uso ha disminuido o eliminado por completo, y se ha optado por buscar nuevas alternativas [7].

El término “pesticida” engloba una amplia gama de productos químicos utilizados tanto en entornos domésticos como industriales, con propiedades químicas y físicas que varían de una clase a otra [8]. La Organización Mundial de la Salud (OMS) ha propuesto clasificarlos según su toxicidad y su estado de uso, ya sea permitido o

prohibido [9]. Actualmente, los métodos de clasificación más empleados se muestran en la Tabla 1 [8,10].

Tabla 1 Métodos de clasificación de pesticidas.

a) Por su estructura química		
Organofosforados	Piretroides	Neonicotinoides
Organoclorados	Azinas	
Carbamatos	Azoles	
b) Vía de ingreso		
Ingestión de toxinas	Exposición por contacto	Expectorantes
Evaporadores		
c) Acción del plaguicida y organismos que mata		
Acaricidas	Alguicidas	Antiincrustantes
Antimicrobianos	Atrayentes	Biocidas
Biopesticidas	Defoliantes	Desecantes
Desinfectantes	Esterilizantes de suelo	Feromonas
Fumigantes	Fungicidas	Herbicidas
Insecticidas	Molusquicidas	Nematicidas
Ovicidas	Plaguicidas microbianos	Preservantes de madera
Reguladores de crecimiento	Rodenticidas	Slimcidas

1.2 Efectos De Los Pesticidas A La Salud

Los pesticidas pueden ingresar al organismo por vía oral, nasal o por absorción dérmica. El compuesto viaja por el torrente sanguíneo para posteriormente ser metabolizado y excretado en la orina o el sudor e ingresado en tejidos adiposos en su misma forma o ya metabolizado [11]. Se sabe, por ejemplo, que estos compuestos pueden acumularse en el bulbo piloso del cabello e ir abandonando el cuerpo a medida que este crece, pudiendo quedar residuos durante décadas como trazas [10].

Al ser compuestos lipofílicos se transportan con los lípidos por todo el cuerpo, resultando en alteraciones metabólicas, causando, entre otros padecimientos: obesidad, dislipidemia, resistencia a la insulina y alteración de la función tiroidea [12,13]. La exposición prenatal y postnatal a pesticidas se relaciona con el desarrollo de padecimientos en la edad adulta, encontrándose presencia de estos compuestos en muestras de leche materna y suero de cordón umbilical, pudiendo los niños expuestos desarrollar defectos después del nacimiento o durante la gestación [14].

Se ha encontrado evidencia que bajas concentraciones de organotiofosforados inhiben la acetilcolinesterasa, causando estrés oxidativo que altera el metabolismo de macronutrientes, observándose malformaciones en ratones e importantes daños neuronales [15,16]. En la exposición a pesticidas se ha visto que son capaces de cambiar la arquitectura del tejido mamario y el equilibrio hormonal en ratas, promoviendo el desarrollo de tumores [17].

En algunos cultivos, por ejemplo, en las manzanas, la segregación de una capa cerosa como medio de protección natural, resulta ser una matriz donde los pesticidas pueden adherirse y así ser consumidos [18]. Por otra parte, en los climas cálidos y secos facilitan la bioacumulación de pesticidas en las plantas, debido a que aumenta la tasa de evaporación [19]. De manera general los síntomas de envenenamiento por organofosforados incluyen: pérdida de consciencia, respiración entrecortada, sudoración excesiva, echar espuma por la boca, resuello, lacrimación excesiva, contracciones musculares, convulsiones, diarrea, vómito, dolor de pecho y calambres abdominales [20].

1.3 ORGANOTIOFOSFORADOS

Los compuestos organofosforados (Figura 1) son ampliamente utilizados en la agricultura como insecticidas y acaricidas, siendo inhibidores de la serina esterasa y proteasa [21]. Son ésteres o tioles derivados del ácido fosfórico, fosfónico, fosfínico o

fosforomídico [22]. La actividad de esta clase de pesticidas y su toxicidad para los organismos no objetivo, se atribuye a la capacidad que tiene para interactuar e inactivar la acetilcolinesterasa, siendo incluso capaces de traspasar la barrera placentaria [14,23]. Esta proteína se ha estudiado ampliamente, siendo blanco de inhibidores como tratamiento para el Alzheimer [24].

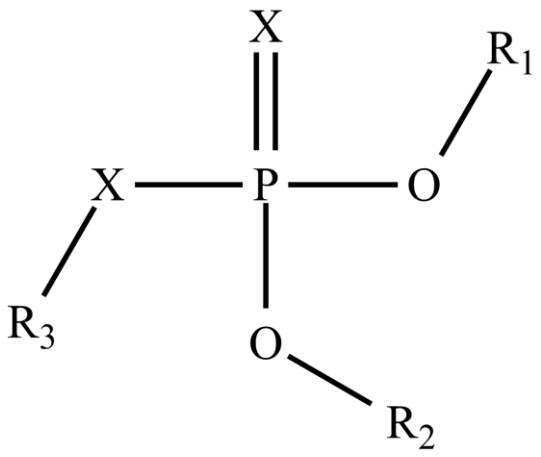

Figura 1 Estructura general de los pesticidas organofosforados (X =S, O y R_1, R_2, R_3 = alquilo o arilo)

La forma P=S es ampliamente manufacturada en comparación con la P=O, debido a que esta última tiene un mayor potencial para interactuar con la acetilcolinesterasa. En el cuerpo, los residuos P=S de los organotiofosfatos son oxidados por el citocromo P450 a su correspondiente forma P=O, esta biotransformación hace que se les considere más seguros [25]. Sin embargo, esta idea es errónea, lo que se puede ver en el caso del paratión (Figura 2), un organotiofosforado, catalogado como extremadamente peligroso que causó en 1986 el envenenamiento de 49 personas en Sierra Leona, de las cuales 14 fallecieron debido a la contaminación de un saco de harina con aproximadamente 10-15 ml del compuesto, se estimó que la ingesta de 5 g de este pan contaminado fueron suficientes para alcanzar la dosis letal en niños [20].

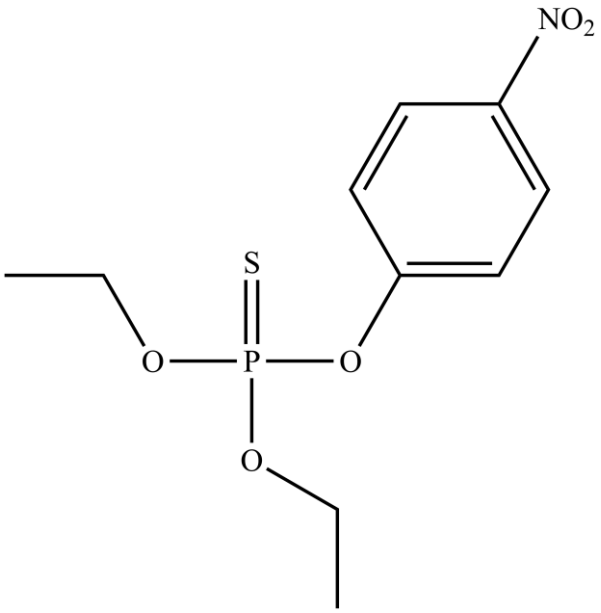

Figura 2 Estructura del paratión.

1.4 MÉTODOS COMPUTACIONALES APLICADOS A PESTICIDAS

Los métodos de química informática y modelado molecular han sido muy utilizados desde hace varias décadas para la selección y optimización de nuevos compuestos que pueden revertir los efectos adversos a causa de la contaminación en diversos organismos [26]. Su aplicación en toxicología predictiva es más reciente, surgiendo como consecuencia de los nuevos requerimientos regulatorios impuestos por las agencias internacionales. La toxicología computacional es actualmente utilizada como 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 las sustancias químicas en la saludad humana y/o en el medio ambiente. La principal ventaja de estos métodos es su capacidad para predecir parámetros estructurales, propiedades electrónicas y, por lo tanto, reducir el tiempo experimental [3,6,27].

Entre los métodos computacionales más efectivos para analizar las propiedades de los pesticidas se emplean los modelos matemáticos de predicción Relación

Cuantitativa Estructura Actividad y Relación Cuantitativa Estructura Propiedad (QSAR y QSPR, por sus siglas en inglés respectivamente), los cuales permiten relacionar las estructuras de los compuestos con su actividad (QSAR), propiedad (QSPR) y toxicidad (QSTR) [28]. Uno de los desafíos al desarrollar un modelo es la representación de la estructura molecular, la cual debe adoptar una forma legible por el software a emplear y retener la mayor cantidad de información estructural posible [6]. Los modelos computacionales tienen la capacidad de predecir las propiedades fisicoquímicas o biológicas de compuestos sin tener que llevar a cabo necesariamente su síntesis química en el laboratorio, es decir, se caracteriza por la aplicabilidad de los modelos resultantes de forma fácil e inmediata a nuevas estructuras. El uso de los enfoques *in silico* representa un ahorro significativo de tiempo, recursos y dinero, que pueden dar la ventaja competitiva a la industria que lo aplique.

1.5 INTELIGENCIA ARTIFICIAL

El empleo de métodos computacionales cumple con los principios de reemplazo, refinamiento y reducción de las pruebas en animales, mejor conocido como las 3 R's. Estas metodologías están basadas en el concepto de que los compuestos comparten similitudes estructurales o que ciertas subestructuras poseen una mayor probabilidad de compartir propiedades toxicológicas [29].

Los modelos QSAR más recientes buscan emplear nuevas herramientas y metodologías a fin de mejorar la capacidad de predicción, entre estas herramientas se encuentra el uso de métodos de inteligencia artificial (IA) como el aprendizaje automático (*machine learning*, ML) y el aprendizaje profundo (*deep learning*). La IA es una rama de la informática que puede ser definida como aquella que tiene el objetivo de crear sistemas o métodos que analicen la información y permitan el manejo de la complejidad de un problema [30].

La IA usada en el desarrollo de fármacos surge de conceptos de ML y quimioinformática. Su uso se vio beneficiado por los avances en temas como el manejo de *big data*, aceleración mediante GPU (unidad de procesamiento gráfico, por sus siglas en inglés), computación en la nube y liberación de herramientas de IA [31]. ML emplea diversas técnicas estadísticas para permitir que las computadoras aprendan de datos químicos, biológicos o toxicológicos, sin intervención humana. Estos algoritmos son capaces de encontrar la compleja relación no lineal entre descriptores relevantes [29].

A grandes rasgos, la IA abarca métodos de aprendizaje, descubrimiento y razonamiento. En los métodos de aprendizaje el objetivo es que la computadora sea capaz de aprender automáticamente, a partir de conocimiento previo, sin asistencia. El descubrimiento y exploración se basa en la creación de algoritmos capaces de identificar información válida, entendible y relevante en bases de datos. Por último, la parte de razonamiento indica la generación de caminos precisos y efectivos para generar inferencias de manera robusta [30].

1.5.1 ALGORITMOS DE APRENDIZAJE

Dentro del ámbito del aprendizaje automático, existen diferentes tipos de algoritmos que se utilizan para hacer predicciones y tomar decisiones. Estos se pueden dividir en algoritmos de tipo "*black box*" y algoritmos explicables, conocidos de manera análoga como "*white box*". Se consideran algoritmos tipo *black box* aquellos en los que se genera una función demasiado compleja para que pueda ser comprendida por cualquier persona, o bien cuando están patentadas. En general, los modelos de aprendizaje profundo (*deep learning*) tienden a ser de este tipo, ya que son altamente recursivos. Por otro lado, los algoritmos tipo *white box* son fácilmente interpretables [32,33]. Es importante aclarar que el concepto de "explicable" o "interpretable" en este contexto se refiere a la capacidad de entender las interacciones y comportamientos de

las variables dentro del modelo, y no de cómo estas variables se interpretan por sí solas en el mundo real.

1.5.1.1 RANDOM FOREST

Random Forest (RF) es un algoritmo de aprendizaje automático utilizado para la clasificación y la regresión. Es un método de conjunto que combina múltiples árboles de decisión para producir un modelo predictivo robusto y preciso. Fue desarrollado por Leo Breiman en 2001 y se ha convertido en uno de los algoritmos más populares en el aprendizaje automático, particularmente para problemas de clasificación [34].

Este algoritmo está basado en árboles de decisión, sin embargo, difiere de los árboles de decisión individuales en que utiliza múltiples árboles en lugar de uno solo, dando lugar a bosques. Los árboles que conforman el bosque se construyen utilizando diferentes subconjuntos de datos de entrenamiento y variables de entrada. Al combinar predicciones de cada árbol individual se logra que el algoritmo sea menos propenso al sobreajuste y mejora la capacidad de generalizar a nuevos datos. Lo anterior se da debido a que, aunque cada árbol individual puede no ser muy preciso, en conjunto, los árboles pueden generar una predicción más robusta y precisa [35].

RF se basa en dos principios: el ensamblaje de los árboles de decisión y la selección aleatoria de las características. En el proceso de entrenamiento se construyen varios árboles independientes, cada uno entrenado con un subconjunto aleatorio de datos de entrenamiento y un subconjunto aleatorio de características. En la etapa de predicción, cada árbol individual realiza su propia predicción basada en el subconjunto de datos y características con el que fue entrenado. La predicción final se calcula como la media de las predicciones de todos los árboles individuales para la clasificación, y como la media ponderada para la regresión. Este proceso promediado reduce el riesgo de sobreajuste y aumenta la precisión y robustez de las predicciones [34–36].

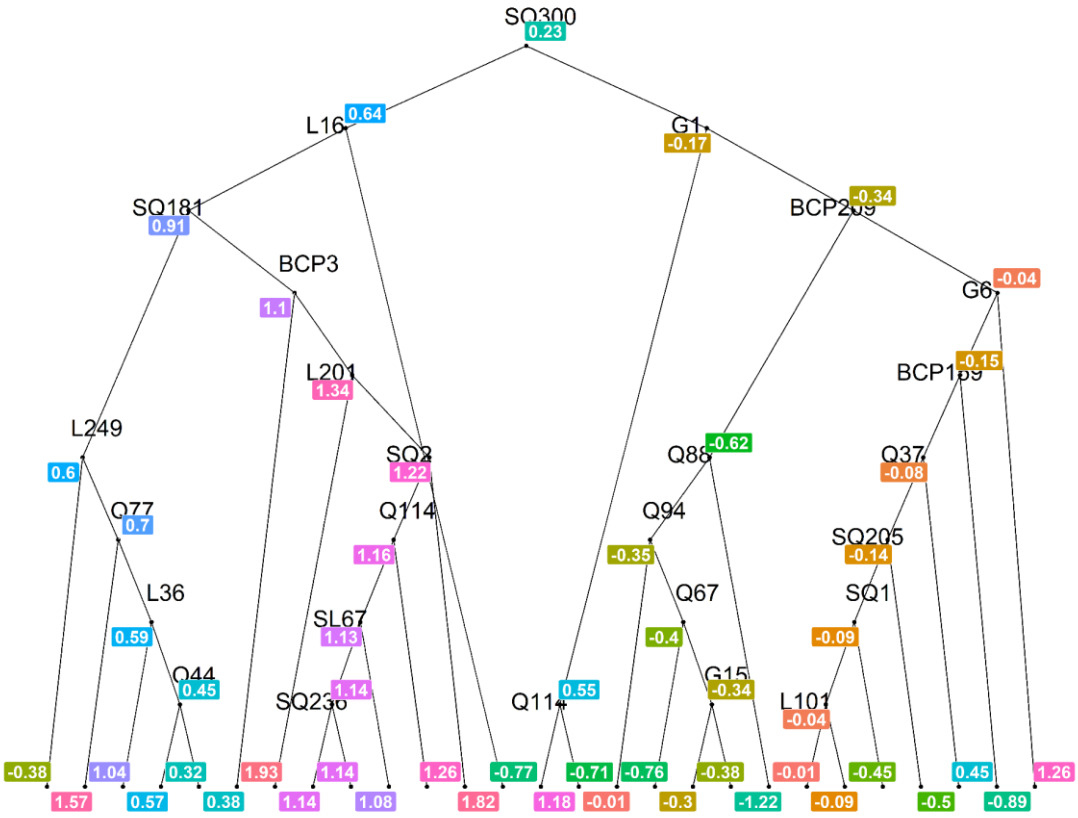

Figura 3 Representación gráfica de un árbol dentro de un modelo de Random Forest.

En la Figura 3 se muestra un ejemplo de un árbol de decisión individual extraído de un modelo RF. Se puede observar cómo el modelo utiliza diferentes variables para tomar decisiones y a su vez cómo carece de interpretabilidad, lo que dificulta la comprensión de cómo el modelo llega a sus predicciones. Las ventajas que, de manera general, este algoritmo presenta se muestran a continuación:

- Puede manejar conjuntos de datos grandes con una gran cantidad de variables de entrada.
- Es robusto frente a valores atípicos y datos faltantes.
- Proporciona una medida de la importancia relativa de las variables en la predicción.

- Es menos propenso al sobreajuste en comparación con otros métodos de aprendizaje automático.
- Puede manejar diferentes tipos de variables, como variables categóricas y numéricas.

Sin embargo, entre sus desventajas se encuentran:

- Puede ser computacionalmente costoso, especialmente con un gran número de árboles de entrada.
- La configuración de hiperparámetros puede requerir ajustes para tener un rendimiento óptimo.
- No es adecuado para datos de series de tiempo o datos con una estructura temporal.

1.5.1.2 LASSO

LASSO, que significa "Operador de Selección y Contracción Mínima Absoluta" (por su significado en inglés: "*Least Absolute Shrinkage and Selection Operator*"), es un algoritmo de aprendizaje automático utilizado para la selección de características y la regularización de modelos de regresión. Fue introducido por Robert Tibshirani en 1996, siendo un método para manejar problemas de alta dimensionalidad en los datos [37].

El objetivo de LASSO es reducir la complejidad de un modelo al reducir el número de características incluidas en el modelo y al reducir los valores de los coeficientes de las características no importantes. Esto se logra a través de una penalización del valor absoluto de los coeficientes de las características en la función de costo del modelo. De esta manera, LASSO fuerza a algunos de los coeficientes de las características a tener el valor de cero, lo que resulta en una selección de características eficiente y en la construcción de modelos más simple [38].

Este algoritmo funciona al agregar la penalización L1 a la función de costo de la regresión lineal ordinaria. La función de costo se define como la suma de los errores cuadráticos entre las predicciones y las respuestas verdaderas más la suma de los valores absolutos de los coeficientes. El parámetro de regularización lambda (λ), que se ajusta a través de validación cruzada, controla el peso de la penalización L1. En la Figura 4 se observa cómo a medida que se aumenta el valor de lambda, más coeficientes se vuelven cero y menos características se incluyen en el modelo final [37,38].

LASSO tiene varias ventajas sobre otros algoritmos de aprendizaje automático, algunas de las más destacables son:

- Es efectivo para seleccionar las características más importantes y descartar las que no lo son, lo que resulta en modelos más simples e interpretables.
- Puede ayudar a prevenir el sobreajuste de los datos de entrenamiento al regularizar los coeficientes de las características.
- Es robusto frente a la colinealidad en los datos.
- Es fácil de implementar y de ajustar, lo que lo hace ideal para su uso en aplicaciones prácticas.
- Limita la cantidad máxima de variables en el modelo a 1/3 de los datos observados.

Sin embargo, también tiene algunas desventajas que considerar:

- La selección de un valor adecuado de lambda es crítica para obtener un modelo preciso y útil. Un valor de lambda demasiado bajo puede resultar en un modelo con demasiadas características irrelevantes.
- En presencia de datos altamente correlacionados tiende a seleccionar sólo una de las características correlacionadas y descartar las demás. Esto puede resultar en un modelo subóptimo.
- Se basa en una función de regresión lineal y no es adecuado para datos no lineales

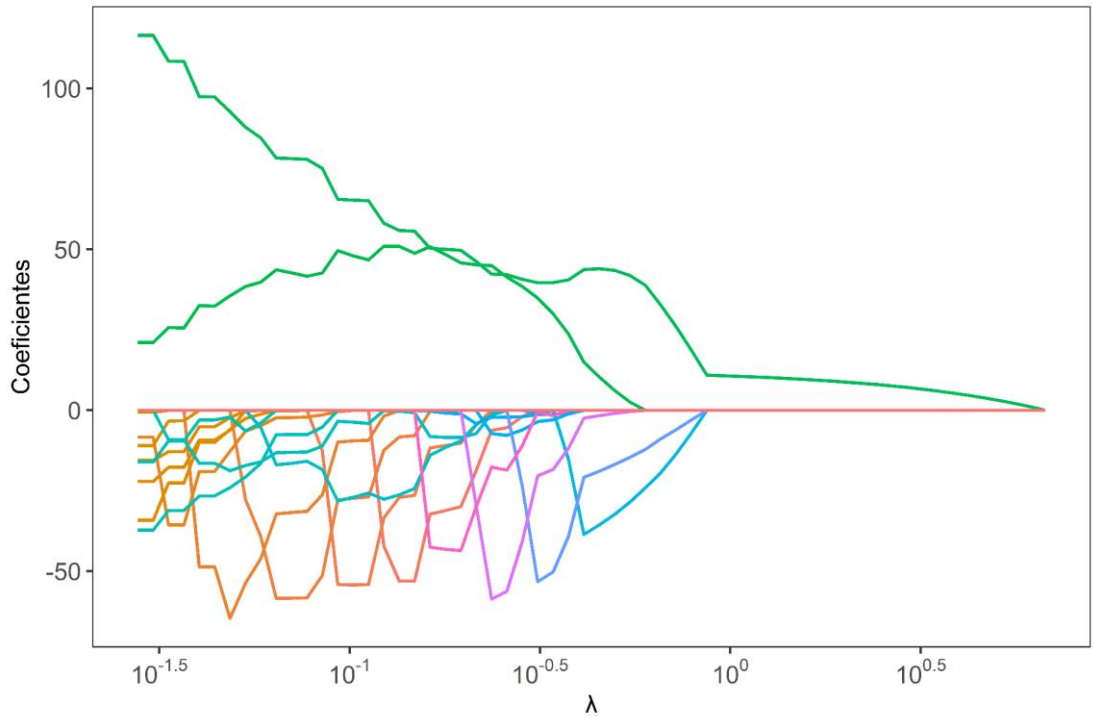

Figura 4 Cambio en los coeficientes de un modelo LASSO a medida que aumenta λ.

1.5.1.3 RIDGE

RIDGE es un algoritmo de aprendizaje automático introducido por Hoerl y Kennard en 1970, utilizado para la selección de características y la regularización de modelos de regresión. Es similar a LASSO, pero utiliza una penalización diferente para los coeficientes de las características. El objetivo de este algoritmo es reducir la complejidad de un modelo al reducir el tamaño de los coeficientes de las características no importantes. Esto se logra a través de la penalización del cuadrado de los coeficientes de las características en la función de costo del modelo. De esta manera RIDGE reduce el valor de los coeficientes sin forzarlos a ser cero [39,40].

El algoritmo funciona al agregar una penalización L2 a la función de costo de la regresión lineal ordinaria. Al igual que con LASSO, el parámetro de regularización lambda se ajusta a través de validación cruzada. En la Figura 5 se muestra que, a medida

que se aumenta el valor de lambda los valores de los coeficientes se reducen, pero nunca se hacen cero [38,39].

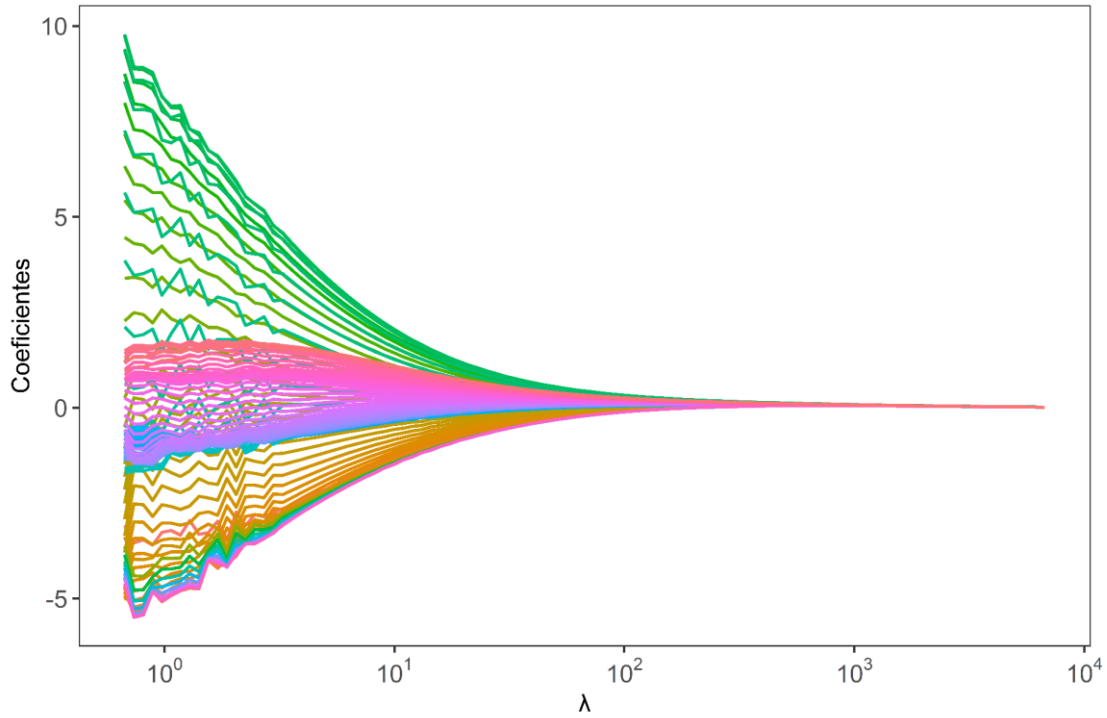

Figura 5 Cambio en los coeficientes de un modelo RIDGE a medida que aumenta λ.

RIDGE tiene varias ventajas sobre otros algoritmos de aprendizaje automático, algunas de las más destacables son:

- Es muy efectivo para prevenir el sobreajuste de los datos de entrenamiento al regularizar los coeficientes de las características.
- Es robusto ante la colinealidad de los datos.
- Es fácil de implementar y ajustar.
- Puede manejar conjuntos de datos con muchas características.

Por otro lado, sus desventajas son:

- No es tan efectivo como LASSO para la selección de características más importantes y el descarte de las menos importantes.

- No fuerza a los coeficientes a ser cero, lo que puede resultar en la inclusión de características irrelevantes en el modelo final.
- La elección de un valor adecuado de lambda es crítica para obtener un modelo útil.

1.5.1.4 ELASTIC NET

Elastic Net (EN) es un algoritmo de aprendizaje automático que combina la regularización L1 de LASSO y la L2 de RIDGE para seleccionar características y reducir el sobreajuste en modelos de regresión. La idea detrás de Elastic Net es tomar lo mejor de ambos mundos: la capacidad de LASSO para seleccionar características importantes y la capacidad de RIDGE para reducir el sobreajuste en un conjunto de datos con alta dimensionalidad. Este algoritmo agrega una penalización L1 y L2 a la función de costo de la regresión ordinaria, donde los parámetros de regularización empleados son α y λ. El parámetro α controla el equilibrio entre la penalización L1 y L2, si es 0 se reduce a una regresión lineal ordinaria; si es 1 se reduce a LASSO y si está entre 0 y 1 se combinan ambas penalizaciones. Por otro lado, λ controla la fuerza de las penalizaciones y se ajusta a través de validación cruzada [41]. En la Figura 6 se ejemplifica un modelo EN para un valor de $\alpha = 0.5$, lo que da el mismo peso a ambas penalizaciones, se observa cómo algunos de los coeficientes se eliminan, mientras que otros tienen a cero.

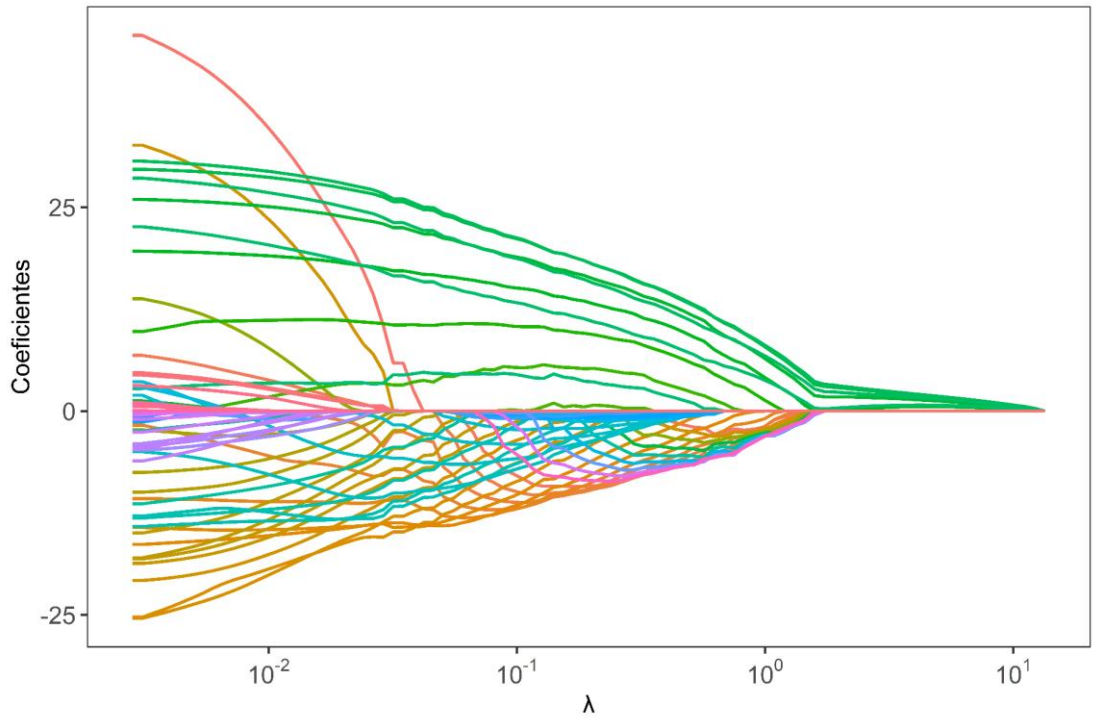

Figura 6 Cambio en los coeficientes de un modelo EN a medida que aumenta λ, para un valor de α = 0.5.

1.5.1.5 SUPPORT VECTOR MACHINES

Support Vector Machines (SVM) es un algoritmo que se basa en encontrar un hiperplano en un espacio de alta dimensionalidad que maximice la separación entre diferentes clases de datos o, en el caso de la regresión, que minimice el error entre las predicciones y los valores reales [42].

El objetivo de Support Vectors Machines es encontrar una función lineal que modele la relación entre las características de entrada y la variable de salida. En lugar de minimizar el error cuadrático medio (MSE), como se hace en la regresión ordinaria, SVM minimiza una función de costo que tiene en cuenta la distancia entre los puntos de los datos y el hiperplano de separación. Los puntos de los datos que están más cerca del hiperplano tienen un mayor peso en la función de costo, a estos puntos se les llama vectores de soporte, de ahí el nombre del algoritmo [43].

Las ventajas más destacables de SVM son:

- Puede manejar datos no lineales empleando una técnica llamada kernel trick, que mapea los datos a un espacio de alta dimensión donde pueden ser linealmente separables.
- Es resistente al sobreajuste gracias a la función de costo que tiene en cuenta la distancia entre los puntos de los datos y el hiperplano de separación.
- Generaliza bien en conjuntos de datos no vistos gracias a la penalización por error de clasificación que se aplica a los puntos de datos que están más cerca del hiperplano de separación.

Por otro lado, las desventajas a considerar son:

- Puede depender en gran medida de la elección adecuada de los parámetros, como el parámetro de regularización y el tipo de kernel.
- SVM puede ser sensible a datos ruidosos o valores atípicos en el conjunto de datos.
- Puede volverse computacionalmente costoso y requerir una gran cantidad de memoria cuando se trabaja con conjuntos de datos muy grandes. Esto se debe a que el tiempo de entrenamiento y la complejidad del algoritmo aumentan considerablemente con el tamaño del conjunto de datos.

1.6 cDFT

La Teoría de los Funcionales de la Densidad conceptual (cDFT, por sus siglas en inglés) es una extensión de la Teoría de los Funcionales de la Densidad en donde los conceptos físicos y matemáticos de DFT son reutilizados y combinados para crear descriptores que permitan elucidar conceptos químicos, en especial aquellos enfocados a la reactividad, mediante el enfoque en el análisis de las propiedades electrónicas de los sistemas moleculares. Esta teoría se basa en el principio del funcional de la densidad

de Hohenberg-Kohn, que establece que la densidad electrónica de un sistema es suficiente para describir completamente su estado físico y químico [44].

Esta herramienta es útil para predecir la reactividad de los compuestos ante diferentes reacciones química, lo que permite diseñar nuevos compuestos con propiedades específicas y mejorar los procesos de síntesis química. Además, es capaz de proporcionar información detallada sobre los mecanismos de reacción, lo que permite una comprensión más profunda de los procesos químicos a nivel molecular. Entre las ventajas de su uso se encuentran su capacidad para predecir con precisión la reactividad química de los compuestos y su flexibilidad para adaptarse a diferentes sistemas moleculares. Sin embargo, una de las principales desventajas de esta teoría es su complejidad matemática y que a menudo requiere la utilización de programas de cálculo complejos y costosos [44–46].

1.7 QTAIM

La Teoría Cuántica de Átomos en Moléculas (QTAIM, por sus siglas en inglés) es un modelo teórico para la descripción de la estructura y reactividad de las moléculas basado en la topología de la densidad electrónica. Su enfoque se centra en los átomos individuales en una molécula y cómo se conectan mediante los enlaces químicos. Esta teoría sostiene que los enlaces químicos no son líneas imaginarias entre los átomos, sino que están definidos por las regiones donde la densidad electrónica es alta y donde hay una energía potencial mínima entre dos átomos [47,48].

QTAIM permite la descripción de la reactividad de las moléculas mediante la identificación de los puntos críticos de la densidad electrónica, como los núcleos atómicos y los puntos de silla en una superficie de potencial electrostático. Estos puntos críticos pueden ser utilizados para calcular propiedades moleculares. Además, con estas propiedades se puede predecir la estabilidad relativa de diferentes moléculas y proporcionar información detallada sobre los procesos de reacción. Una de las ventajas

de emplear QTAIM es que proporciona una descripción más detallada de la densidad electrónica y la topología de los enlaces químicos que los métodos tradicionales de química cuántica, como la teoría de los funcionales de la densidad. Además de ser aplicable a una amplia gama de sistemas químicos, incluyendo moléculas orgánicas e inorgánicas en una variedad de condiciones. Sin embargo, una de sus desventajas es que su aplicación a sistemas moleculares grandes y complejos puede ser computacionalmente costosa y requiere una gran cantidad de tiempo de cálculo. A su vez, la interpretación de los resultados puede ser difícil al ser muy técnicos [48].

1.8 Descriptores de Reactividad

La búsqueda de formas en las que la reactividad de una especie pueda ser inferida a partir de cálculos teóricos se ha convertido en uno de los temas más relevantes dentro de la química computacional. La reactividad se puede estimar mediante el empleo de parámetros, que son ecuaciones que describen alguna propiedad, ya sea local o global, y que resultan útiles para describir la reactividad global y la selectividad local, respectivamente [49].

1.8.1 Descriptores Globales

GAP:

Es la diferencia energética entre el orbital molecular más bajo desocupado (LUMO) y el orbital molecular más alto ocupado (HOMO):

$$GAP = E_{LUMO} - E_{HOMO} \quad (1)$$

Potencial de ionización:

Energía necesaria para remover un electrón de un átomo en estado gaseoso [50].

$$PI = E_C^{N-1} - E_0^N \quad (2)$$

Donde:

E_C^{N-1} : energía del sistema al perder un electrón.

E_0^N : energía del sistema en estado basal.

AFINIDAD ELECTRÓNICA:

Capacidad de un átomo para atraer los electrones [51]:

$$AE = E_0^N - E_C^{N+1} \quad (3)$$

ELECTRONEGATIVIDAD:

Capacidad de un átomo para atraer los electrones de un enlace químico [52]:

$$\chi = \frac{PI + AE}{2} \quad (4)$$

POTENCIAL QUÍMICO:

Variación de la energía respecto al número de electrones a un potencial externo constante [53]:

$$\mu = -\frac{PI + AE}{2} = -\chi \quad (5)$$

DUREZA:

Resistencia de una especie a cambiar su configuración electrónica [54]:

$$\eta = PI - AE \quad (6)$$

BLANDURA:

Inverso de la dureza [52,54]:

$$S = \frac{1}{\eta} \tag{7}$$

ÍNDICE DE ELECTROFILICIDAD:

Capacidad de aceptación máxima de un flujo de electrones [54,55]:

$$\omega = \frac{\mu^2}{2\eta} \tag{8}$$

ÍNDICE DE NUCLEOFILICIDAD:

Capacidad de donación máxima de un flujo de electrones [54,56,57]:

$$N_{Nu} = E_{HOMO_{Nu}} - E_{HOMO_{TCE}} \tag{9}$$

Donde:

$E_{HOMO_{Nu}}$: energía del HOMO del sistema.

$E_{HOMo_{TCE}}$: energía del HOMO del tetracianoetileno.

PODER ELECTRODONADOR:

Tendencia a donar carga [55,58]:

$$\omega^- = \frac{(\mu^-)^2}{\eta} \tag{10}$$

Donde:

$$\mu^- = -\frac{3PI + AE}{4} \tag{11}$$

PODER ELECTROATRACTOR:

Tendencia a aceptar carga [55,58]:

$$\omega^+ = \frac{(\mu^+)^2}{\eta} \tag{12}$$

Donde:

$$\mu^+ = -\frac{PI + 3AE}{4} \tag{13}$$

1.8.2 DESCRIPTORES LOCALES

FUNCIONES DE FUKUI:

Se utilizan para caracterizar la capacidad de un átomo específico en una molécula para aceptar o donar electrones en una reacción química. En particular, se refieren a la capacidad de un átomo para actuar como un nucleófilo o un electrófilo en una reacción química. Estas ecuaciones se derivan del análisis de la densidad electrónica de la molécula y se utilizan para predecir la selectividad de las reacciones químicas, así como la regioselectividad y la quimioselectividad. Los valores de las funciones de Fukui se pueden calcular para cada átomo de una molécula, lo que permite una descripción detallada de la reactividad local de la molécula en su conjunto, además de analizar el comportamiento ácido-base de diferentes especies a nivel local, incluso si no comparten los mismos grupos funcionales [59]:

$$\mu_{k\tau}^- = -PIf_k^- \; para\; \omega < 0 \tag{14}$$

$$\mu_{k\tau}^+ = -AEf_k^+ \; para\; \omega > 0 \tag{15}$$

$$\mu_{k\tau}^o = -\frac{PIf_k^- + AEf_k^+}{2} \; para\; \omega = 0 \tag{16}$$

$$DD_k = f_k^+ - f_k^- \tag{17}$$

Donde:

q_k^N : carga de Mulliken para el *k* átomo para la especie con N electrones.

q_k^{N-1}: carga de Mulliken para el *k* átomo para la especie con N-1 electrones.

q_k^{N+1} : carga de Mulliken para el *k* átomo para la especie con N+1 electrones.

ELECTROFILICIDAD LOCAL:

Capacidad de aceptar carga del *k* átomo [54,60]:

$$\omega_k^{loc} = \omega f_k^+ \quad (18)$$

NUCLEOFILICIDAD LOCAL:

Capacidad de donar carga del *k* átomo [54,60]:

$$N_k^{loc} = N_{Nu} f_k^- \quad (19)$$

BLANDURA CONDENSADA

Capacidad del k-átomo para deformarse cuando se somete a una perturbación externa [54,60]:

$$S_k^+ = S f_k^+ \quad (20)$$

$$S_k^- = S f_k^- \quad (21)$$

$$S_k^o = S f_k^o \quad (22)$$

ELECTROFILICIDAD RELATIVA [61]:

$$\omega_k = \frac{S_k^+}{S_k^-} \quad (23)$$

NUCLEOFILICIDAD RELATIVA [54,61]:

$$N_k = \frac{S_k^-}{S_k^+} \tag{24}$$

POTENCIAL QUÍMICO LOCAL:

Analiza el comportamiento ácido-base de diferentes especies a nivel local, incluso si no comparten los mismos grupos funcionales [59]:

$$\mu_{k\tau}^- = -PIf_k^- \; para \; \omega < 0 \tag{25}$$

$$\mu_{k\tau}^+ = -AEf_k^+ \; para \; \omega > 0 \tag{26}$$

$$\mu_{k\tau}^o = -\frac{PIf_k^- + AEf_k^+}{2} \; para \; \omega = 0 \tag{27}$$

DUREZA LOCAL:

Permite comparar sitios equivalentes en ambientes químicos diferentes [59]:

$$\eta_\tau^- = \eta_k - \frac{\eta_k DD_k}{2} \; para \; \omega = -1 \tag{28}$$

$$\eta_\tau^+ = \eta_k + \frac{\eta_k DD_k}{2} \; para \; \omega = 1 \tag{29}$$

$$\eta_{k\tau}^o = \eta_k \; para \; \omega = 0 \tag{30}$$

Donde:

$$\eta_k = PIf_k^- - AEf_k^+ \tag{31}$$

FUKUI KERNEL:

Es empleado como un indicador de la reactividad/estabilidad de los enlaces [59]:

$$f_\tau^- = f_k^- f_{k'}^- \; para \; \omega < 0 \tag{32}$$

$$f_\tau^+ = f_k^+ f_{k'}^+ \; para \; \omega > 0 \tag{33}$$

$$f_\tau^o = \frac{f_k^- f_{k'}^- + f_k^+ f_{k'}^+}{2} \; para \; \omega = 0 \tag{34}$$

Donde:

k' : es el átomo vecino a k.

KERNEL DUAL:

Indica el equilibrio entre las características electrofílicas y nucleofílicas de un enlace [59]:

$$\Delta f_\tau^- = \Delta f_{kk'} - \frac{DD_k DD_{k'}}{2} \; para \; \omega = -1 \tag{35}$$

$$\Delta f_\tau^+ = \Delta f_{kk'} + \frac{DD_k DD_{k'}}{2} \; para \; \omega = 1 \tag{36}$$

$$\Delta f_\tau^o = \Delta f_{kk'} \; para \; \omega = 0 \tag{37}$$

Donde:

$$\Delta f_{kk'} = f_k^+ f_{k'}^+ - f_k^- f_{k'}^- \tag{38}$$

2 OBJETIVOS

2.1 OBJETIVO GENERAL

Generar modelos QSTR para compuestos organotiofosforados y obtener descriptores de reactividad mediante la aplicación de la Teoría de los Funcionales de la Densidad y técnicas de inteligencia artificial con el fin de identificar las variables que tienen el mayor impacto en la toxicidad y desarrollar un modelo predictivo que explique la reactividad de estos compuestos.

2.2 OBJETIVOS ESPECÍFICOS

- Investigar en acervos bibliográficos y revistas especializadas datos sobre toxicidad aguda LD_{50} y parámetros fisicoquímicos de compuestos químicos derivados de organotiofosforados para construir una base de datos relevante en este proyecto.
- Seleccionar y optimizar la geometría de los compuestos que contengan el grupo tiofosfato, utilizando el funcional ωB97XD y la base 6-311++G**, en presencia de solvente agua, para preparar adecuadamente los sistemas de estudio.
- Aplicar la teoría de los funcionales de la densidad (DFT) para determinar los parámetros estructurales y electrónicos de las moléculas optimizadas, para obtener información detallada sobre sus propiedades electrónicas, mediante el uso de la teoría de los funcionales de la densidad conceptual (cDFT) y el software AIMALL.
- Relacionar los descriptores globales, locales y de QTAIM con los valores de toxicidad mediante la obtención de diferentes modelos matemáticos (Relación Cuantitativa Estructura-Toxicidad) aplicando inteligencia artificial.

3 PROCEDIMIENTO

3.1 SELECCIÓN DE LA BASE DE DATOS

La creciente cantidad de información almacenada en bases de datos ha generado un desafío en la comunidad científica debido a la dificultad que representa su análisis. Actualmente, se estima que cada dos años la cantidad de datos disponibles se duplica, lo que ha llevado a la era del "*big data*". Sin embargo, el término "*big data*" no sólo se refiere a la cantidad de información a procesar, sino también a la creciente complejidad de los datos. En este contexto, resulta imprescindible el desarrollo de nuevas técnicas y herramientas que permitan analizar y extraer información valiosa de los datos, a fin de realizar avances significativos [62,63]. Con lo anterior, la aplicación de técnicas de inteligencia artificial y aprendizaje automático se presenta como una alternativa para el procesamiento y análisis de grandes volúmenes de datos complejos. De manera inicial, en este proyecto se decidió emplear el lenguaje de programación R4.2.1 [64] junto con la API de ChemIDplus [65,66] donde se realizó una búsqueda de compuestos organotiofosforados que presentaran la estructura mostrada en la Figura 7.

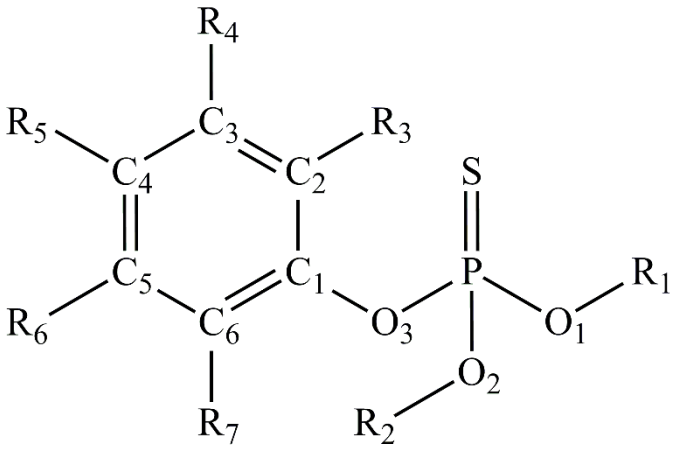

Figura 7 Estructura base de los organotiofosforados estudiados.

En este estudio se definió la subestructura de los organotiofosforados utilizando la línea de SMILES: "[P](=S)(OC)(OC)OC1=CC=CC=C1". Posteriormente, se empleó el paquete "jsonlite" para buscar en la base de datos de ChemIDPlus compuestos que contengan dicha subestructura y valores de toxicidad LD_{50} en ratas vía oral [67]. Se realizó un curado manual de las estructuras obtenidas, siguiendo las buenas prácticas en la generación de modelos [68]. En total, se obtuvieron 62 compuestos que cumplían con los criterios establecidos, los cuales se muestran en la Figura 8. La aplicación rigurosa de las buenas prácticas es crucial para garantizar la calidad y confiabilidad de los modelos generados, estas prácticas consideran los siguientes puntos:

- Eliminación de compuestos inorgánicos y mezclas.
- Conversión y limpieza estructural.
- Normalización de quimotipos específicos.
- Remoción de duplicados.
- Revisión manual final.

A continuación, para los 62 compuestos se obtuvieron los valores del coeficiente de partición octanol-agua empleando la calculadora de propiedades Molinspiration [69]. Con fundamento en el punto de vista regulatorio, los datos de toxicidad LD_{50} en las bases de datos se presentan en mg/kg como lo requiere el Registro, Evaluación, Autorización y Restricción de Químicos (REACH, por sus siglas en inglés) y la Clasificación, Etiquetado y Envasado (CLP, por sus siglas en inglés) [70]. Sin embargo, con el fin de proporcionar una correcta interpretación química, se optó por transformar los datos de toxicidad LD_{50} a mmol/kg, lo cual se muestra en la Tabla 2 junto con los IDs asignados y los identificadores CAS correspondientes. Esta transformación se consideró en los modelos de regresión generados en este estudio.

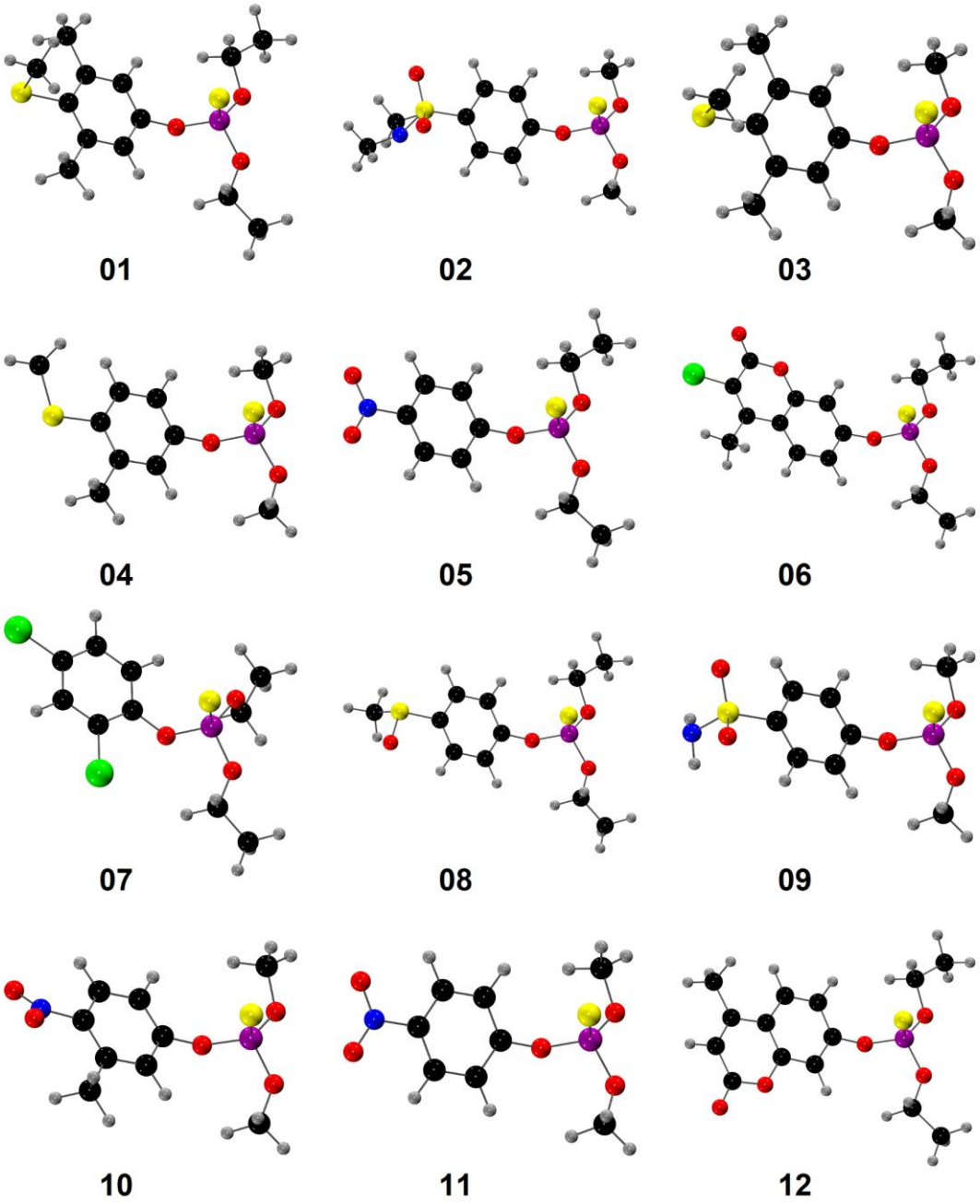

Figura 8 Estructuras en 3D de los compuestos organotiofosforados estudiados. El código de colores es: gris (H), negro (C), azul (N), amarillo (S), rojo (O), verde (Cl), cian (F), magenta (P), café (Br) y púrpura (I).

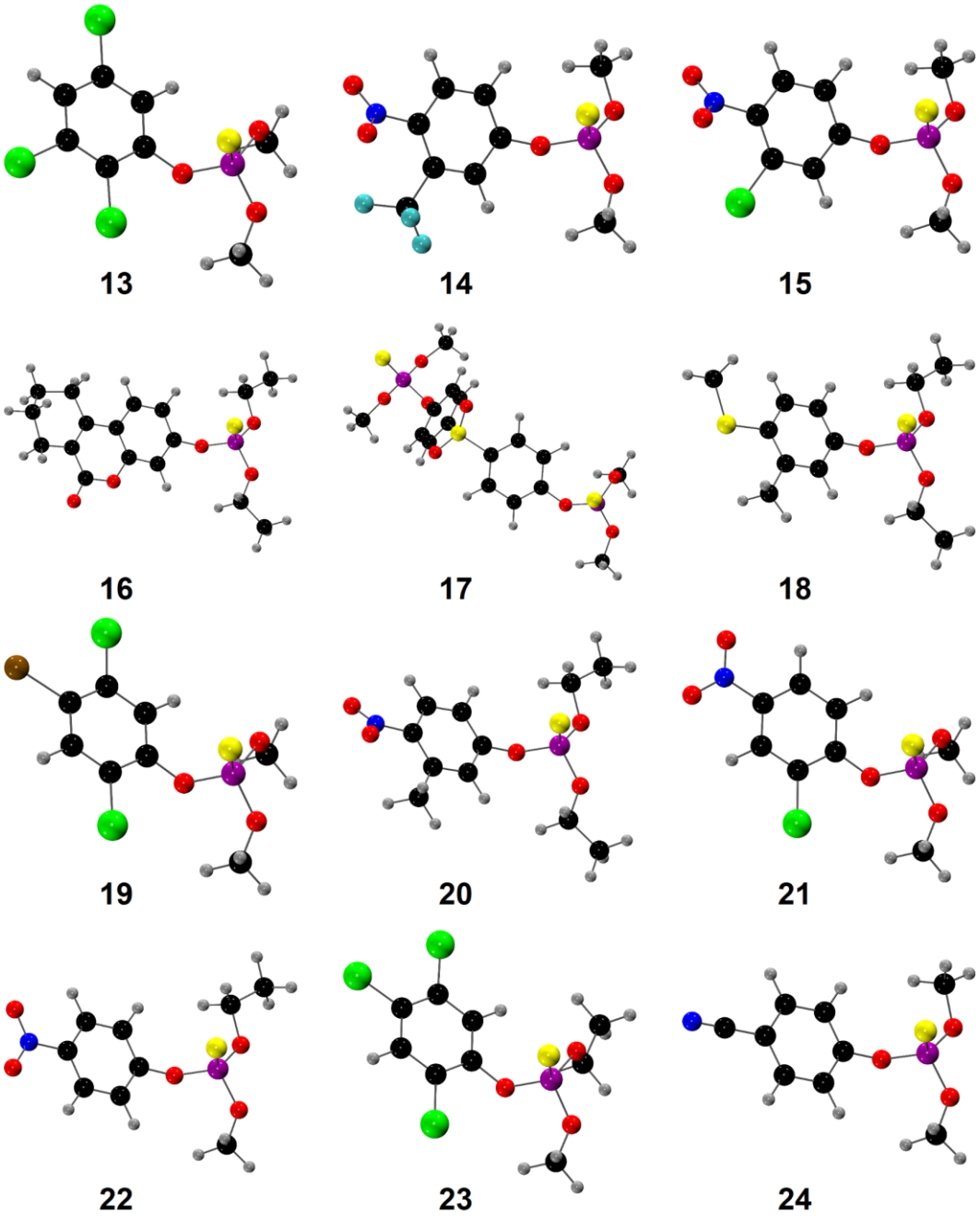

Figura 8 continuación.

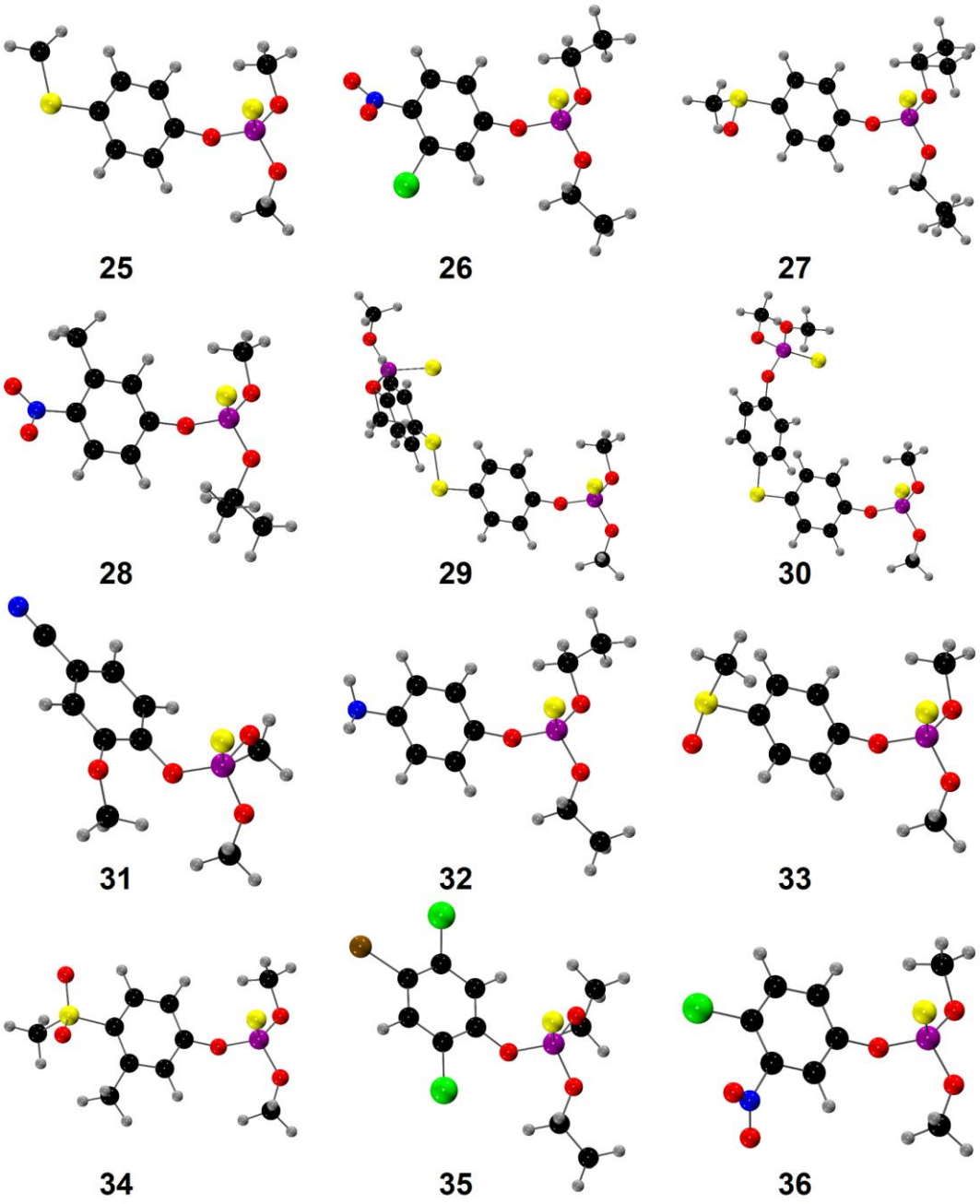

Figura 8 continuación.

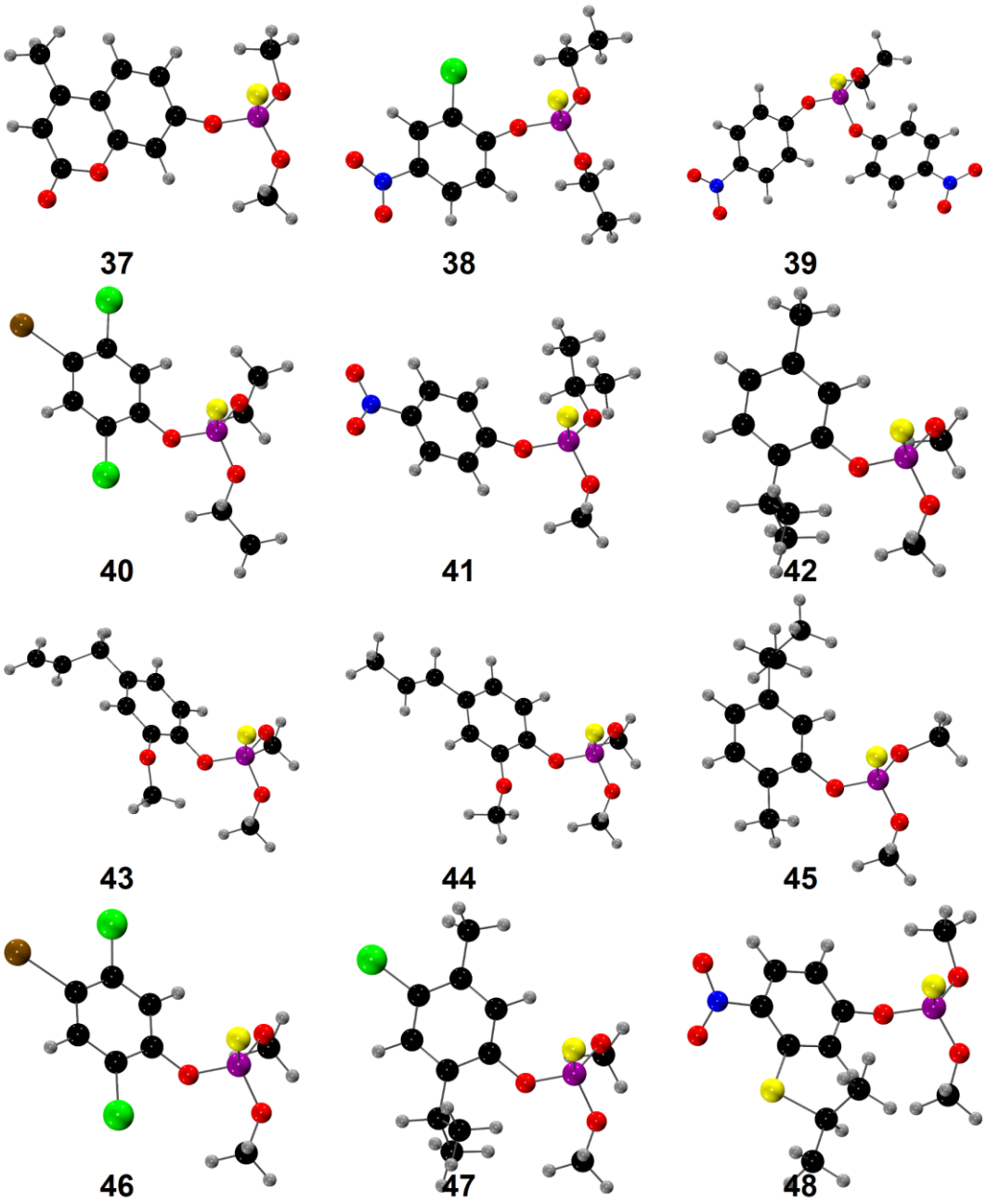

Figura 8 continuación.

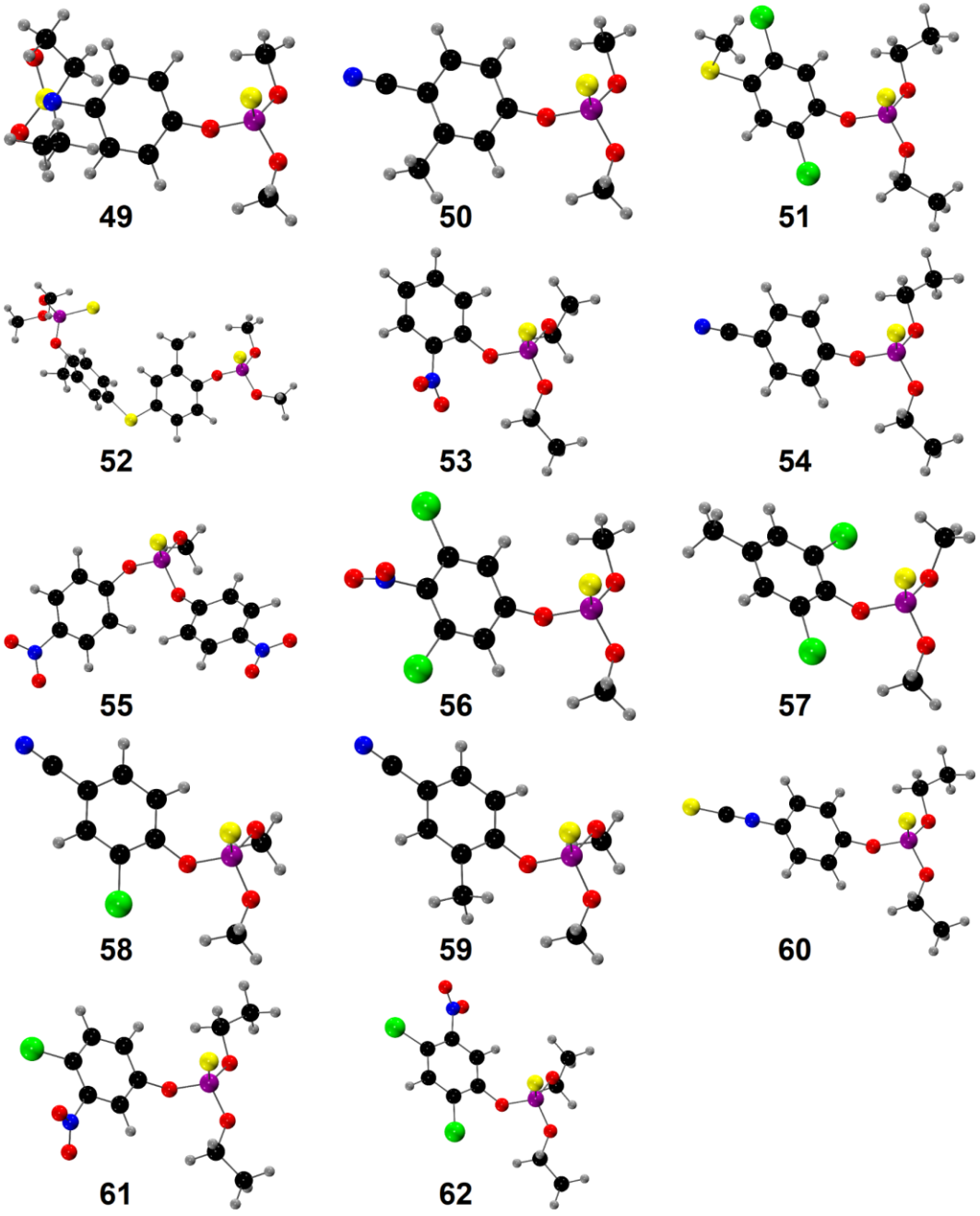

Figura 8 continuación.

Tabla 2 Etiqueta, identificador CAS y toxicidad en mmol/kg de los organotiofosforados estudiados.

ID	CAS	LD_{50} (mmol/kg)	ID	CAS	LD_{50} (mmol/kg)
01	52-60-8	1.1704	**32**	3735-01-1	1.7223
02	52-85-7	0.0861	**33**	3761-41-9	0.4247
03	55-37-8	3.4205	**34**	3761-42-0	0.4028
04	55-38-9	0.6467	**35**	4824-78-6	0.1320
05	56-38-2	0.0069	**36**	5826-76-6	1.6798
06	56-72-4	0.0358	**37**	5826-85-7	0.0520
07	97-17-6	0.5458	**38**	6012-87-9	0.0307
08	115-90-2	0.0071	**39**	7508-73-8	0.1741
09	115-93-5	0.5382	**40**	7533-79-1	0.3174
10	122-14-5	0.9018	**41**	13955-12-9	0.0172
11	298-00-0	0.0228	**42**	15037-62-4	2.7340
12	299-45-6	0.0448	**43**	15037-64-6	0.8671
13	299-84-3	1.9437	**44**	15037-65-7	0.8671
14	363-97-3	0.7548	**45**	15043-63-7	2.7340
15	500-28-7	0.9575	**46**	18181-70-9	5.6417
16	572-48-5	0.1819	**47**	18361-12-1	1.6194
17	1174-83-0	4.0123	**48**	19645-42-2	3.3792
18	1716-09-2	0.0457	**49**	21410-51-5	0.1289
19	2104-96-3	4.3716	**50**	21840-65-3	3.8873
20	2425-15-2	0.0328	**51**	21923-23-9	0.0360
21	2463-84-5	0.9541	**52**	24303-87-5	0.5055
22	2591-57-3	0.0103	**53**	29689-00-7	0.0858
23	2633-54-7	0.9357	**54**	33841-12-2	0.0369
24	2636-26-2	0.8840	**55**	39004-94-9	0.8426
25	3070-16-4	0.0378	**56**	50590-01-7	2.2584
26	3070-19-7	0.1535	**57**	57018-04-9	16.6042
27	3254-62-4	0.1486	**58**	63981-11-3	0.3601
28	3305-08-6	0.0819	**59**	76211-54-6	1.9436
29	3356-57-8	0.0241	**60**	84197-34-2	0.8274
30	3383-96-8	2.1437	**61**	84197-35-3	0.3070
31	3581-11-1	9.9177	**62**	84197-36-4	0.2777

3.2 CÁLCULO DE LOS DESCRIPTORES CUÁNTICOS

Para todos los compuestos se llevó a cabo una optimización completa de la geometría considerando el efecto solvente, mediante el empleo del campo de reacción autoconsistente (SCRF) disponible en el programa Gaussian (Rev. C.01) [71], y como solvente agua, utilizando el funcional ωB97XD [72] y la base orbital 6-311++G** [73]. Cabe destacar que el conjunto funcional/base elegido ha proporcionado resultados confiables en una amplia variedad de sistemas químicos [74–76]. La base LANL2DZ [77], que incluye los potenciales de núcleo efectivo, se empleó para los compuestos que contuvieran átomos de yodo en su estructura.

Mientras que, para todos los sistemas se llevó a cabo un estudio de frecuencias para garantizar que se obtuviera un mínimo real sobre la superficie de energía potencial. Los descriptores cuánticos globales y locales se calcularon a través de cDFT [78], que pueden utilizarse para determinar una relación estructura-toxicidad, como lo son: la energía del orbital molecular más alto ocupado (E_{HOMO}), la energía del orbital molecular más bajo desocupado (E_{LUMO}), potencial de ionización (I), afinidad electrónica (A), electronegatividad absoluta (χ), dureza (η), blandura (S), electrofilicidad (ω), nucleofilicidad (N_{Nu}), poder electrodonador (ω^-), poder electroaceptor (ω^+), volumen (V); las propiedades locales [45] como son el potencial químico local, la dureza local, Fukui kernel, dual descriptor kernel y los descriptores topológico QTAIM de Bader [46], para investigar la variación de algunas características electrónicas de los enlaces de los compuestos organotiofosforados empleando el programas AIMALL [79]. Se determinó la densidad electrónica, ρ(r), su laplaciano, $\nabla^2\rho(r)$, la densidad de energía cinética electrónica, G(r), la densidad de energía potencial electrónica, V(r), y la densidad electrónica de energía, H(r), de los puntos críticos de enlace (BCP). En total se obtuvieron 1232 propiedades que se utilizaron como posibles descriptores asociados a la reactividad química para generar los modelos QSTR.

3.3 Algoritmos de Aprendizaje Automatizado y la Generación de Modelos

En la actualidad, el uso de algoritmos de aprendizaje automatizado ha cobrado gran relevancia en la generación de modelos predictivos en diversas áreas de investigación. Estos algoritmos permiten que las máquinas aprendan a partir de datos y desarrollen patrones y reglas para predecir resultados futuros con mayor precisión. En particular, para este proyecto se construyeron tres tipos de modelos generados mediante algoritmos de aprendizaje: modelos de regresión lineal, modelos de clasificación lineal y modelos de regresión no lineal.

Los modelos de regresión lineal se utilizan para predecir valores numéricos de una variable de interés a partir de un conjunto de descriptores. Los modelos de clasificación lineal, por otro lado, se utilizan para predecir la pertenencia a una clase determinada de una variable categórica a partir de un conjunto de descriptores. Finalmente, los modelos de clasificación no lineal se utilizan para modelar relaciones más complejas entre variables, que no pueden ser representadas por una función lineal simple.

3.3.1 MODELOS QSTR DE REGRESIÓN LINEAL

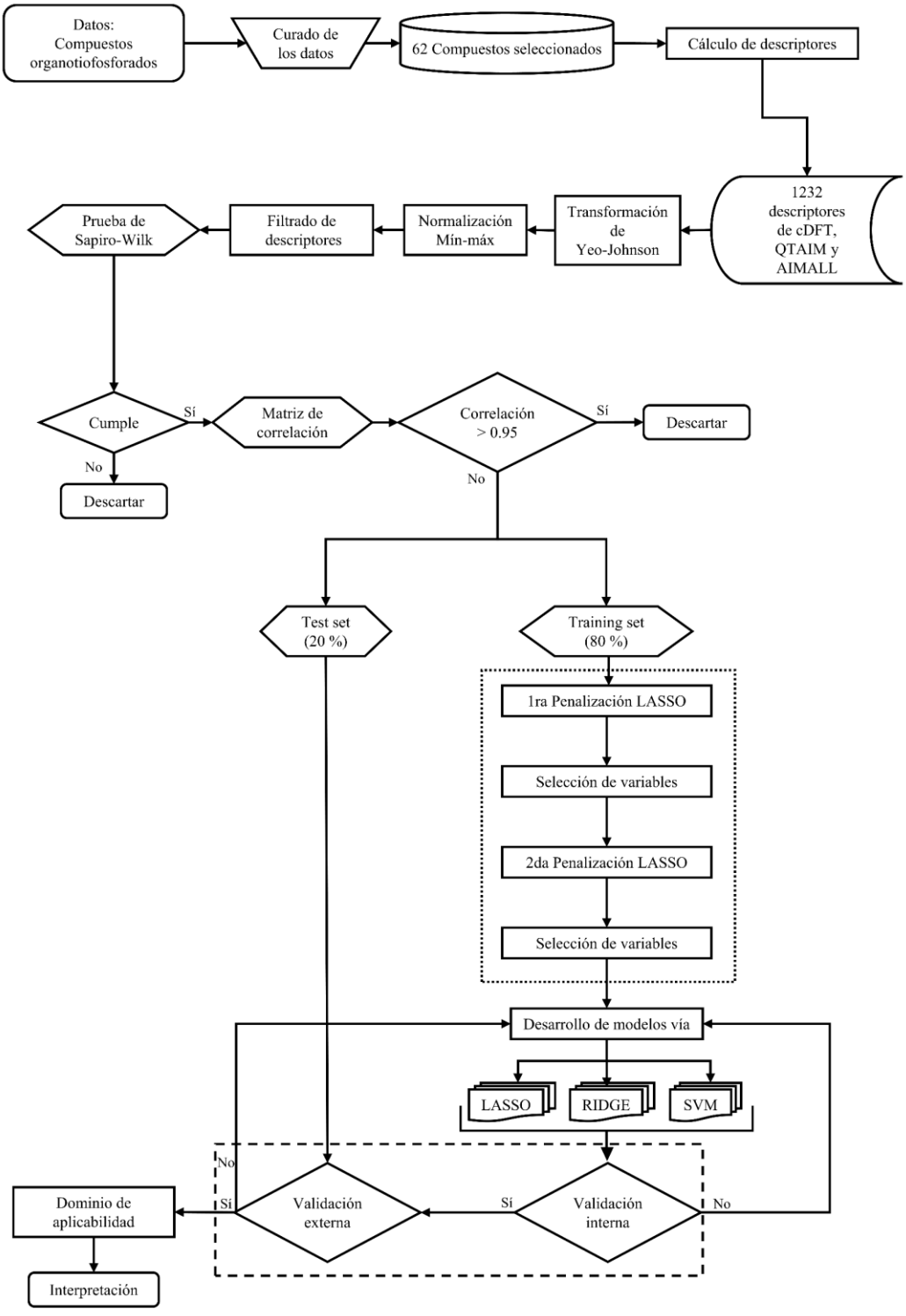

Figura 9 Diagrama de flujo para el desarrollo de los modelos de regresión lineal.

Se realizó un análisis de los datos de toxicidad LD_{50} en mmol/kg mediante la visualización de gráficos de cajas y violín. Para asegurar la fiabilidad de los datos, se eliminaron los valores atípicos, dejando un total de 44 compuestos de los 62 originales. Además, como se observa en la Figura 9, se aplicaron técnicas de filtrado de registros y atributos para descartar datos irrelevantes y se procedió a transformar y normalizar los datos. Para corregir la asimetría se utilizó el método de Yeo-Johnson [80,81], tanto para los 1232 descriptores como para los datos de toxicidad. A continuación, se escaló el conjunto de datos transformados mediante el algoritmo mín-máx, lo que resultó en valores normalizados entre 0 y 1. De la Figura 10 a 12 se presentan los gráficos de cajas y violín que muestran el cambio en la distribución de los datos después de la transformación previamente descrita. Para asegurar que las variables siguieran una distribución normal, se aplicó la prueba de Shapiro-Wilks [82] y se seleccionaron 364 variables que pasaron la prueba. Para reducir la redundancia en los datos, se eliminaron los descriptores con una correlación superior a 0.95 mediante el uso de una matriz de correlación, lo que dejó 154 descriptores. Finalmente, se dividieron los datos conformados por los 44 compuestos y 154 descriptores se dividieron en un conjunto de entrenamiento y prueba en una proporción 80/20, seleccionados aleatoriamente.

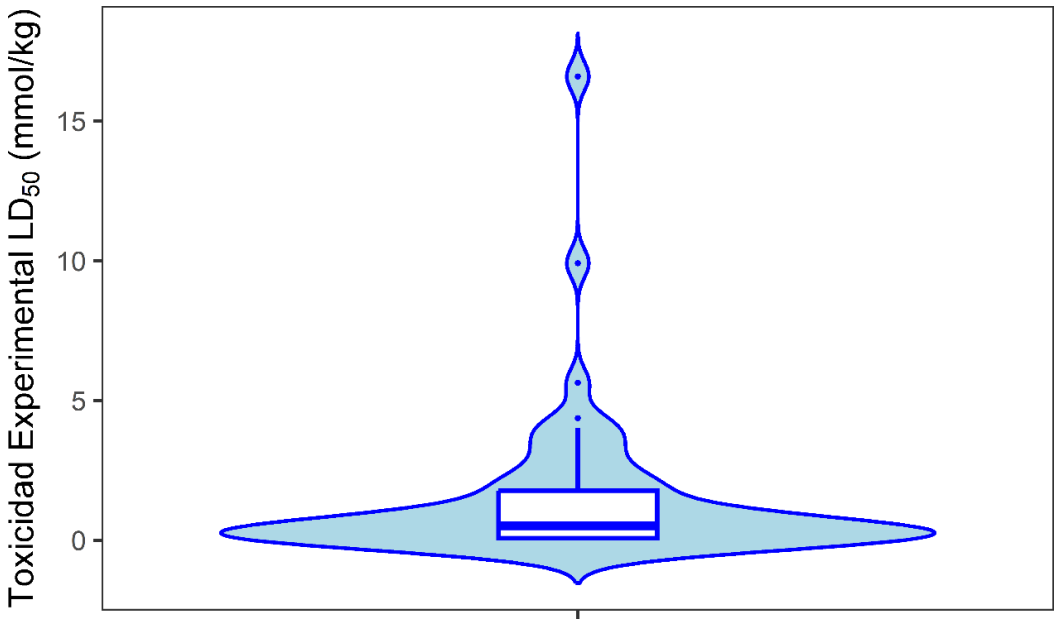

Figura 10 Gráficos de caja y violín para la toxicidad LD_{50} en mmol/kg para los 62 compuestos organotiofosforados.

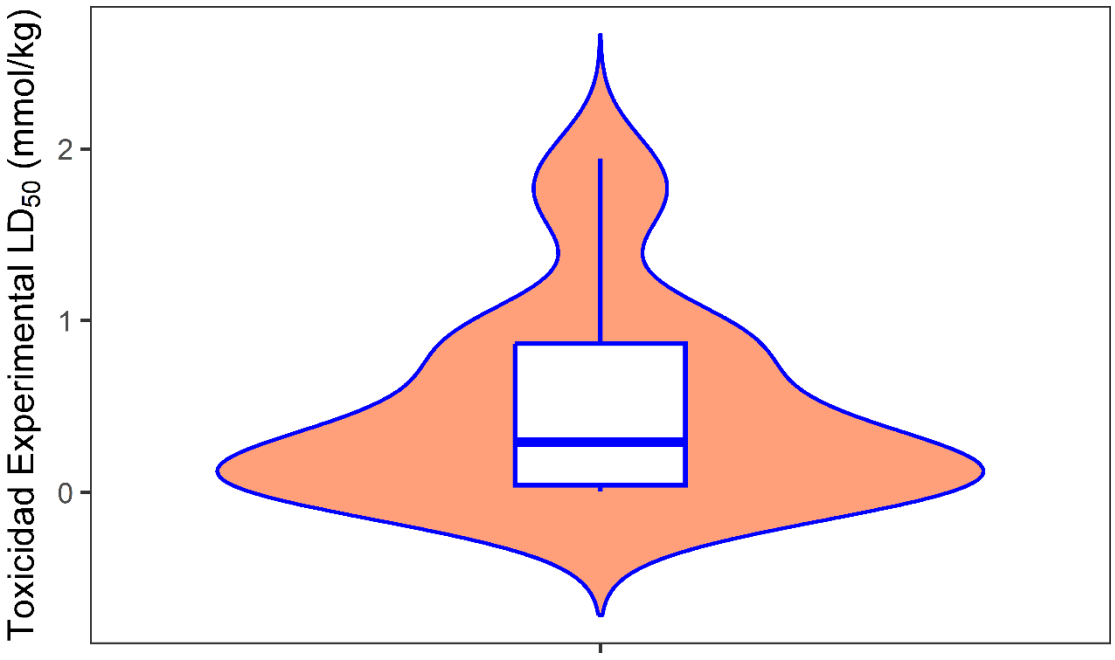

Figura 11 Gráficos de caja y violín para la toxicidad LD_{50} en mmol/kg para los compuestos sin valores de toxicidad atípicos.

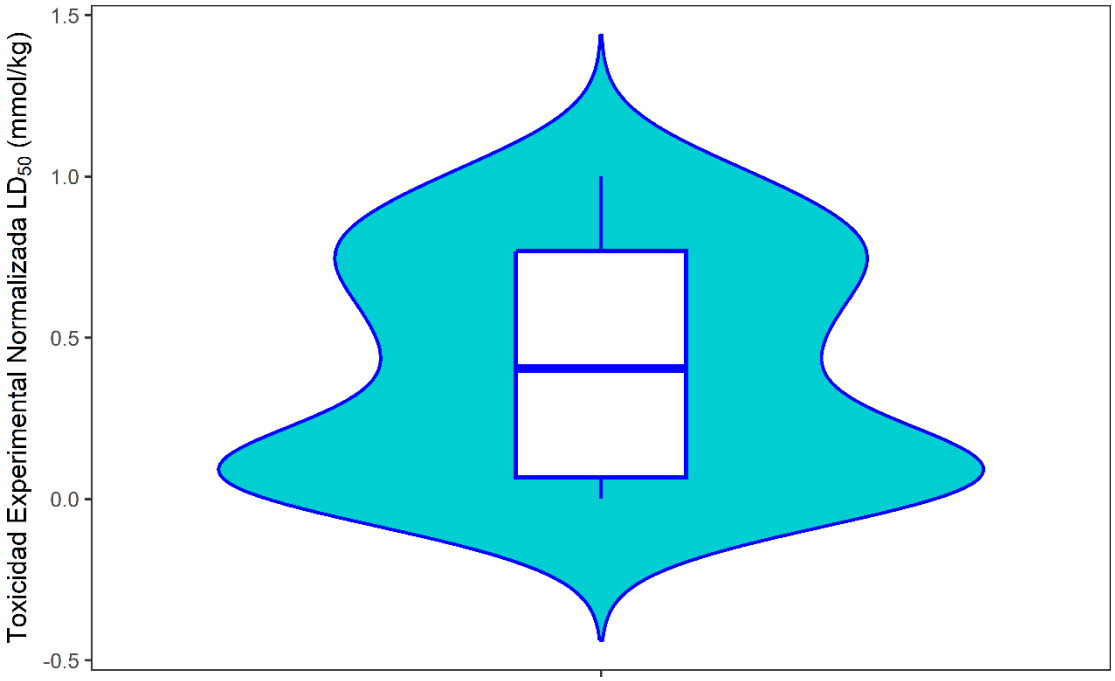

Figura 12 Gráficos de caja y violín para la toxicidad LD_{50} en mmol/kg la toxicidad transformada y normalizada.

Para simplificar la complejidad de los modelos generados, se utilizó una doble penalización con el método LASSO. En la primera penalización, se buscó el valor de λ óptima que minimizara el RMSE mediante un grid de 0 a 1 con 100 puntos equidistantes, utilizando una validación interna cruzada de 5 folds y 100 repeticiones. El gráfico de RMSE vs λ se muestra en la Figura 13, y las variables óptimas de este paso se presentan en la Figura 14, las cuales se usaron en la segunda penalización. Para esta última, se utilizó un grid de 0 a 0.5 con 500 puntos equidistantes y una validación interna cruzada de 10 folds y 100 repeticiones, cuyo gráfico se muestra en la Figura 15. De este paso, se obtuvieron 7 variables, mostradas en la Figura 16, que se emplearon para construir los modelos de regresión.

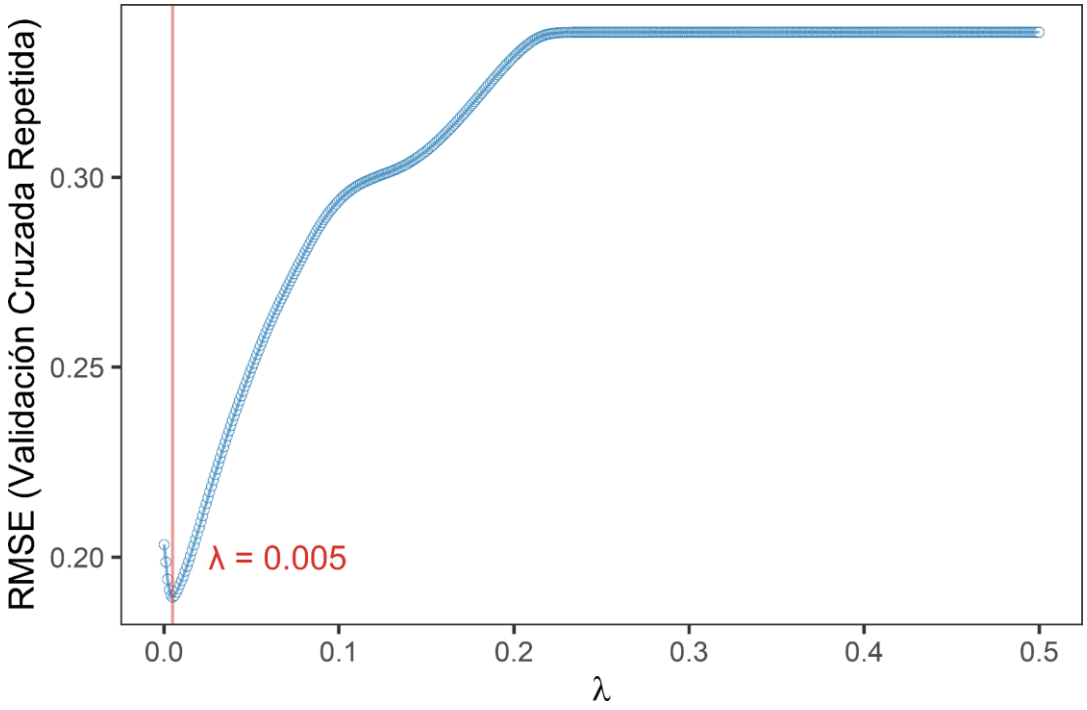

Figura 13 Gráfico de RMSE vs λ para la primera penalización LASSO.

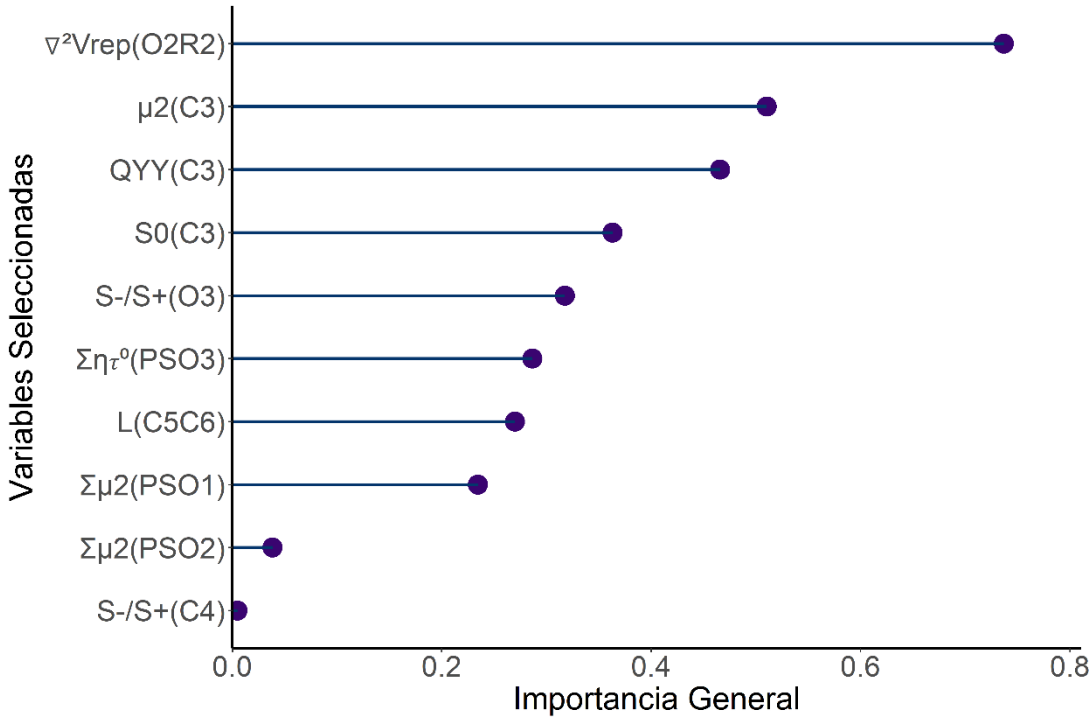

Figura 14 Importancia de las variables óptimas para la primera penalización LASSO al mejor valor de λ.

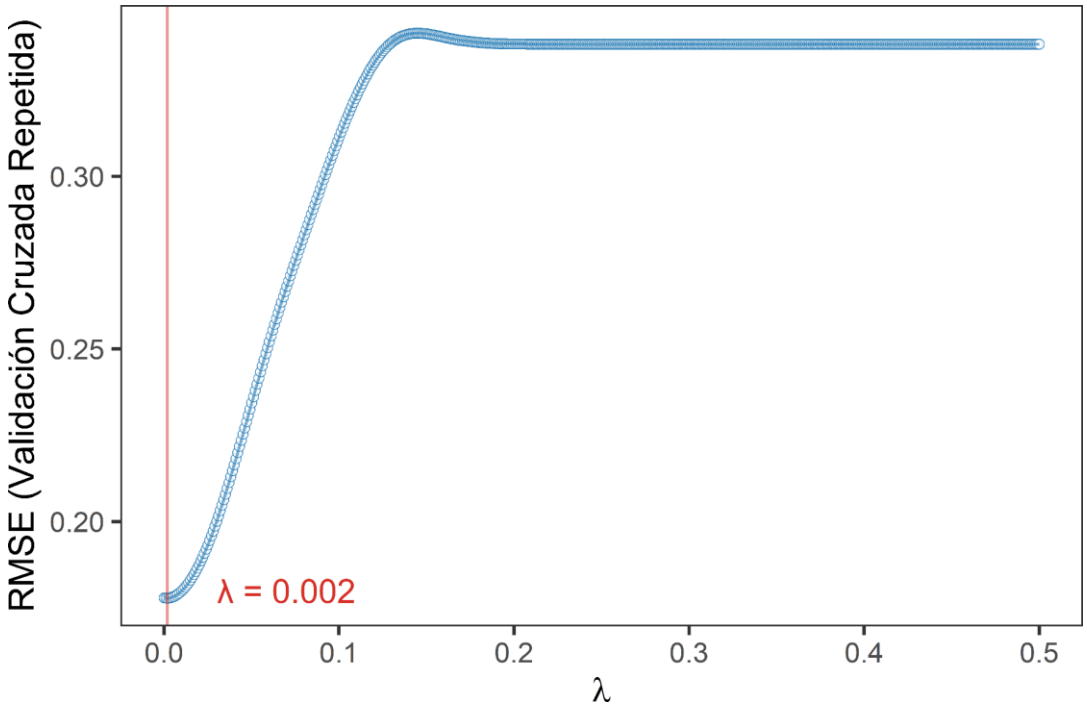

Figura 15 Gráfico de RMSE vs λ para la segunda penalización LASSO.

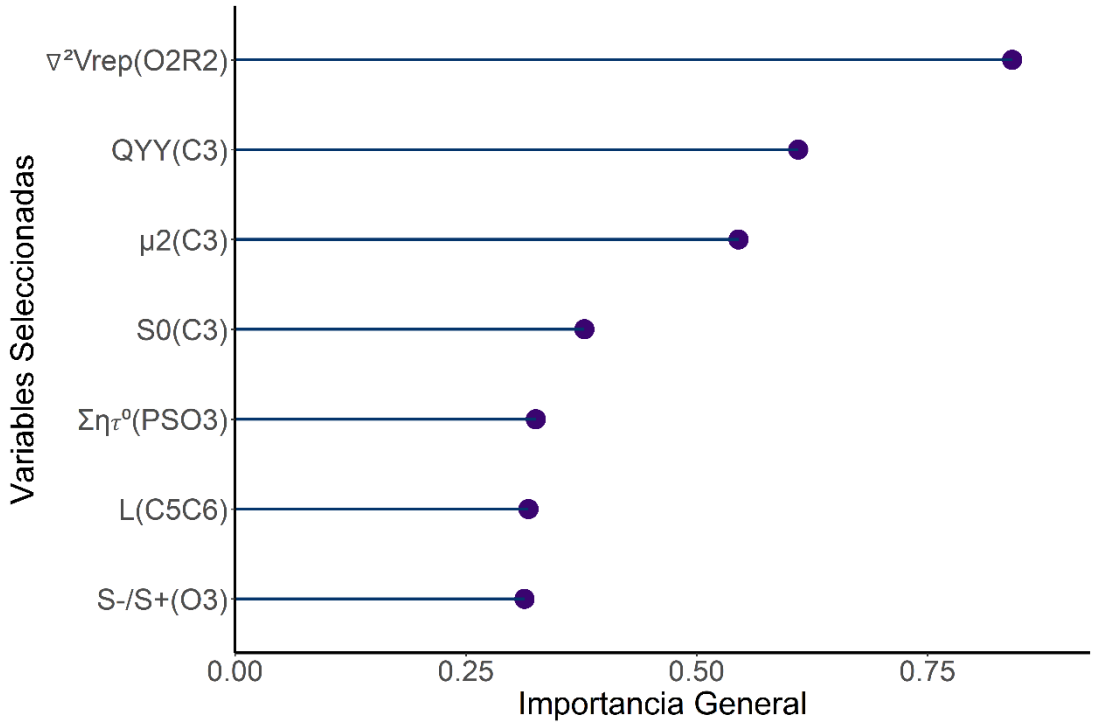

Figura 16 Importancia de las variables óptimas para la segunda penalización LASSO al mejor valor de λ.

El valor de λ óptimo de la segunda penalización, así como los mismos parámetros de validación interna se emplearon para construir el modelo LASSO, mostrándose en las Figuras 17 y 18 el gráfico de regresión y residuales del modelo. Por otro lado, se generaron los modelos RIDGE y SVM mediante una validación interna cruzada de 10 folds y 100 repeticiones. Se llevó a cabo una búsqueda de los parámetros λ y C a través de un grid de 0 a 0.5 y 0.01 a 0.5, respectivamente, con 500 puntos equidistantes para ambos casos. Los resultados de esta búsqueda se reflejan en las Figuras 19 y 20, donde se pueden observar los valores óptimos de las penalizaciones. A su vez, los gráficos de regresión y residuales de ambos modelos se presentan en las Figuras 21, 26, 27 y 24.

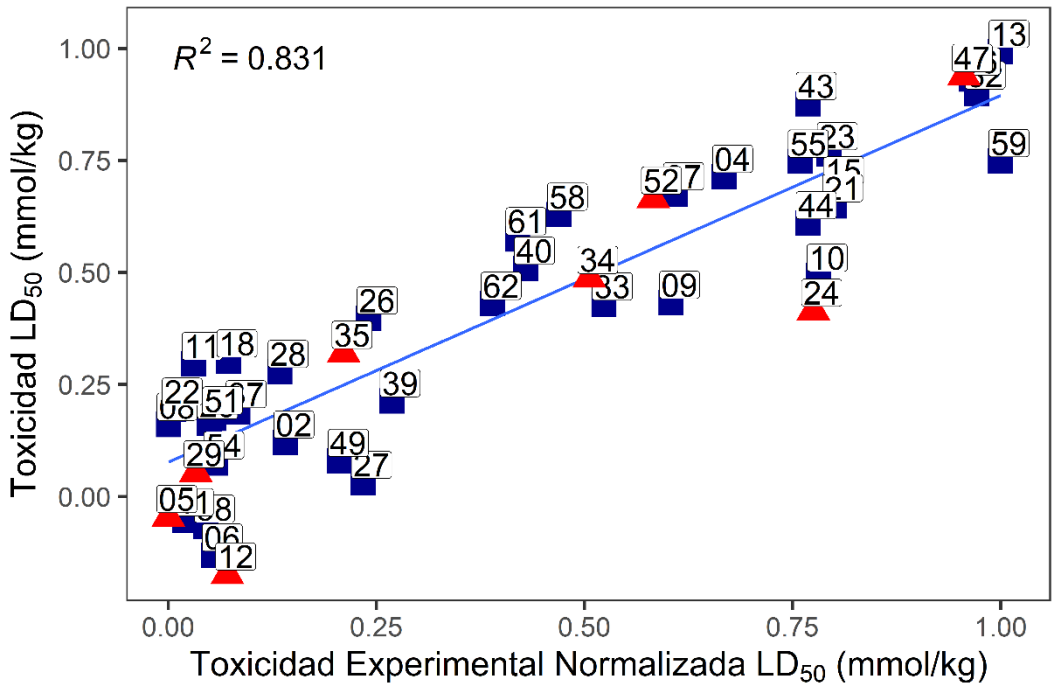

Figura 17 Gráfico de dispersión para los valores predichos por el modelo vs los valores de toxicidad experimental normalizada para el modelo LASSO.

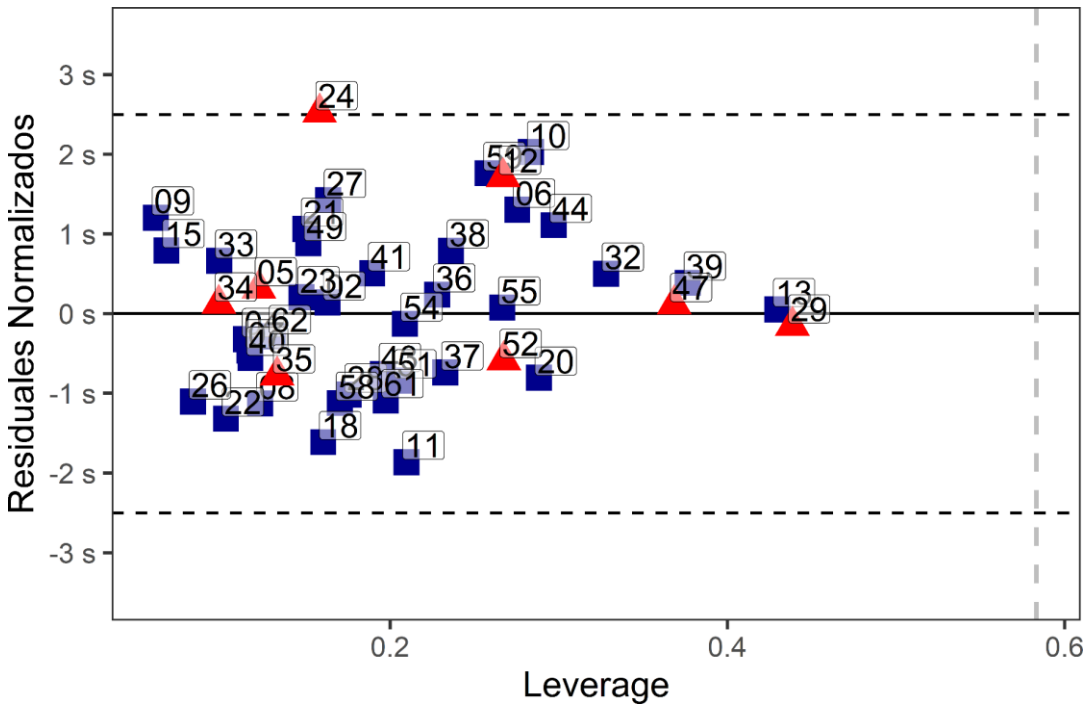

Figura 18 Gráfico de los residuales predichos por el modelo vs leverage para el modelo LASSO. La línea vertical representa el dominio de aplicabilidad.

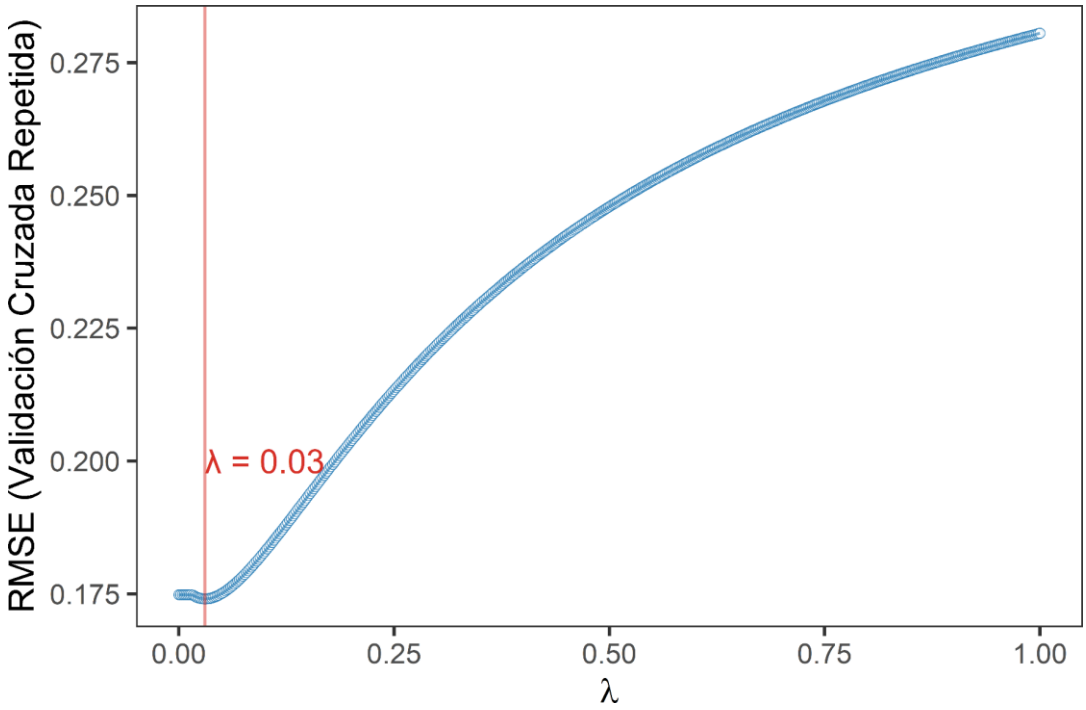

Figura 19 Gráfico de RMSE vs λ para el modelo RIDGE.

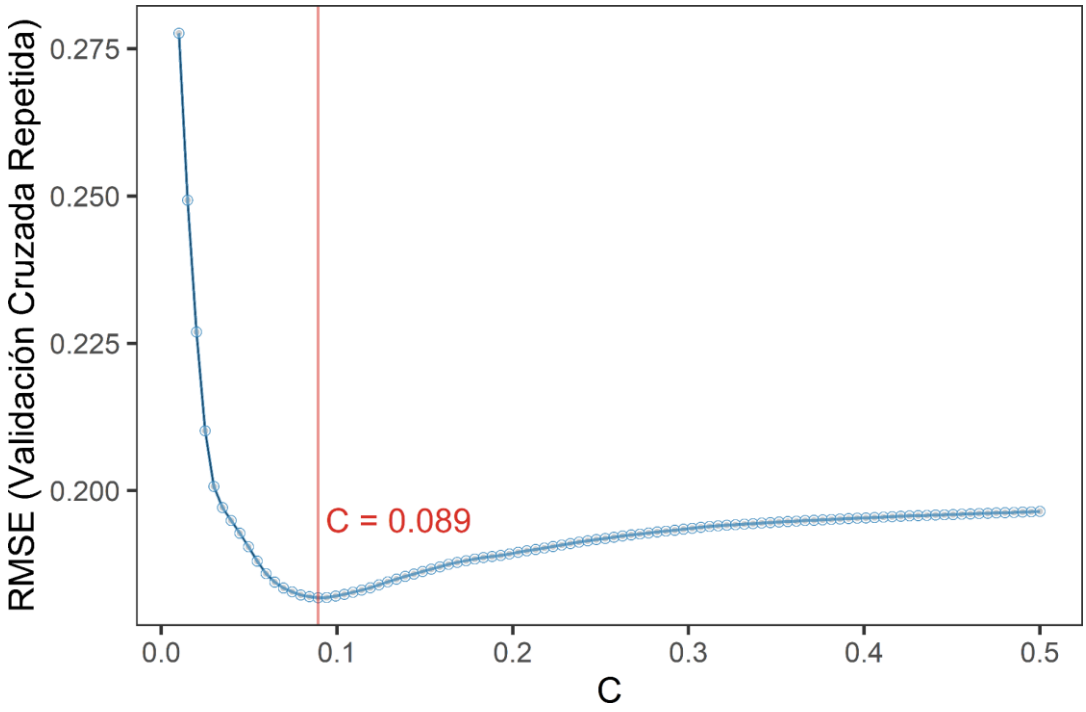

Figura 20 Gráfico de RMSE vs C para el modelo SVM.

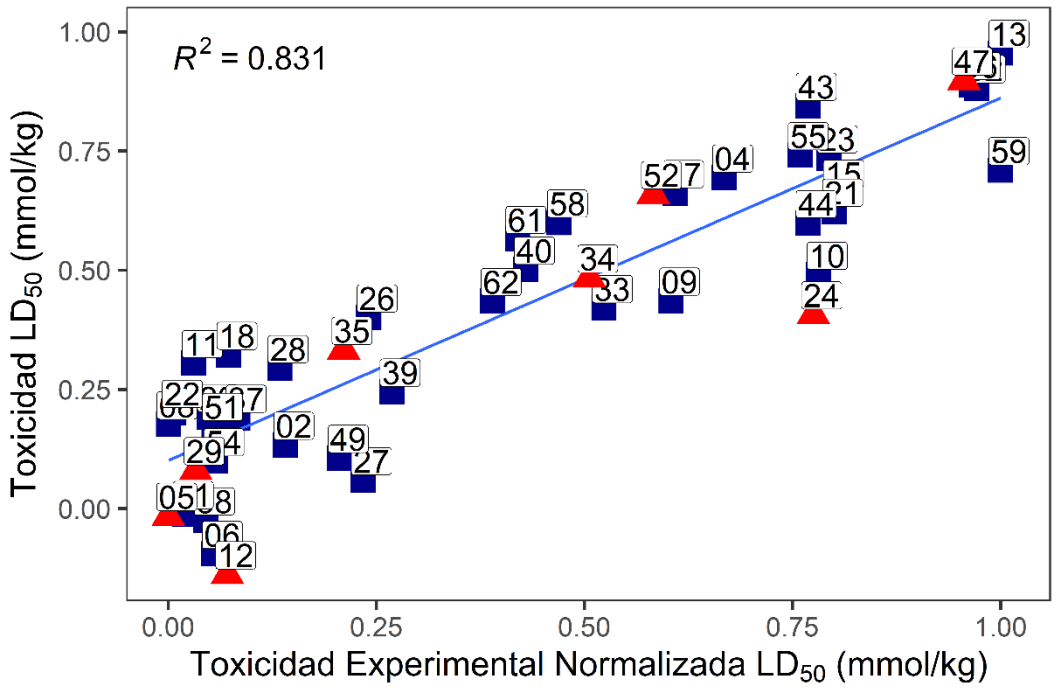

Figura 21 Gráfico de dispersión para los valores predichos por el modelo vs los valores de toxicidad experimental normalizada para el modelo RIDGE.

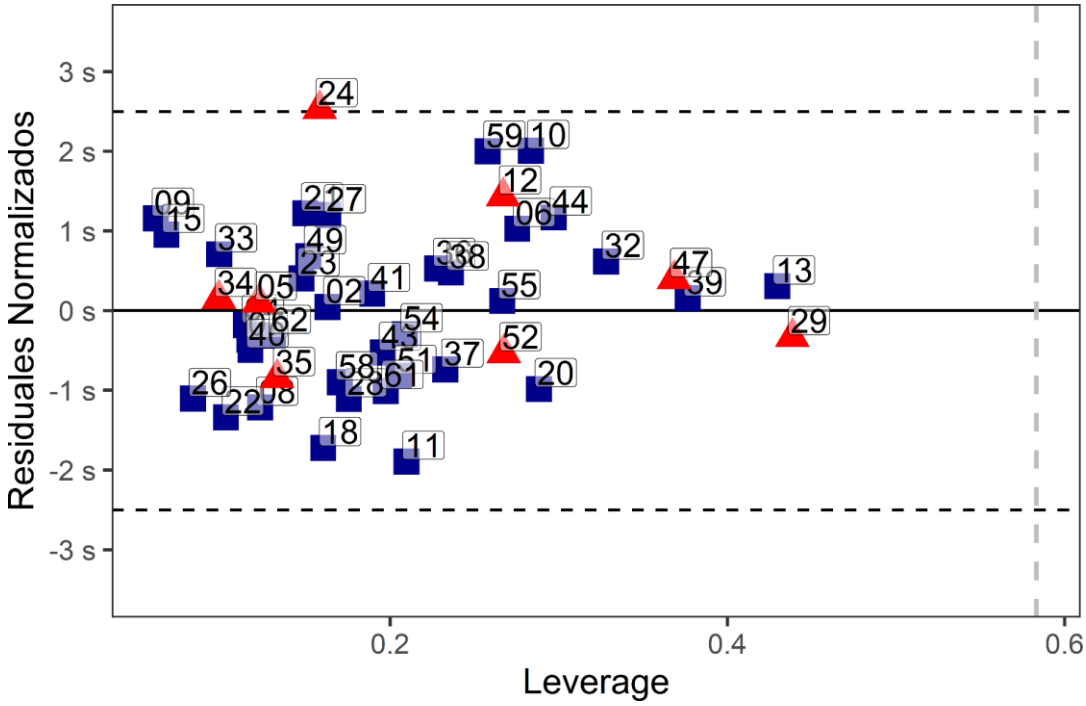

Figura 22 Gráfico de los residuales predichos por el modelo vs leverage para el modelo RIDGE. La línea vertical representa el dominio de aplicabilidad.

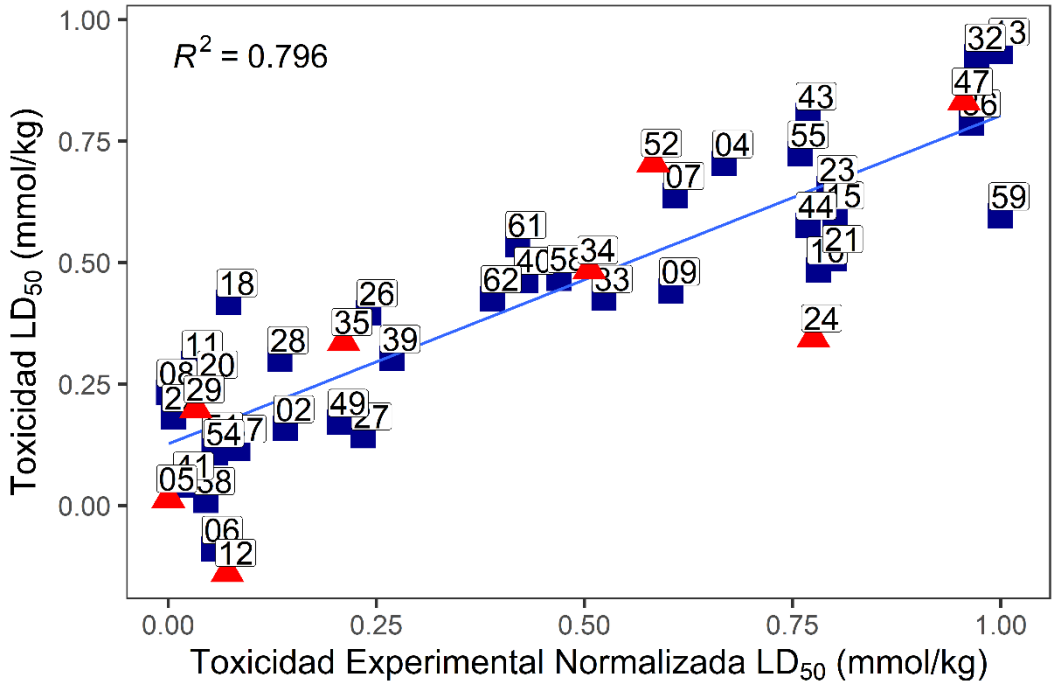

Figura 23 Gráfico de dispersión para los valores predichos por el modelo vs los valores de toxicidad experimental normalizada para el modelo SVM.

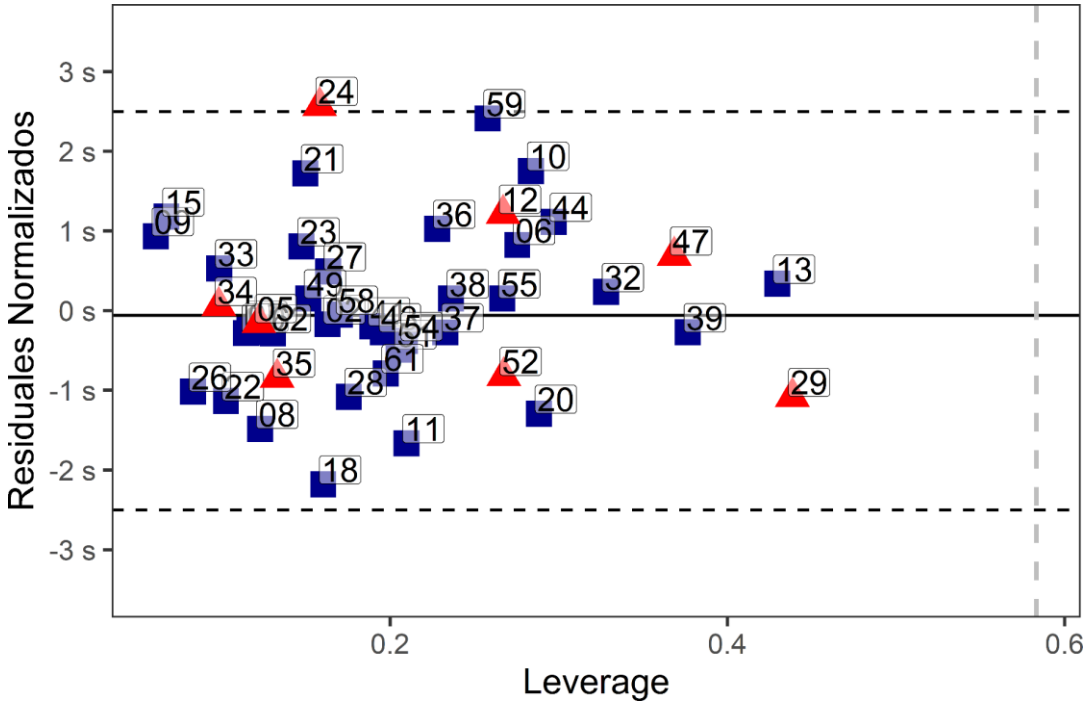

Figura 24 Gráfico de los residuales predichos por el modelo vs leverage para el modelo SVM. La línea vertical representa el dominio de aplicabilidad.

Tras la construcción de los modelos de regresión lineal, se evaluó su desempeño mediante el análisis de los parámetros de validación propuestos por Todeschini [83]. Los resultados mostraron que tanto el modelo LASSO como RIDGE cumplían con los parámetros de validación establecidos en la Figura 25. Sin embargo, dado que el modelo LASSO fue el primero en ser obtenido mediante la metodología previamente descrita y que presentó un buen desempeño estadístico, fue seleccionado como el mejor modelo.

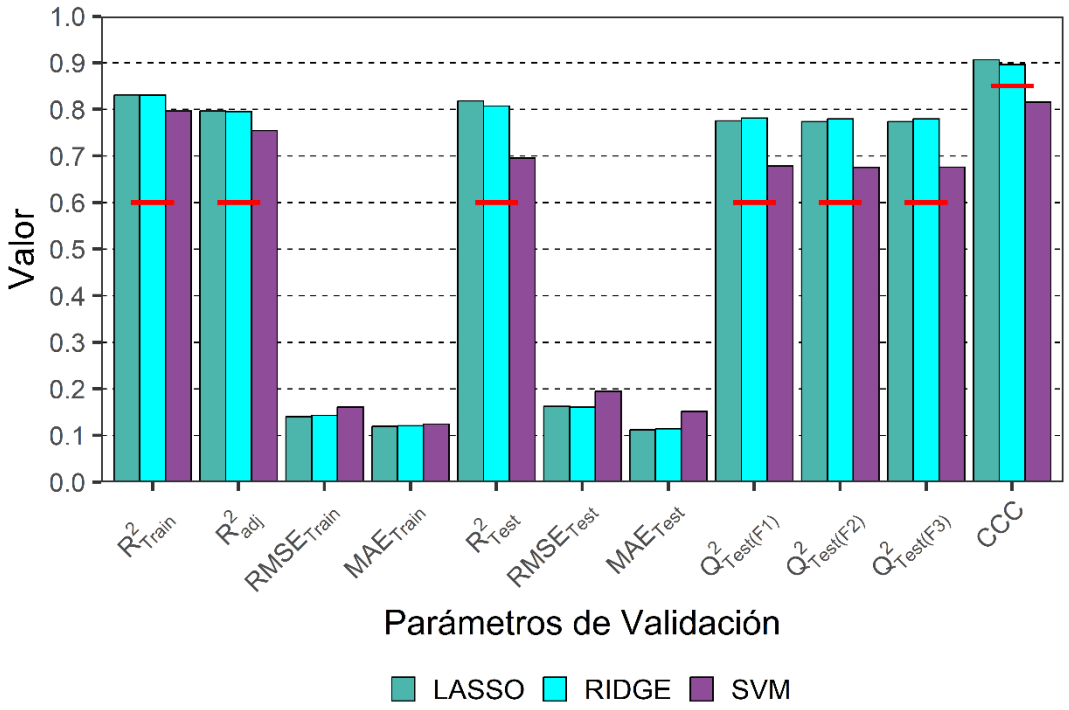

Figura 25 Parámetros de validación para los modelos de regresión QSTR lineales, elaborados con las metodologías LASSO, RIDGE y SVM. En rojo se presentan las características mínimas con las que un modelo se considera estadísticamente válido.

Las 7 variables involucradas en los modelos son: el laplaciano de la contribución repulsiva al momento virial del enlace entre los átomos O_2-R_2 ($\boldsymbol{\nabla^2 V_{rep}}_{(\boldsymbol{O_2-R_2})}$), el lagrangiano de la densidad del enlace C_5-C_6 ($\boldsymbol{L}_{(\boldsymbol{C_5-C_6})}$), el índice de nucleofilicidad relativa de O_3 ($\boldsymbol{S^+/S^-_{(O_3)}}$), blandura local condensada para un ataque radical de O_3 ($\boldsymbol{S^\circ}_{(\boldsymbol{C_3})}$), el momento cuadrupolar en el plano Y de C_3 ($\boldsymbol{Q_{YY}}_{(\boldsymbol{c_3})}$), el

segundo eigenvalor del momento cuadrupolar para C_3 ($\boldsymbol{\mu 2_{(C_3)}}$) y la suma de la dureza relativa para ω=0 del fragmento P-S-O_3 ($\sum \boldsymbol{\eta^{\tau\circ}_{(P,S,O_3)}}$). En la Figura 26 se muestra una comparativa gráfica de los coeficientes obtenidos en los modelos de regresión.

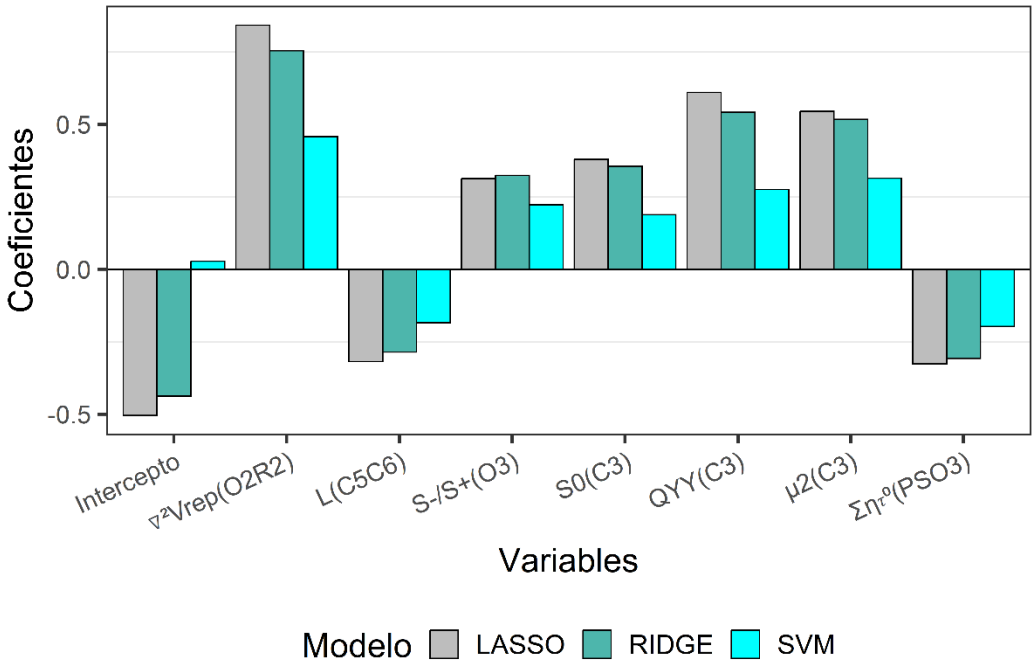

Figura 26 Comparación de los coeficientes de las variables obtenidas en los modelos de regresión.

El modelo de regresión LASSO sigue la siguiente ecuación:

$$\boldsymbol{LD_{50(Normalizado)}} = -0.5042 + 0.8421\boldsymbol{\nabla^2 V_{rep_{(O_2-R_2)}}} - 0.3181\boldsymbol{L_{(C_5-C_6)}} + 0.3181\,\boldsymbol{S^+/S^-_{(O_3)}} + 0.3788\boldsymbol{S^\circ_{(C_3)}} + 0.6103\boldsymbol{Q_{YY_{(C_3)}}} + 0.5456\boldsymbol{\mu 2_{(C_3)}} - 0.3261\sum \boldsymbol{\eta^{\tau\circ}_{(P,S,O_3)}}$$

En el modelo de regresión LASSO se identificaron como variables más importantes $\boldsymbol{\nabla^2 V_{rep_{(O_2-R_2)}}}$, $\boldsymbol{Q_{YY_{(C_3)}}}$ y $\boldsymbol{\mu 2_{(C_3)}}$, todas con una relación positiva con la disminución de la toxicidad de los compuestos. Por el contrario, las variables $\boldsymbol{L_{(C_5-C_6)}}$ y $\boldsymbol{\eta^{\tau\circ}_{(P,S,O_3)}}$, así como el intercepto, presentaron una relación negativa y, por lo tanto, asociación con un aumento de la toxicidad. Se destaca que el momento cuadrupolar es

una medida de la distribución de la carga eléctrica en una molécula y en el caso de $\boldsymbol{Q_{YY_{(C_3)}}}$ y $\boldsymbol{\mu 2_{(C_3)}}$, ambas están relacionadas con el momento cuadrupolar de C_3, lo que sugiere que la distribución de carga eléctrica en esta región de la molécula es importante para la reactividad. Específicamente, $\boldsymbol{Q_{YY_{(C_3)}}}$ puede indicar una asimetría en la distribución de carga eléctrica en esta región, lo que puede influir en la interacción de la molécula con otros compuestos o biomoléculas. El segundo eigenvalor del momento cuadrupolar $\boldsymbol{\mu 2_{(C_3)}}$, puede indicar una orientación específica de la distribución de carga eléctrica en esta región, lo que también puede influir en la reactividad.

En la relación $\boldsymbol{\nabla^2 V_{rep_{(O_2-R_2)}}}$ se destaca que esta variable se relaciona con la contribución repulsiva al momento virial del enlace O_2-R_2, lo que sugiere que la distancia y la orientación relativa entre el átomo de fósforo y los sustituyentes en los oxígenos pueden ser importantes para la reactividad. Asimismo, se identificó que la región alrededor de C_3 puede ser crítica para la reactividad, mientras que $\boldsymbol{\sum \eta^{\tau\circ}_{(P,S,O_3)}}$ sugiere que la región cercana al enlace P=S también es importante.

3.3.2 MODELOS QSTR DE CLASIFICACIÓN LINEAL

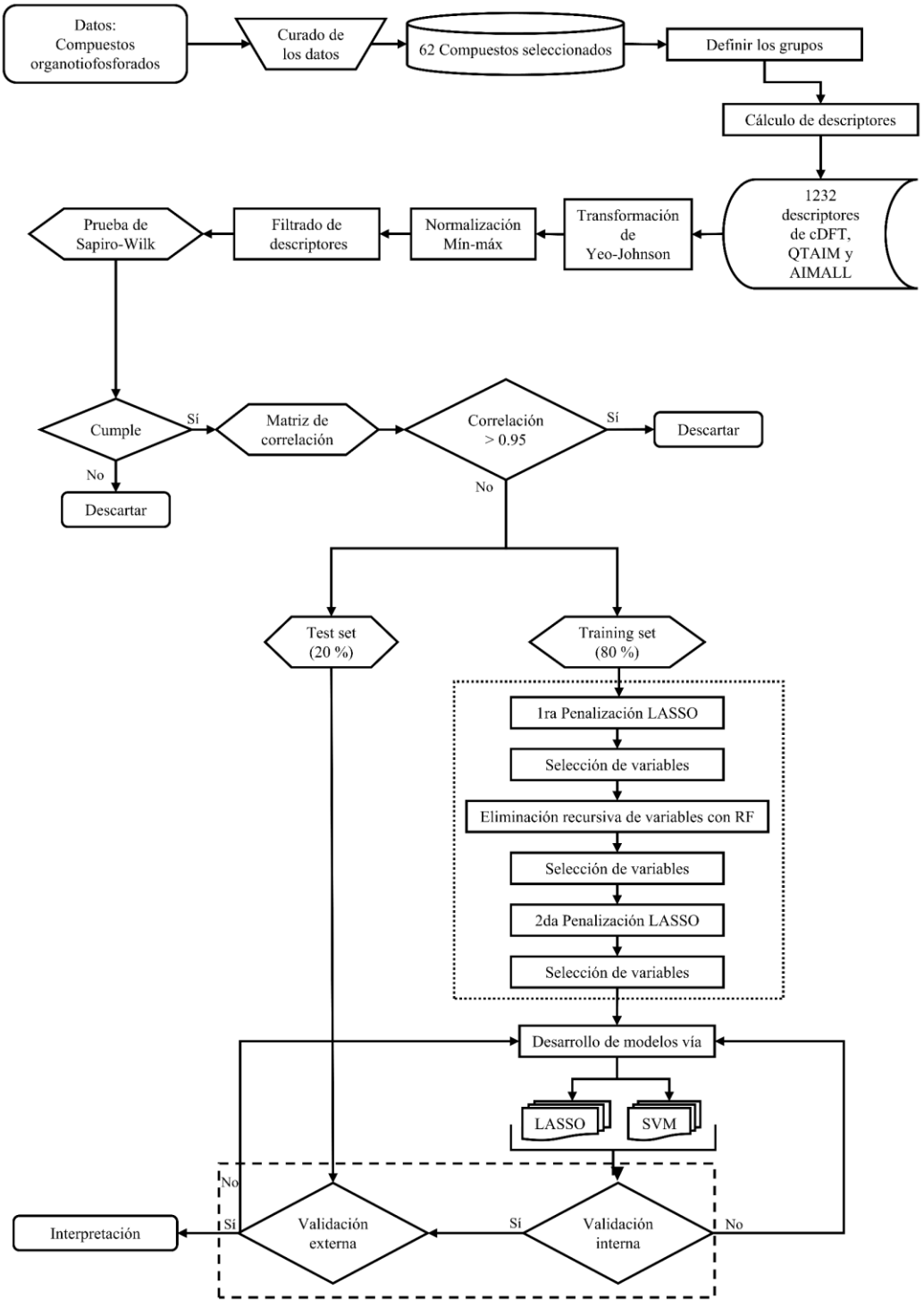

Figura 27 Diagrama de flujo para el desarrollo de los modelos de clasficación.

Tabla 3 Clasificación de los compuestos mediante intervalos de toxicidad propuestos por la OMS.

Intervalo LD_{50}/(mg/kg)	**Clasificación**	**Etiqueta**
$(\mathbf{0-5}]$	Extremadamente tóxico	ET
$(\mathbf{5-50}]$	Altamente tóxico	HT
$(\mathbf{50-2000}]$	Medianamente tóxico	MT
$(\mathbf{2000-5000}]$	Ligeramente tóxico	ST
$(\mathbf{5000}-\infty)$	No tóxico	NT

La OMS [9] establece intervalos en los valores de toxicidad LD_{50} en mg/kg para catalogar a los pesticidas, los cuales se muestran en la Tabla 3. Tomando en cuenta lo anterior, se tomaron los valores de toxicidad en mg/kg de los 62 compuestos organotiofosforados estudiados y se procedió a clasificarlos acorde a estos parámetros, sin embargo, la cantidad de datos que se encuentran en los extremos es insuficiente para hacer un análisis de agrupación pertinente sobre ellos, por lo que se procedió a colapsar los intervalos de la siguiente manera: $(0-50]$ HT y $(50-\infty)$ MT, como se muestra en la Figura 28.

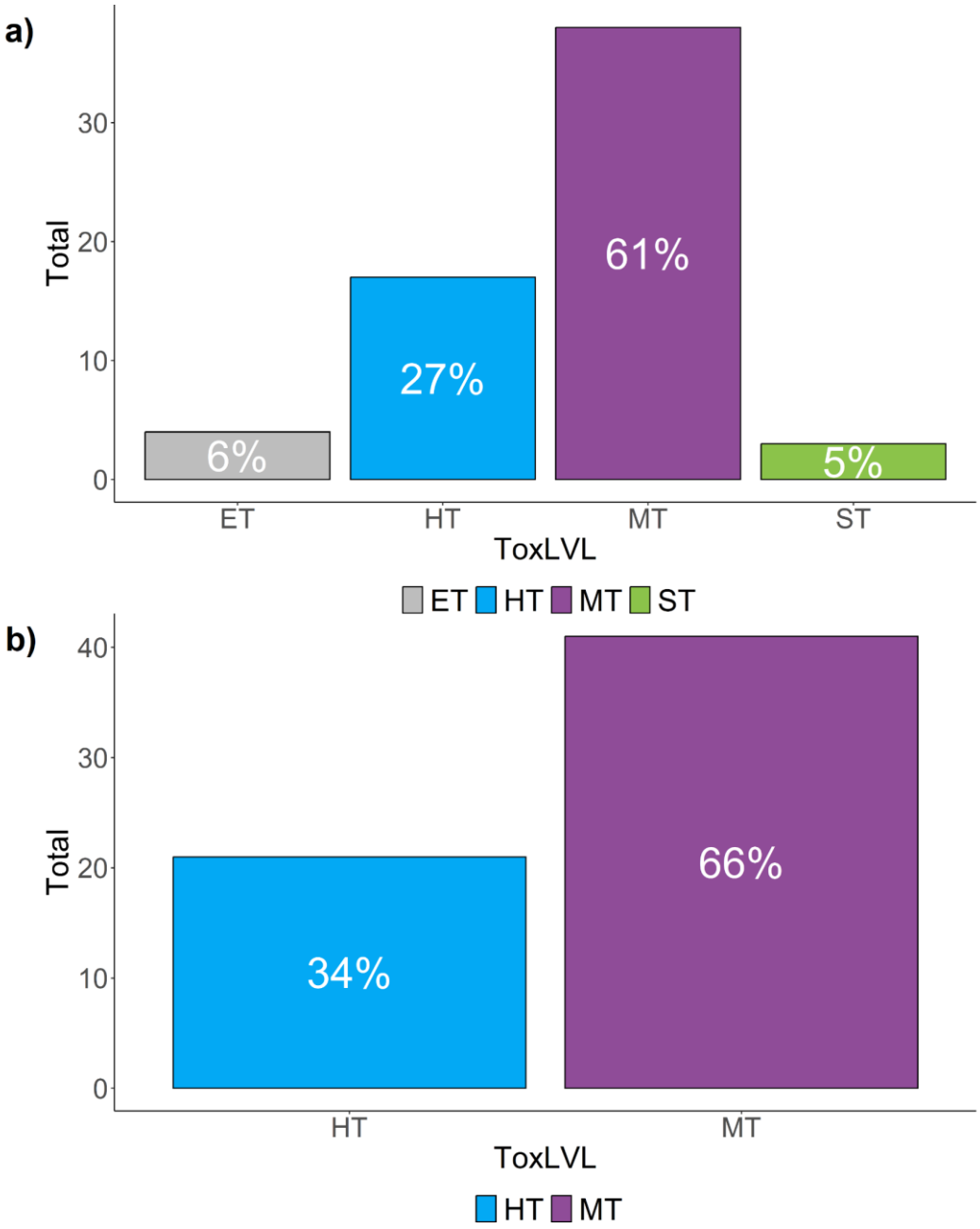

Figura 28 a) Compuestos organotiofosforados estudiados divididos por los parámetros de toxicidad propuestos por la OMS. b) Clasificación colapsada empleada para generar los modelos.

Se siguió la misma metodología para la transformación, normalización y eliminación de las variables intercorrelacionadas que en la obtención del modelo lineal de regresión, como se observa en el diagrama de flujo de la Figura 27. En total, se

obtuvieron 128 descriptores para los 62 compuestos que se dividieron aleatoriamente en un conjunto de entrenamiento y otro de prueba en una proporción 80/20. Para reducir el número de variables y mejorar la capacidad predictiva del modelo, se realizaron tres pasos: una penalización inicial de tipo LASSO, una eliminación recursiva de variables utilizando Random Forest [84] y una segunda penalización con LASSO. La primera penalización se llevó a cabo mediante una validación interna cruzada de 5-folds y 100 repeticiones, y se utilizó una búsqueda en grid de 0 a 1 con 100 puntos equidistantes para obtener el valor óptimo del parámetro de penalización λ, resultando en 24 descriptores, como se muestra en las Figuras 29 y 30. Posteriormente, en la Figura 31 se muestra la eliminación recursiva de variables con Random Forest que se realizó con una validación interna cruzada de 5-folds y 10 repeticiones, y una predicción con una validación interna de 10 repeticiones. Se utilizó una búsqueda en grid de mtry de 2-10 con 5 puntos equidistantes, lo que resultó en la selección de los 9 descriptores. Finalmente, se aplicó una segunda penalización con LASSO utilizando una validación interna cruzada de 10-folds, 100 repeticiones y una búsqueda de λ de 0 a 1 con 500 puntos equidistantes, lo que permitió seleccionar 8 variables, como se observa en las Figuras 32 y 33. Para el modelo RIDGE, se emplearon las variables seleccionadas anteriormente, y se utilizó una validación interna cruzada de 10-folds y 100 repeticiones, y una búsqueda de los valores óptimos del parámetro de penalización λ en un grid de 0 a 1 con 500 puntos equidistantes, Figura 34. En la Figura 35 se muestran los parámetros de validación para los modelos.

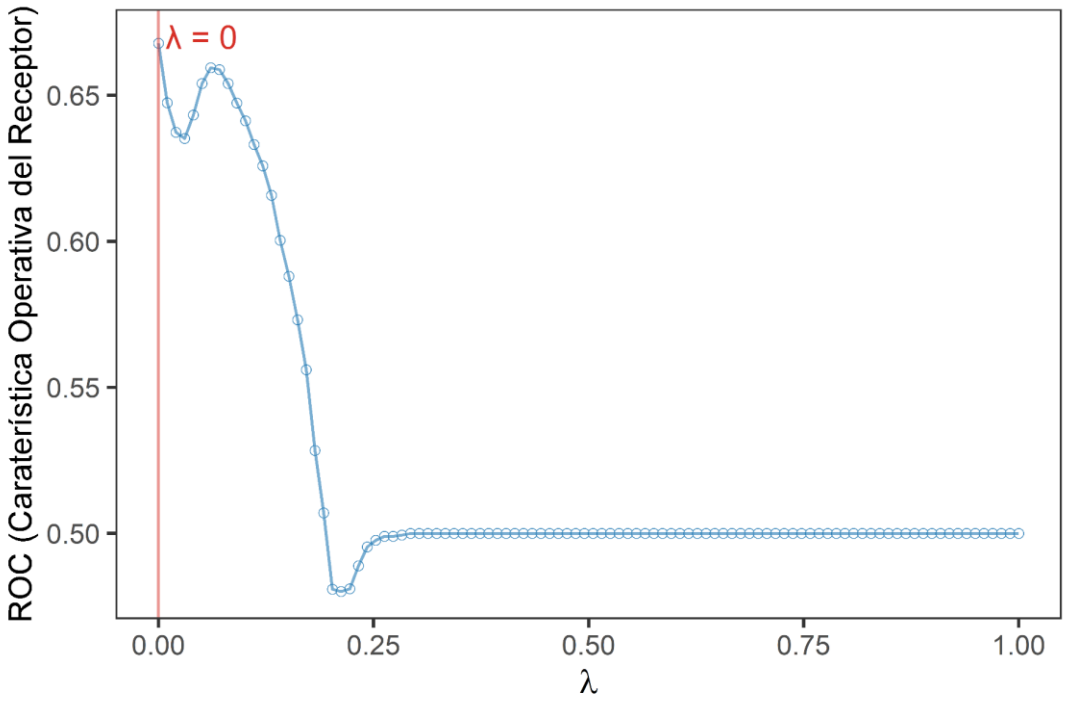

Figura 29 Gráfico de ROC vs λ para la primera penalización LASSO.

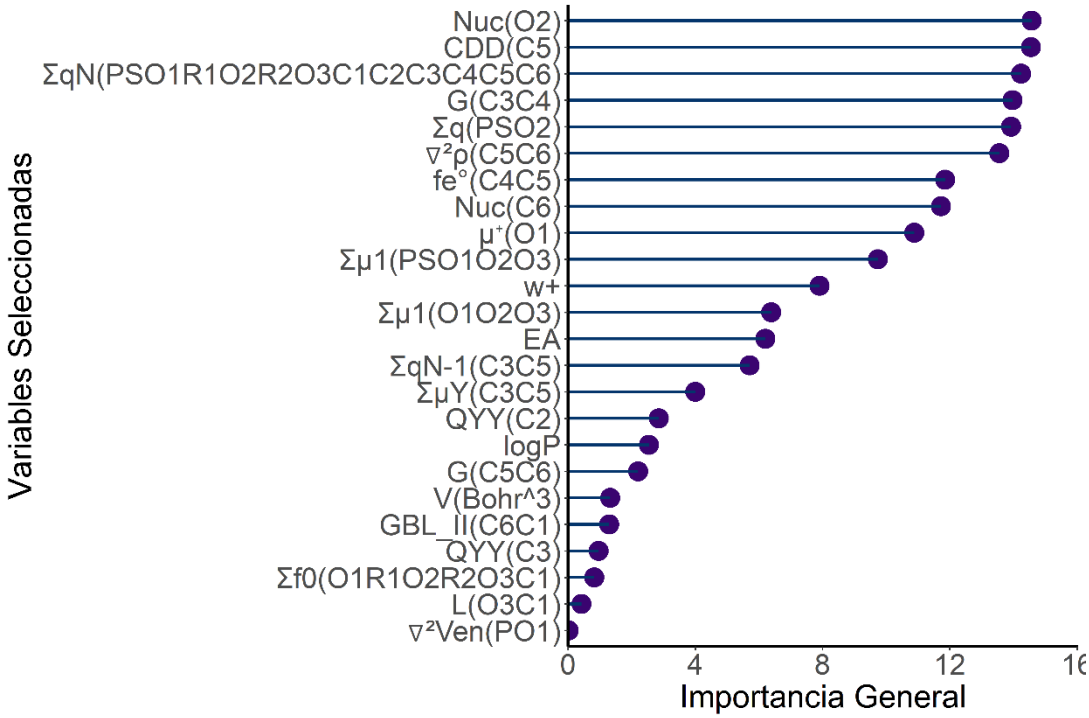

Figura 30 Importancia de las variables óptimas para la primera penalización LASSO al mejor valor de λ.

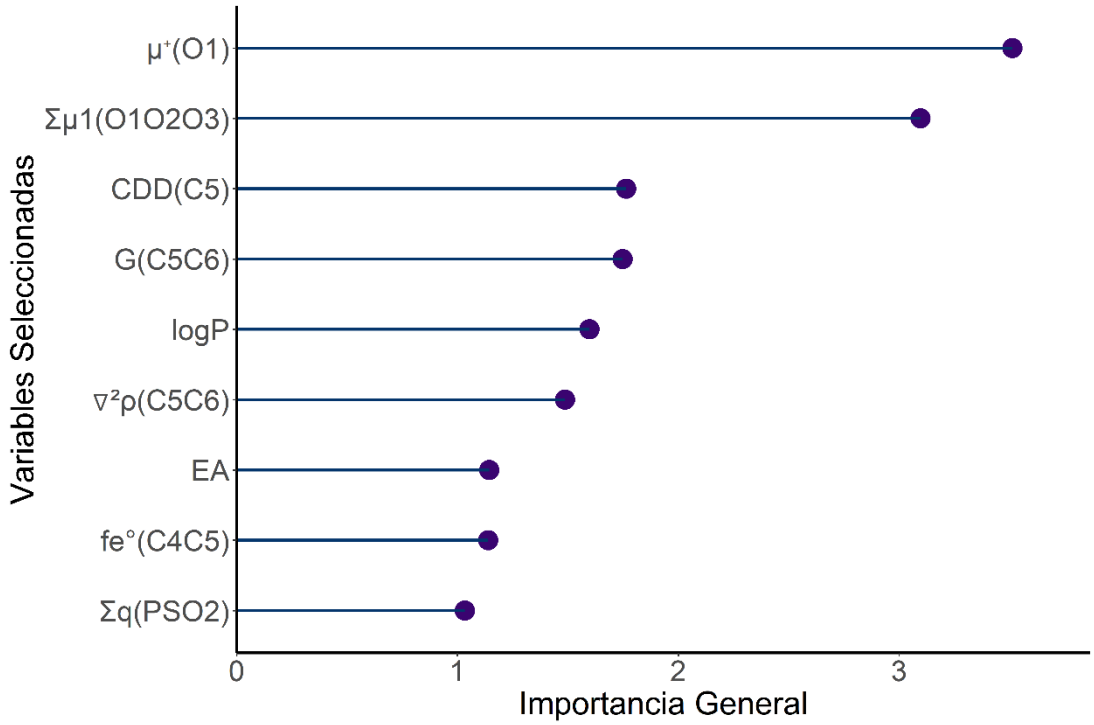

Figura 31 Importancia de las variables óptimas para la eliminación recursiva de variables empleando Random Forest.

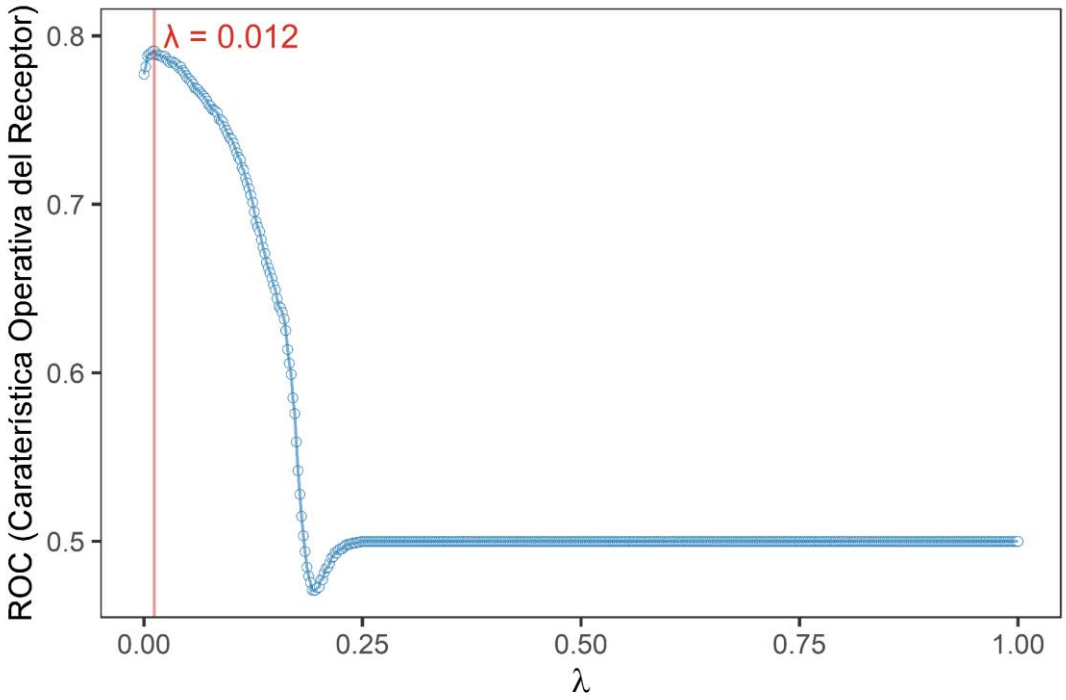

Figura 32 Gráfico de ROC vs λ para la segunda penalización LASSO.

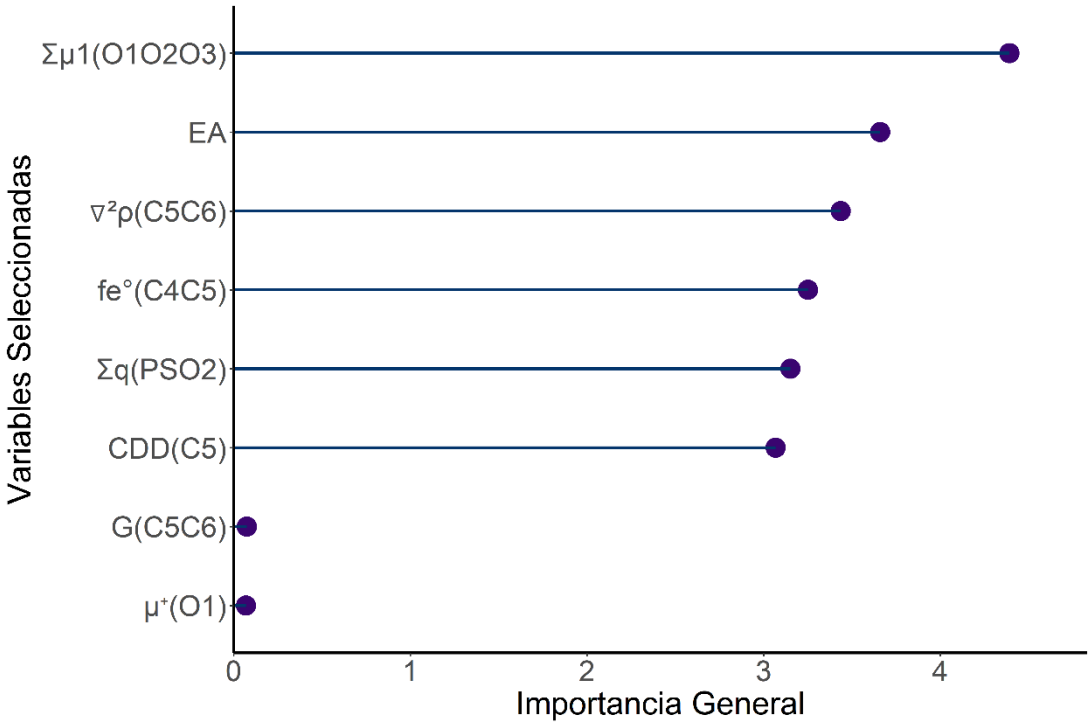

Figura 33 Importancia de las variables óptimas para la segunda penalización LASSO al mejor valor de λ.

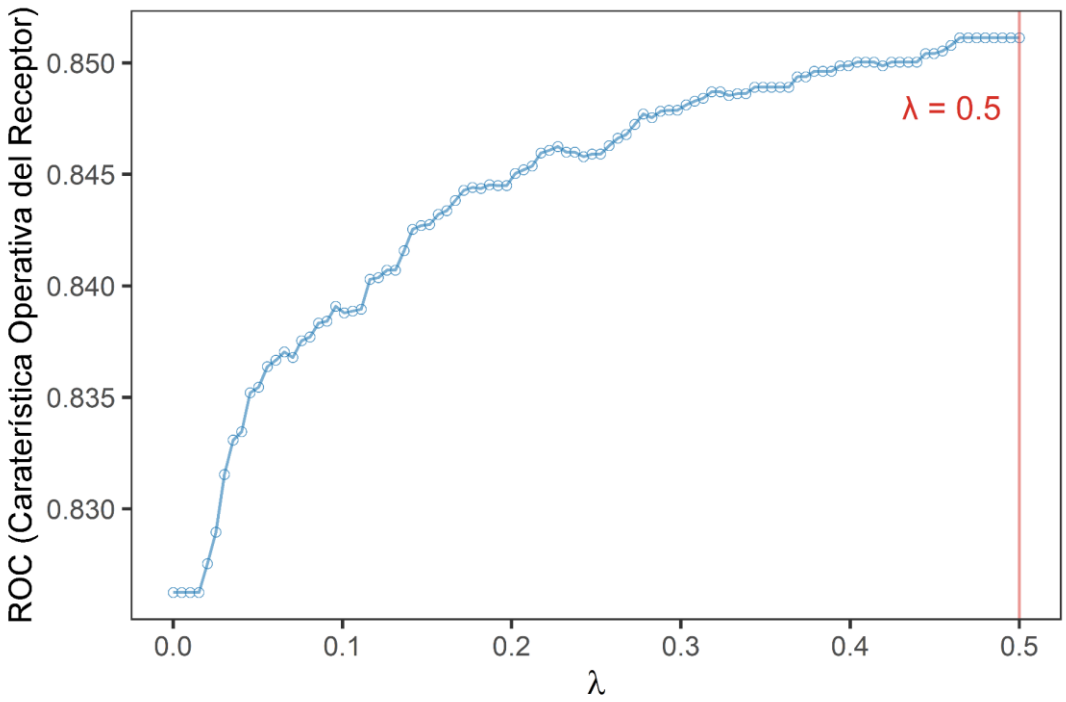

Figura 34 Gráfico de ROC vs λ para el modelo RIDGE.

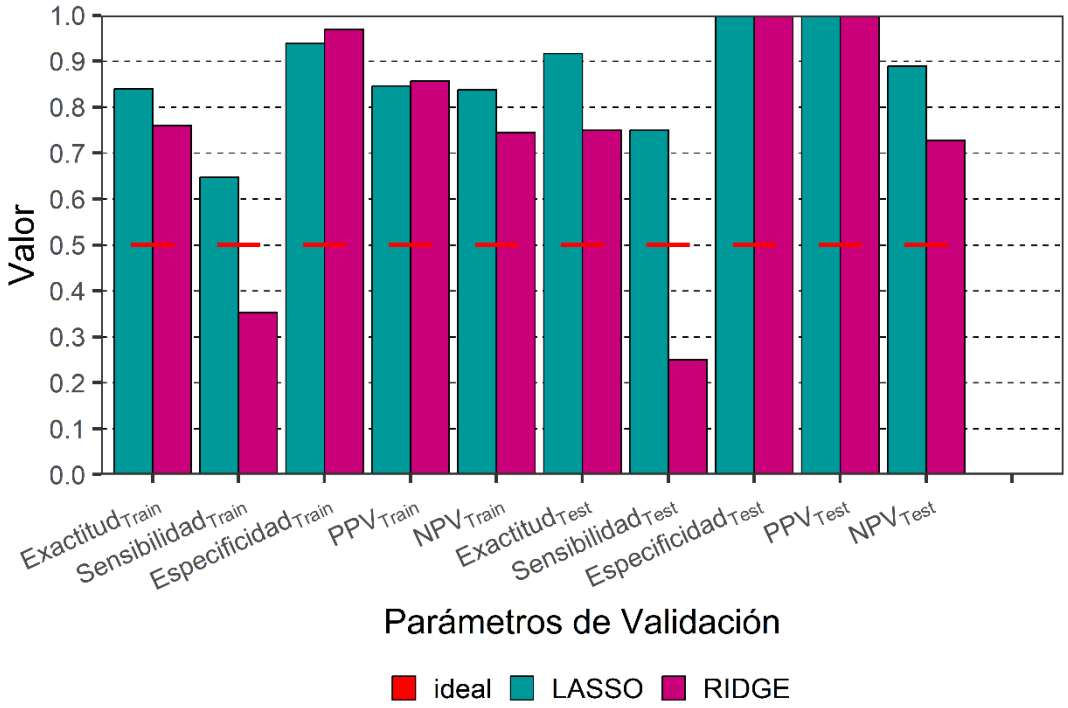

Figura 35 Parámetros de validación para los modelos de clasificación QSTR lineales, elaborados con las metodologías LASSO y RIDGE. En rojo se presentan las características mínimas con las que un modelo se considera estadísticamente válido

Las 8 variables involucradas en el modelo LASSO, el cual presentó el mejor rendimiento, son: el laplaciano de la densidad del enlace entre C_5-C_6 ($\boldsymbol{\nabla^2\rho_{(C_5-C_6)}}$), la energía cinética del enlace C_5-C_6 ($\boldsymbol{G_{(C_5-C_6)}}$), la afinidad electrónica ($\boldsymbol{EA}$), el descriptor dual de C_5 ($\boldsymbol{CDD_{(C_5)}}$), la Fukui kernel para ω=0 del enlace C_4-C_5 ($\boldsymbol{fe^{\circ}_{(C_4-C_5)}}$), la dureza local para ω>0 de O_1 ($\boldsymbol{\mu^{+}_{(O_1)}}$), la sumatoria del primer eigenvalor del momento cuadrupolar para los oxígenos unidos al fósforo ($\sum \boldsymbol{\mu 1_{(O_1,O_2,O_3)}}$) y la sumatoria de la carga para el fragmento P, S, O_2 ($\sum \boldsymbol{q_{(P,S,O_2)}}$). En la Figura 36 se muestra una comparativa gráfica de los coeficientes obtenidos en los modelos de clasificación.

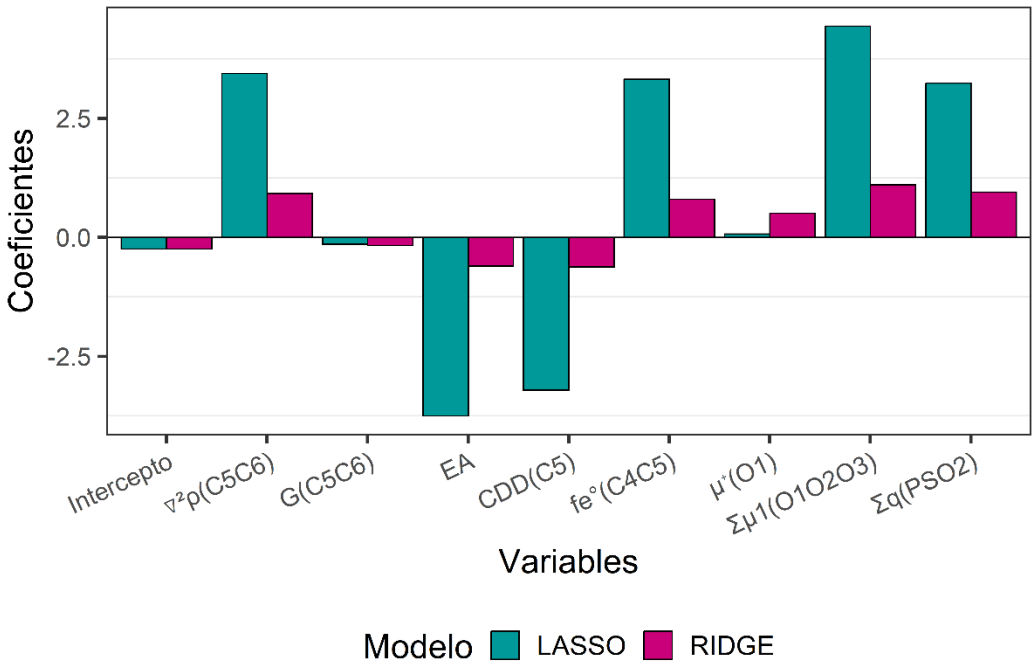

Figura 36 Comparación de los coeficientes de las variables obtenidas en los modelos de clasificación.

El modelo de clasificación LASSO sigue la siguiente ecuación, donde un valor $Y \leq 0$ tiene una mayor probabilidad de ser asignado al grupo HT y $Y > 0$ a MT:

$$Y = -0.2416 + 3.4437\boldsymbol{\nabla^2\rho}_{(\boldsymbol{C_5-C_6})} - 0.1395\boldsymbol{G}_{(\boldsymbol{C_5-C_6})} - 3.7487\boldsymbol{EA} - 3.2029\boldsymbol{CDD}_{(\boldsymbol{C_5})} + 3.3186\boldsymbol{fe}^{\circ}_{(\boldsymbol{C_4-C_5})} + 0.0688\boldsymbol{\mu}^{+}_{(\boldsymbol{O_1})} + 4.4363\sum \boldsymbol{\mu 1}_{(\boldsymbol{O_1,O_2,O_3})} + 3.2419\sum \boldsymbol{q}_{(\boldsymbol{P,S,O_2})}$$

El modelo propone que la densidad del enlace C_5-C_6 está relacionada con la reactividad en esa región de la molécula. La afinidad electrónica es un indicador de la capacidad de un compuesto para interactuar con otros compuestos, mientras que el descriptor dual de C_5 indica la capacidad del átomo para aceptar o donar electrones. La Fukui kernel para C_4-C_5 refleja la capacidad del átomo de C_5 para aceptar o donar electrones a otros átomos en la molécula, y la dureza local de O_1 está relacionada con la estabilidad del compuesto. La suma del primer eigenvalor del momento cuadrupolar para O_1, O_2 y O_3 indica la capacidad de la molécula para interactuar con otros

compuestos, mientras que la suma de la carga de P, S y O_2 está relacionada con la reactividad. Se destacan las zonas de mayor interés en la región alrededor del enlace C_5-C_6 y la región cercana al átomo de fósforo y los átomos de oxígeno (O_2 y O_3). El valor constante (–0.2416) representa el valor promedio de la probabilidad de clasificación para un conjunto de datos en el que todas las variables son cero.

Las variables con coeficientes positivos en la ecuación indican que valores altos de estas variables están relacionados con una mayor probabilidad de clasificación en la clase HT. Por otro lado, las variables con coeficientes negativos en la ecuación señalan que valores altos de estas variables están relacionados con una menor probabilidad de clasificación en la clase HT, lo que implica una mayor probabilidad de clasificación en la clase MT.

3.3.3 MODELOS QSTR DE REGRESIÓN NO LINEAL

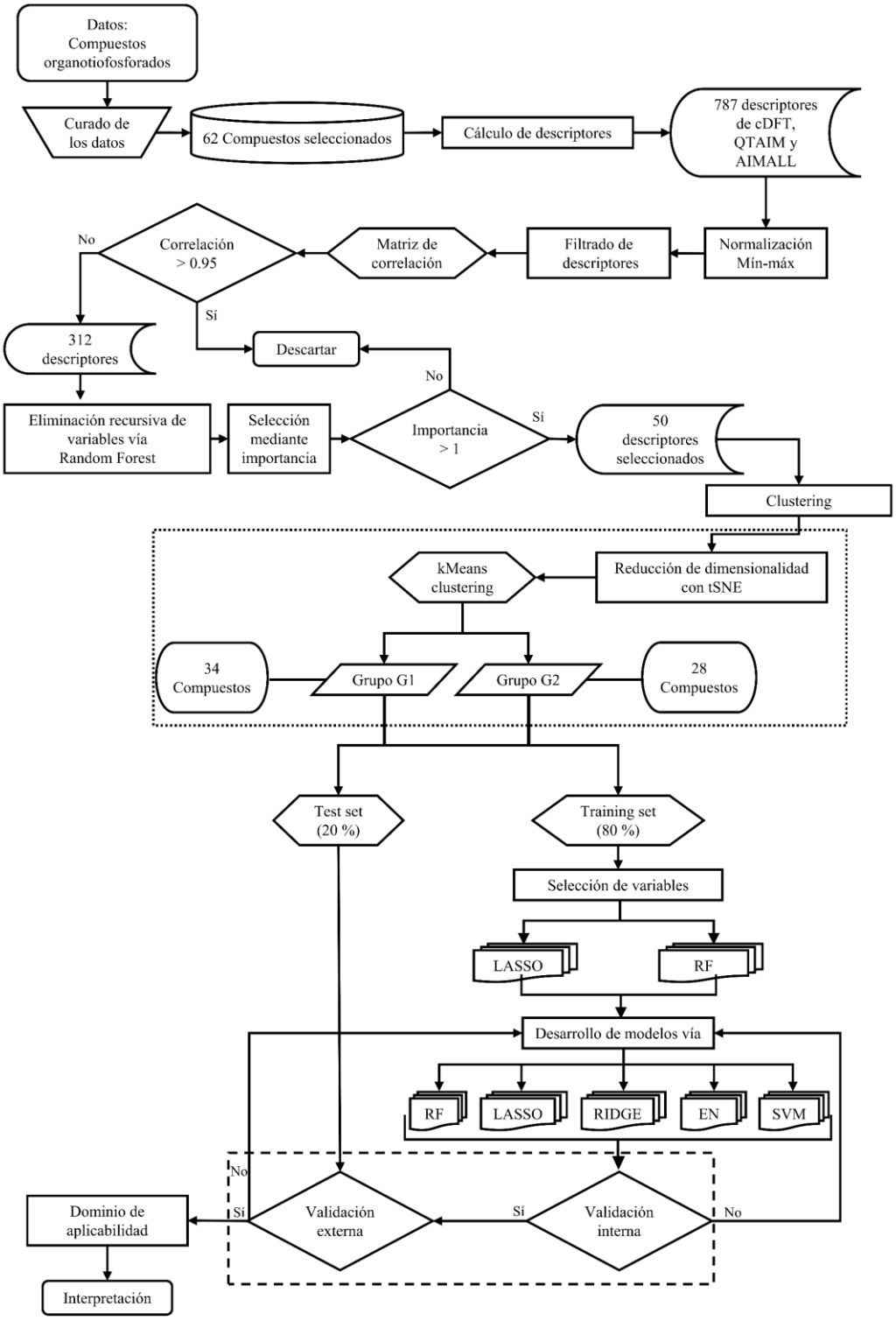

Figura 37 Diagrama de flujo para el desarrollo de los modelos.

En la Figura 37 se muestra de manera general la metodología seguida. Los datos de toxicidad LD_{50} de los 62 compuestos en mmol/kg se transformaron al logaritmo de la toxicidad. Posteriormente, se normalizaron los descriptores con el algoritmo mín-máx y se realizó un filtrado utilizando una matriz de correlación, eliminándose todos aquellos descriptores con una correlación mayor a 0.95. Como resultado de este proceso, se obtuvieron 312 descriptores.

Para abordar el desafío de lidiar con un número de variables mucho mayor que el número de observaciones, se empleó la eliminación recursiva de variables utilizando el algoritmo Random Forest. Este enfoque se llevó a cabo utilizando una validación cruzada con 5-folds y 10 repeticiones, seleccionando las variables con un puntaje overall mayor a 1. Como resultado, se obtuvieron 50 variables relevantes para el análisis.

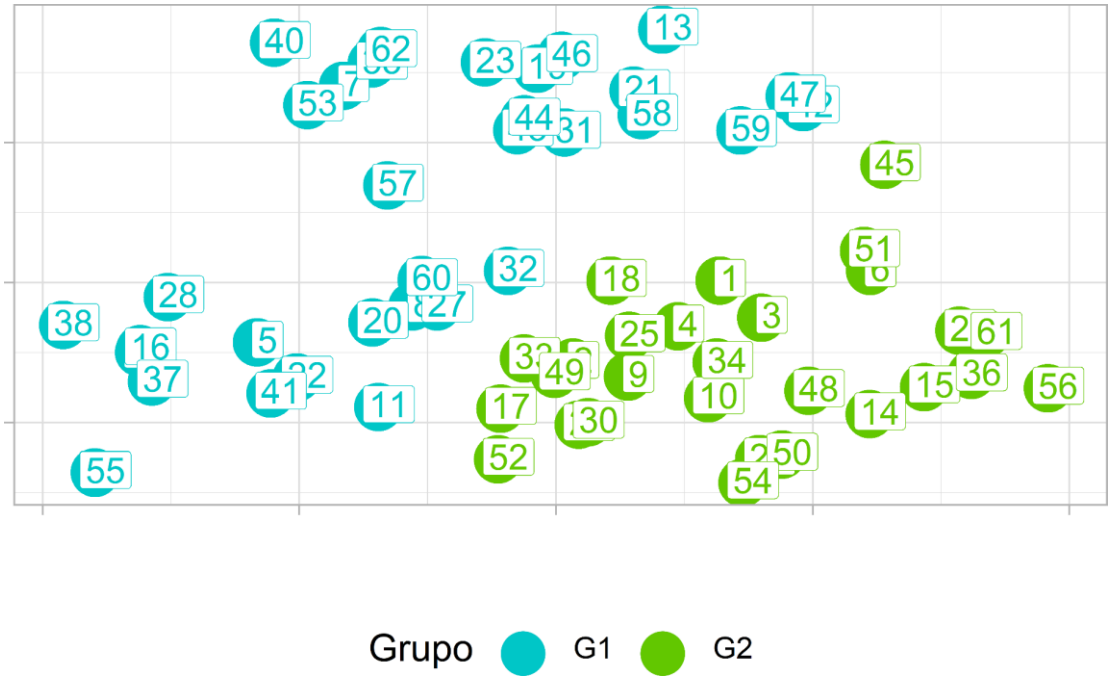

Figura 38 Gráfico de tSNE para los 62 compuestos organotiofosforados.

La nueva matriz de datos, compuesta por los 62 compuestos y las 50 variables seleccionadas, se sometió a una reducción de dimensionalidad mediante la técnica de

t-SNE (t-Distributed Stochastic Neighbor Embedding). Se utilizaron parámetros de perplejidad de 20, theta de 0.1 y 10000 iteraciones para obtener una representación visual de los compuestos en un espacio de menor dimensión, como se muestra en la Figura 38. Se realizó una clasificación de los compuestos en dos grupos, utilizando un número máximo de agrupaciones (k) igual a 2, validado mediante el coeficiente de silueta, que se muestra en la Figura 39. En la Tabla 4 se presentan los resultados de la clasificación, con 34 compuestos asignados al grupo 1 (G1) y 28 compuestos asignados al grupo 2 (G2).

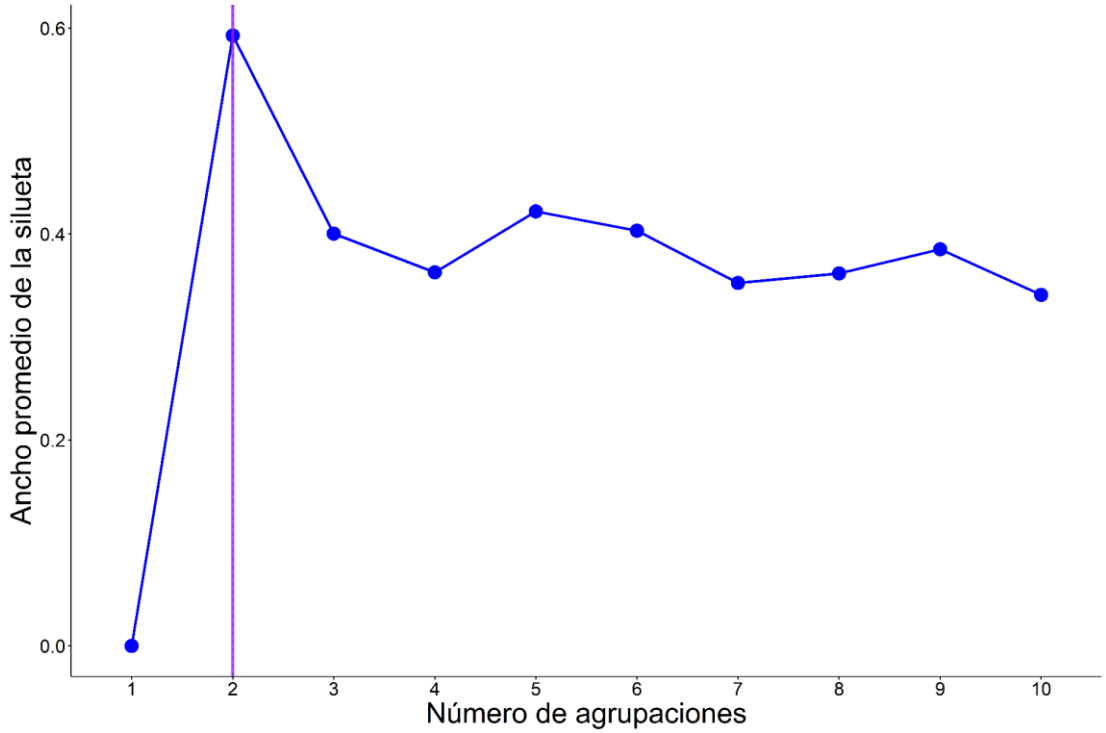

Figura 39 Coeficiente de silueta para las agrupaciones producto de tSNE utilizando kmeans.

Tabla 4 ID y grupo al que pertenecen los organotiofosforados estudiados producto del tSNE.

ID	Grupo	ID	Grupo	ID	Grupo	ID	Grupo
01	G2	**17**	G2	**33**	G2	**49**	G2
02	G2	**18**	G2	**34**	G2	**50**	G2
03	G2	**19**	G1	**35**	G1	**51**	G2
04	G2	**20**	G1	**36**	G2	**52**	G2
05	G1	**21**	G1	**37**	G1	**53**	G1
06	G2	**22**	G1	**38**	G1	**54**	G2
07	G1	**23**	G1	**39**	G1	**55**	G1
08	G1	**24**	G2	**40**	G1	**56**	G2
09	G2	**25**	G2	**41**	G1	**57**	G1
10	G2	**26**	G2	**42**	G1	**58**	G1
11	G1	**27**	G1	**43**	G1	**59**	G1
12	G1	**28**	G1	**44**	G1	**60**	G1
13	G1	**29**	G2	**45**	G2	**61**	G2
14	G2	**30**	G2	**46**	G1	**62**	G1
15	G2	**31**	G1	**47**	G1		
16	G1	**32**	G1	**48**	G2		

La generación de modelos QSTR se llevó a cabo sobre los grupos obtenidos para relacionar las propiedades locales y globales seleccionadas con los valores experimentales de LD_{50}, utilizando los algoritmos RF, LASSO, Ridge, Elastic Net y SVM. Para cada grupo se propuso la selección de variables de manera independiente a través de los métodos de RF y LASSO, las variables obtenidas mediante cada método se emplearon como punto de partida para alimentar los siguientes algoritmos de aprendizaje automático. Todos los modelos utilizaron el 80 % de los elementos en el conjunto de entrenamiento y el 20 % en el conjunto de prueba, realizando una validación cruzada con 10 folds y 10 repeticiones, así como los hiperparámetros mostrados en la Tabla 5.

Tabla 5 Hiperparámetros optimizados empleados para los modelos generados.

Modelo	Algoritmo	G1	G2
A	**RF**	ntree = 5000	ntree = 5000
	↳ LASSO	λ = 0.042	λ = 0.059
	↳ Ridge	λ = 0.075	λ = 1.186
	↳ Elastic Net	α = 0, λ = 0.126	α = 0, λ = 0.021
	↳ SVM	C = 0.105	C = 0.316
B	**LASSO**	λ = 0.056	λ = 0.064
	↳ RF	ntree = 5000	ntree = 5000
	↳ Ridge	λ = 0.441	λ = 101.414
	↳ Elastic Net	α = 0, λ = 0.042	α = 0, λ = 0.147
	↳ SVM	C = 0.105	C = 0.053

En las Figuras 40 y 41 se muestra que los parámetros de validación para G1 indican una fuerte correlación entre las variables independientes y la variable dependiente, con valores de R^2_{Train} de 0.943 y R^2_{Test} de 0.991 obtenidos para el modelo A-RF. El gráfico de regresión lineal para este modelo se muestra en la Figura 42. Para el modelo A-RF en G1 se utilizaron las siguientes cinco variables cuánticas: 1) el Laplaciano de la energía cinética del enlace C_1-C_6 ($\boldsymbol{\nabla^2 G_{(C_1-C_6)}}$); 2) el segundo eigenvalor del tensor del momento cuadrupolar para el átomo R_2 ($\boldsymbol{\mu 2_{(R_2)}}$); 3) la Fukui kernel para el enlace C_3-C_4 ($\boldsymbol{f_e{}^{\circ}{}_{(C_3-C_4)}}$); 4) la función de Fukui electrofílica para el sustituyente R_1 ($\boldsymbol{f^{+}_{(R_1)}}$) y 5) el primer eigenvalor del momento cuadrupolar para O_2 ($\boldsymbol{\mu 1_{(O_2)}}$).

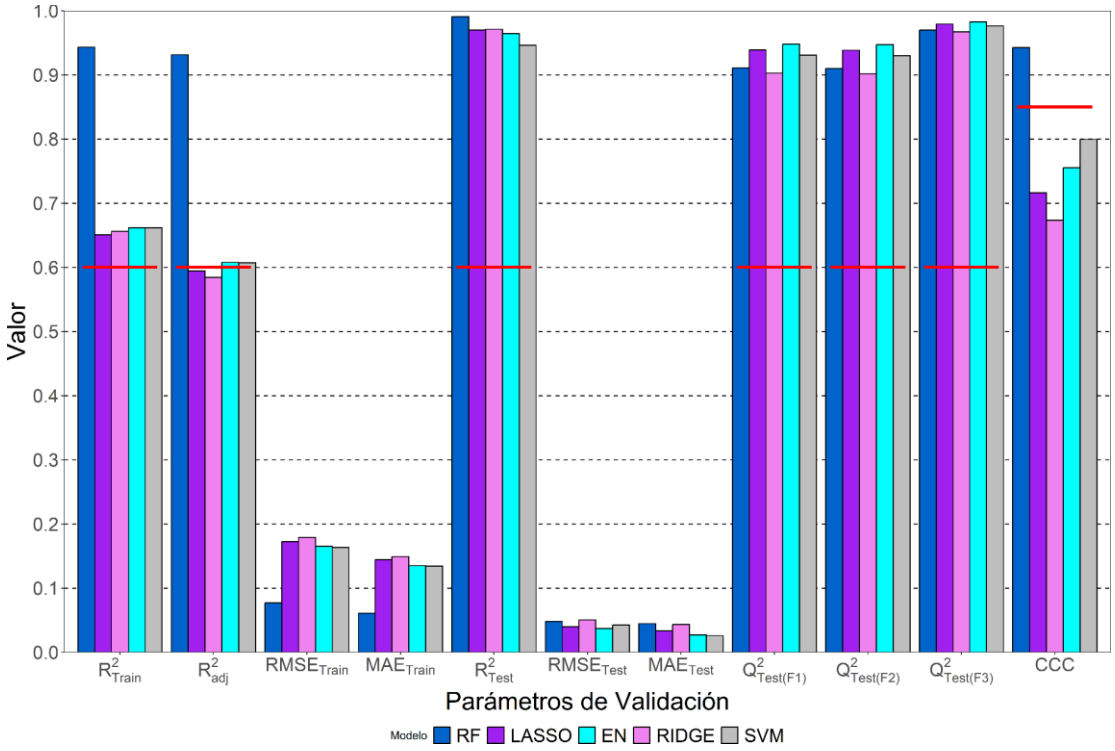

Figura 40 Parámetros de validación para los modelos A del grupo G1. La línea roja representa el valor mínimo necesario para considerar a un modelo como estadísticamente válido.

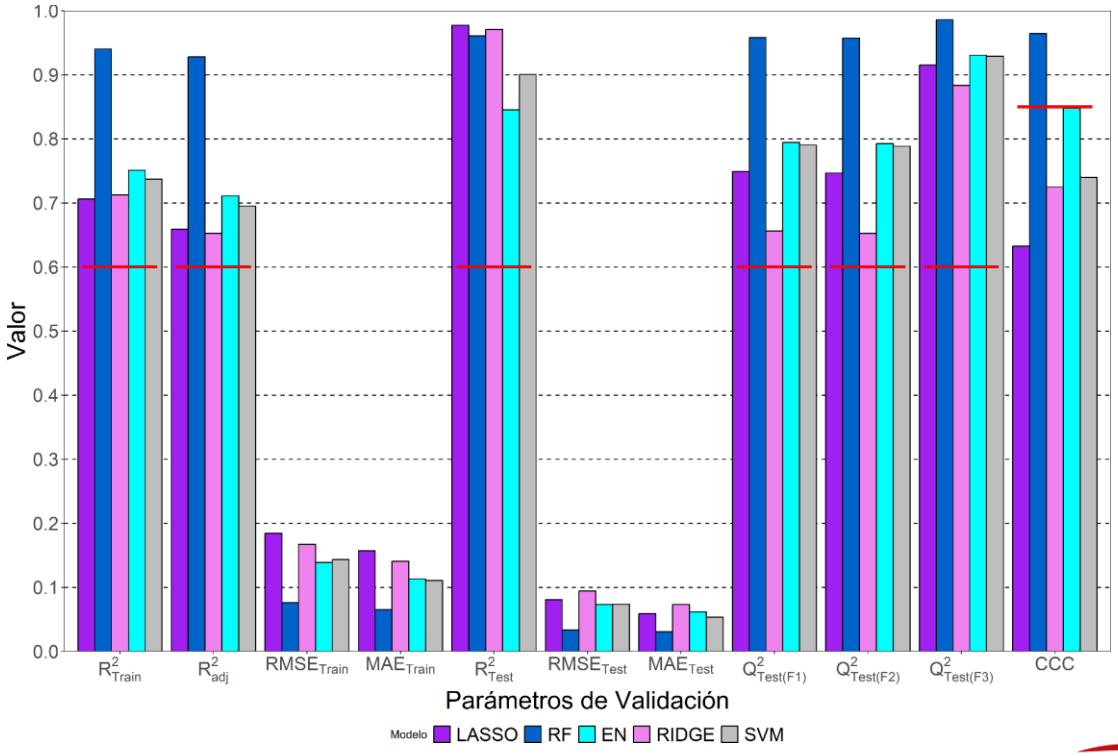

Figura 41 Parámetros de validación para los modelos B del grupo G1. La línea roja representa el valor mínimo necesario para considerar a un modelo como estadísticamente válido.

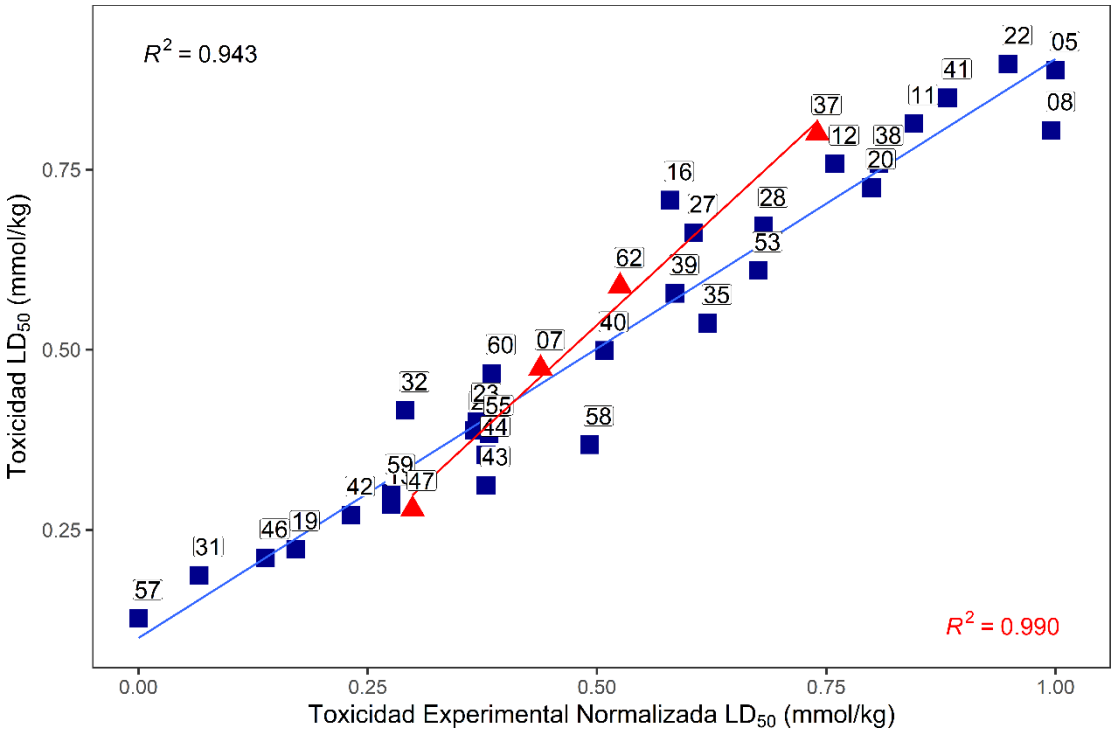

Figura 42 Gráfico de dispersión para el modelo A-RF para G1.

De igual manera, para el grupo G2 se muestran los parámetros de validación para los modelos en las Figuras 43 y 44, donde el modelo A-RF presenta valores de R^2_{Train} de 0.876 y R^2_{Test} de 0.967, como se puede observar en la Figura 45. En este modelo se presentaron los siguientes cinco descriptores: 1) el primer eigenvalor del momento cuadrupolar del O_1 ($\boldsymbol{\mu 1}_{(O_1)}$); 2) el laplaciano de la densidad electrónica para el enlace O_1-R_1 ($\boldsymbol{\nabla^2 \rho}_{(O_1-R_1)}$); 3) el primer eigenvalor del momento cuadrupolar del átomo O_2 ($\boldsymbol{\mu 1}_{(O_2)}$); 4) la carga de Hishfled para el carbono C_3 ($\mathbf{q}^{\mathbf{N+1}}_{(C_3)}$) y 5) el laplaciano de la densidad electrónica del enlace C_5-C_6 ($\boldsymbol{\nabla^2 \rho}_{(C_5-C_6)}$).

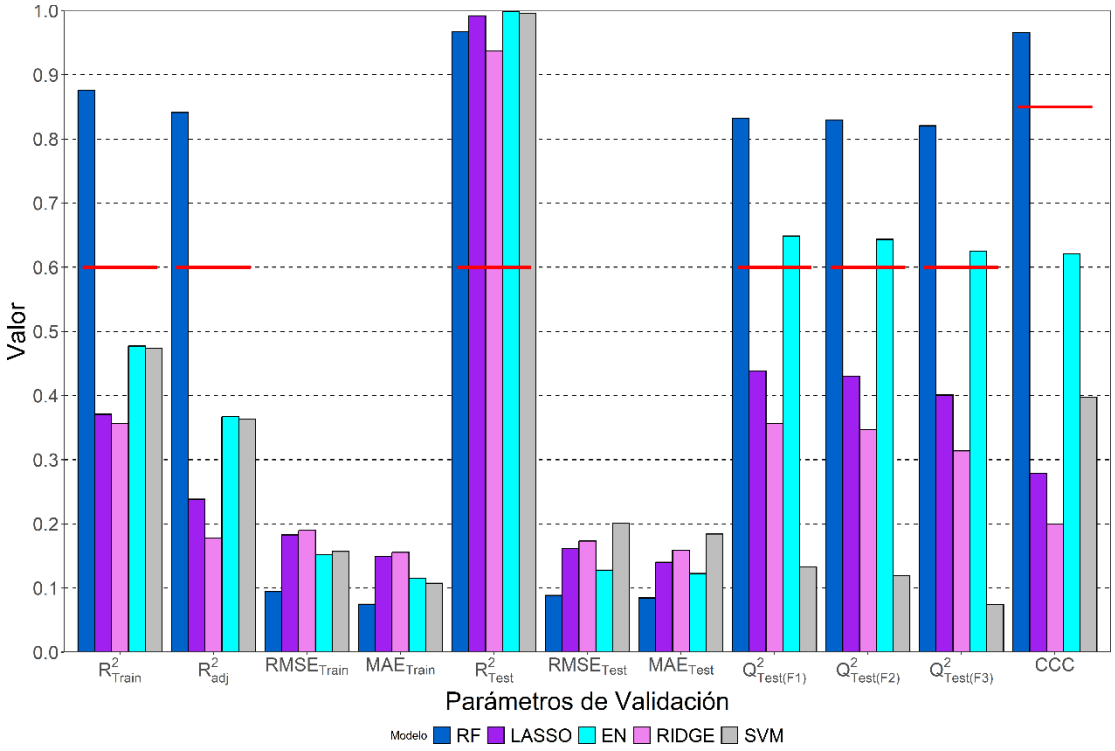

Figura 43 Parámetros de validación para los modelos A del grupo G2. La línea roja representa el valor mínimo necesario para considerar a un modelo como estadísticamente válido.

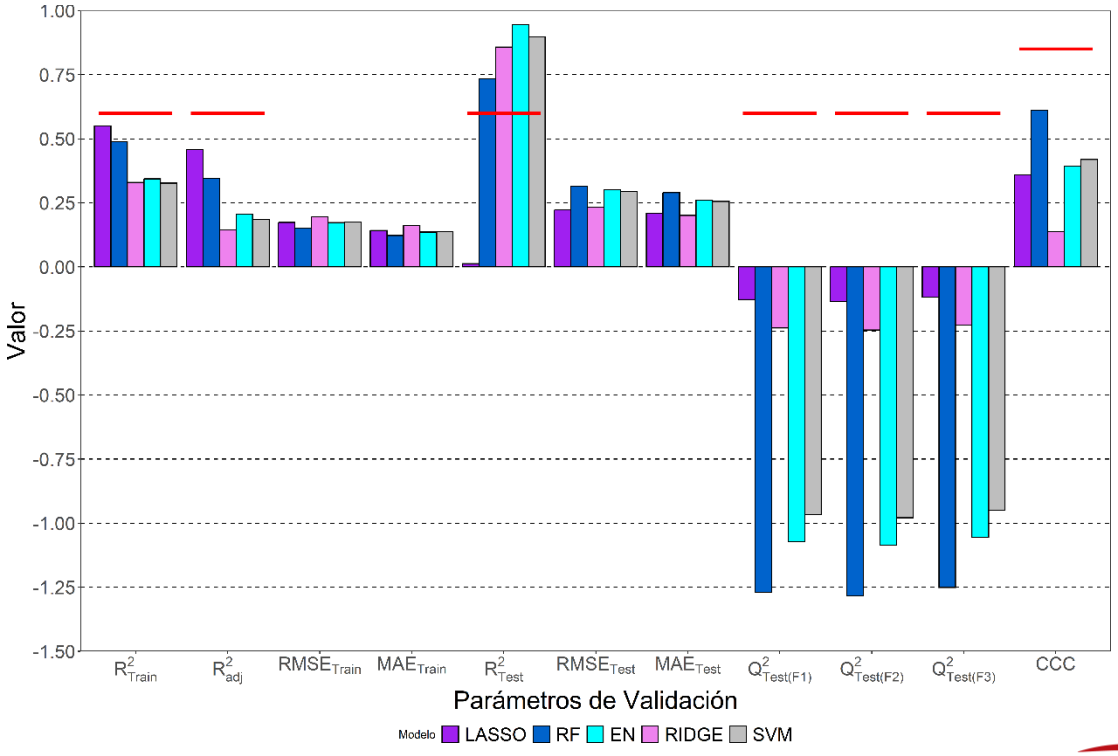

Figura 44 Parámetros de validación para los modelos B del grupo G2. La línea roja representa el valor mínimo necesario para considerar a un modelo como estadísticamente válido.

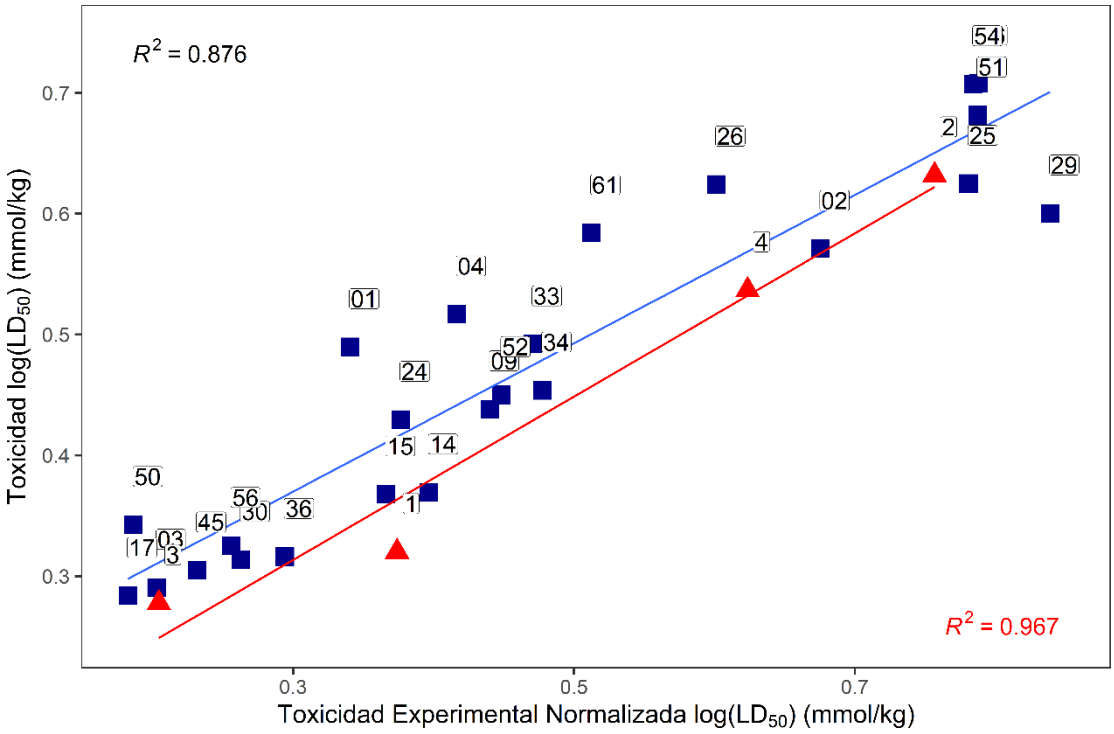

Figura 45 Gráfico de dispersión para el modelo A-RF para G2.

Los modelos obtenidos en ambas agrupaciones presentan variables cuánticas que describen los sitios de interés tanto en la zona del anillo aromático como en los sustituyentes R_1, R_2 y los átomos de oxígeno vecinos al enlace P=S. De hecho, las variables indican una partición significativa de los sustituyentes en el anillo aromático, así como de los grupos R_1 y R_2, lo que sugiere una participación importante en posibles mecanismos biológicos. De las Figuras 46 a 49, se muestran a manera de resumen gráfico la importancia relativa de las variables involucradas en cada uno de los modelos obtenidos.

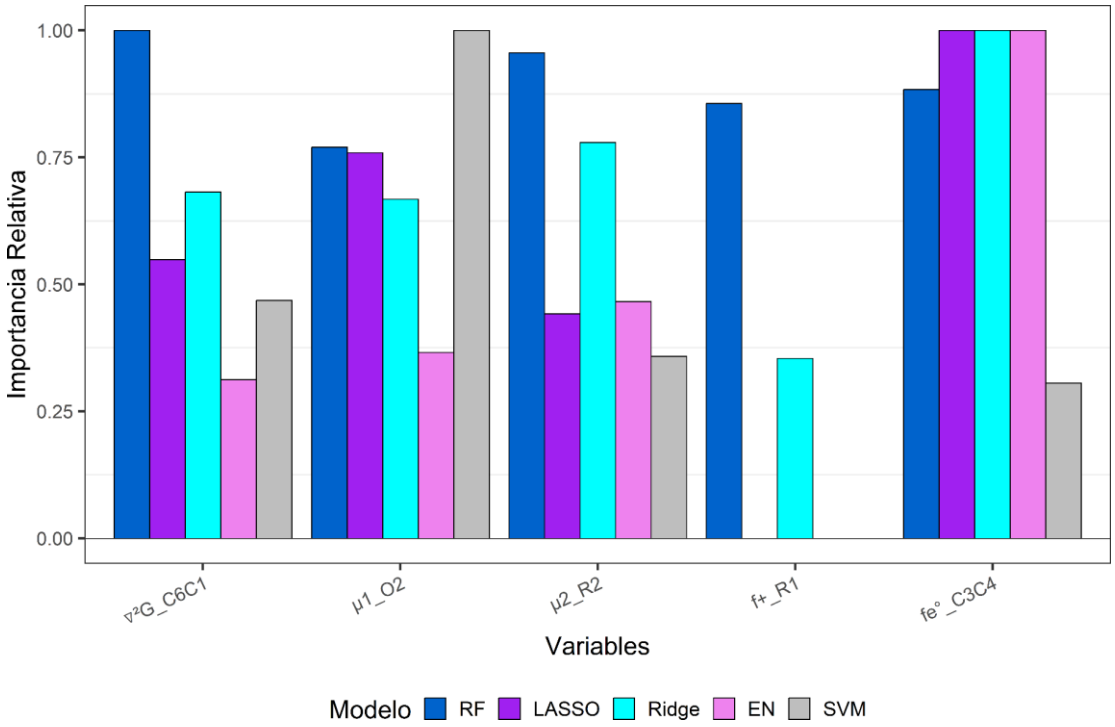

Figura 46 Importancia relativa de las variables presentes en los modelos A del grupo G1.

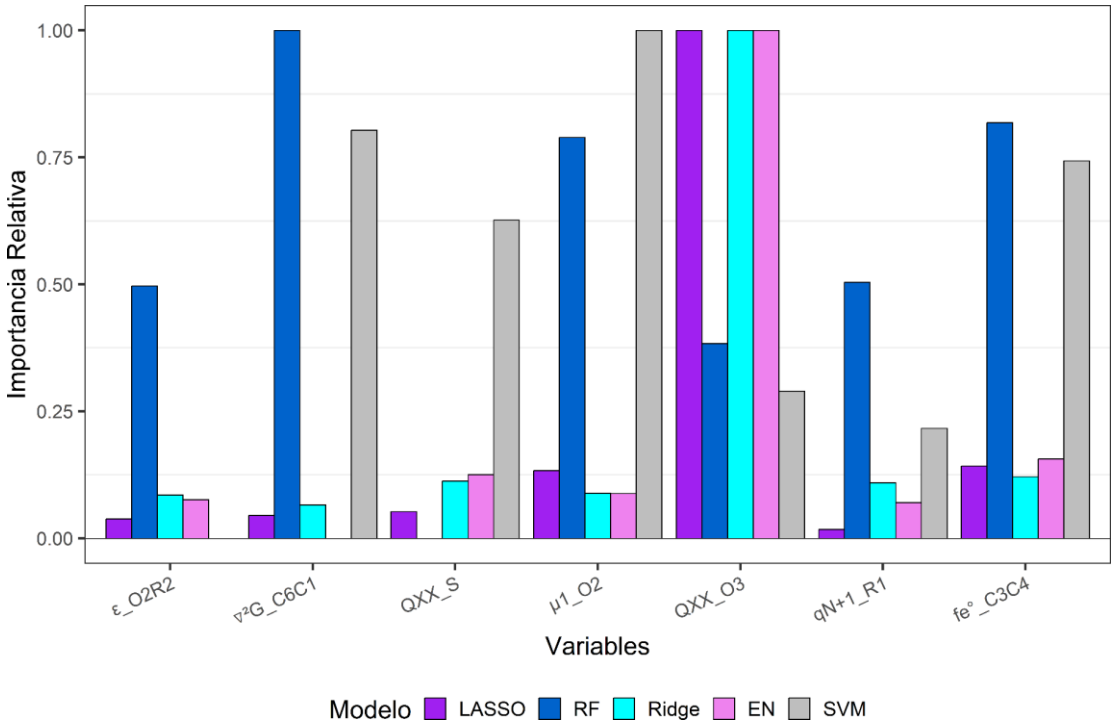

Figura 47 Importancia relativa de las variables presentes en los modelos B del grupo G1.

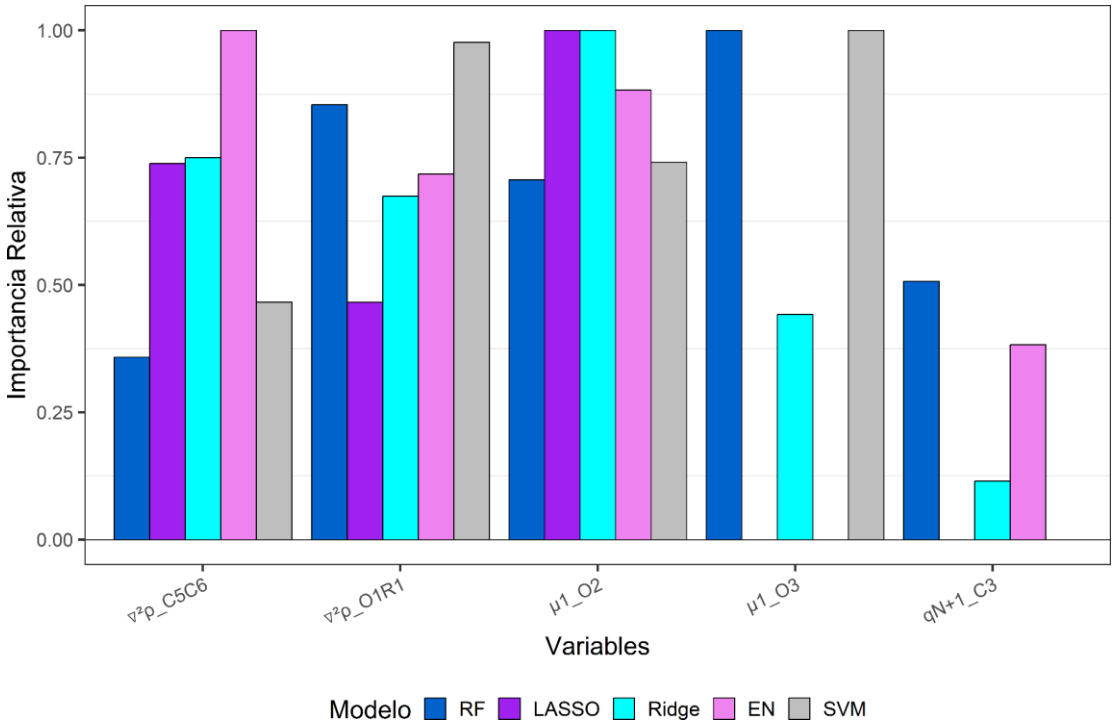

Figura 48 Importancia relativa de las variables presentes en los modelos A del grupo G2.

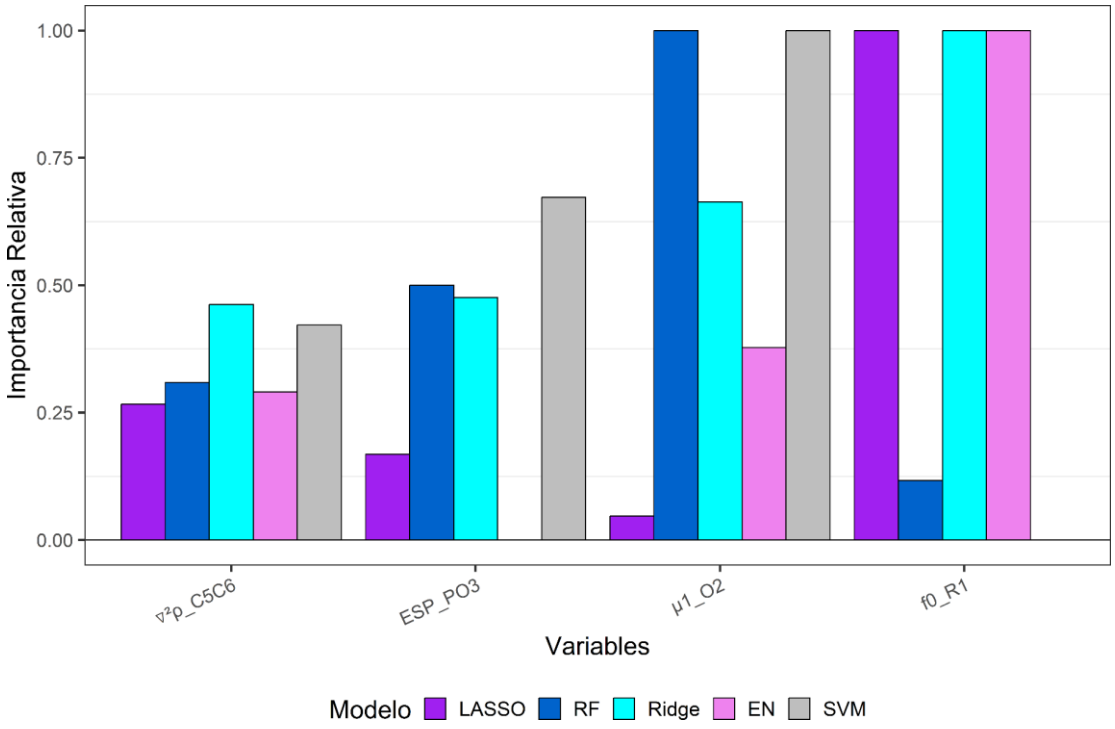

Figura 49 Importancia relativa de las variables presentes en los modelos B del grupo G2.

4 DISCUSIÓN

Mediante el empleo de descriptores de reactividad obtenidos a partir de la teoría de los funcionales de la densidad conceptual y de la teoría cuántica de átomos en moléculas se logró identificar sitios de interés para el conjunto de compuestos organotiofosforados estudiados. Además, se descubrieron propiedades específicas que se correlacionan con la toxicidad de dichas moléculas, proporcionando información crucial sobre su reactividad. Estos enfoques basados en inteligencia artificial permitieron la selección de las variables de interés mediante un análisis exhaustivo y detallado de las características químicas y estructurales que influyen en la toxicidad. En conjunto, estos hallazgos contribuyen a un mejor entendimiento de los compuestos estudiados y sientan las bases para el desarrollo de estrategias más eficientes en el diseño de compuestos menos tóxicos y más seguros.

Se identificaron variables importantes en los modelos de regresión y clasificación lineales que se relacionan con la toxicidad de los compuestos. En el modelo de regresión, las variables $\boldsymbol{\nabla^2 V_{rep_{(O_2-R_2)}}}$, $\boldsymbol{Q_{YY_{(C_3)}}}$ y $\boldsymbol{\mu 2_{(C_3)}}$ se asociaron positivamente con la disminución de la toxicidad, mientras que las variables $\boldsymbol{L_{(C_5-C_6)}}$ y $\boldsymbol{\sum \eta^{\tau\circ}_{(P,S,O_3)}}$ y el intercepto se relacionaron negativamente con la toxicidad. En el modelo de clasificación, la densidad del enlace C_5-C_6, la afinidad electrónica, el descriptor dual de C_5, la Fukui kernel para C_4-C_5, la dureza local de O_1, la suma del primer eigenvalor del momento cuadrupolar para O_1, O_2 y O_3, y la suma de la carga de P, S y O_2 se identificaron como variables importantes.

Estos resultados sugieren que la distribución de carga eléctrica en la molécula y la orientación relativa de los átomos son importantes para la reactividad y la toxicidad. Además, se destacan las regiones alrededor del enlace C_5-C_6, del átomo de fósforo y los átomos de oxígeno (O_2 y O_3) como zonas de mayor interés.

Se desarrollaron dos modelos de regresión no lineal, A-RF-G1 y A-RF-G2, basados en descriptores cuánticos utilizando cDFT y QTAIM, con el objetivo de descubrir qué parámetros electrónicos están asociados a la toxicidad a través de los valores de LD_{50} de sesenta y dos organotiofosfatos a través de una relación no lineal. Ambos modelos utilizan RF como algoritmo para desarrollar los modelos y presentan un valor estadístico aceptable de $R^2 > 0.943$. Esto indica que se pueden utilizar para determinar la toxicidad de los compuestos organotiofosforados.

Los descriptores más relevantes obtenidos para cada modelo principal son $\boldsymbol{\nabla^2 G_{(C_6-C_1)}}$, $\boldsymbol{\mu 2_{(R_2)}}$, $\boldsymbol{fe^{\circ}_{(C_3-C_4)}}$, $\boldsymbol{f^{+}_{(R_1)}}$, $\boldsymbol{\mu 2_{(O_2)}}$ y $\boldsymbol{\mu 1_{(O_3)}}$, $\boldsymbol{\nabla^2_{(O1-R1)}}$, $\boldsymbol{\mu 2_{(O_2)}}$, $\boldsymbol{q^{N+1}_{(C_3)}}$, $\boldsymbol{\nabla^2_{(C_5-C_6)}}$. Estos descriptores incluyen principalmente propiedades relacionadas con la densidad electrónica, la carga de Hirshfeld, los momentos cuadrupolares de átomos y enlaces alrededor del grupo principal. Sin embargo, se observa la ausencia directa del azufre en estos descriptores. Esto sugiere que las propiedades electrónicas del azufre no tienen una influencia fuerte en la toxicidad de los compuestos estudiados.

5 CONCLUSIONES

En este trabajo se llevó a cabo una investigación exhaustiva en el campo de la toxicidad de compuestos organotiofosforados, aplicando una combinación de enfoques teóricos y técnicas de inteligencia artificial. Los objetivos específicos establecidos al inicio de la investigación han sido plenamente cumplidos, y los resultados obtenidos han proporcionado un valioso entendimiento sobre los factores determinantes en la reactividad y toxicidad de estos compuestos.

Mediante la exploración de acervos bibliográficos y revistas especializadas, se recopilaron datos relevantes sobre toxicidad LD_{50} y parámetros fisicoquímicos de compuestos los compuestos estudiados. Esta base de datos se convirtió en el fundamento para la selección y optimización de la geometría de las moléculas de

estudio mediante un nivel de teoría ωB97XD/6-311++G**, permitiendo una representación precisa de los sistemas de estudio.

La aplicación de la teoría de los funcionales de la densidad conceptual (cDFT) y la teoría cuántica de átomos en moléculas (QTAIM) ha arrojado resultados significativos. Los descriptores de reactividad obtenidos a partir de estas teorías no solo identificaron sitios de interés dentro de las moléculas, sino también propiedades específicas que están correlacionadas con la toxicidad. Estas propiedades proporcionan información crucial para comprender la reactividad de los compuestos estudiados y su posible impacto en la toxicidad.

La implementación de técnicas de inteligencia artificial ha permitido no solo relacionar descriptores globales, locales y de QTAIM con los valores de toxicidad, sino también desarrollar modelos matemáticos predictivos. Estos modelos, basados en una relación cuantitativa entre la estructura y la toxicidad, presentan un alto grado de correlación. Los resultados de estos modelos sugieren que la distribución de carga eléctrica en la molécula y la orientación relativa de los átomos son aspectos cruciales para la reactividad y toxicidad.

En resumen, este estudio ha logrado un enfoque integral que abarca desde la recopilación de datos hasta la implementación de técnicas teóricas y de inteligencia artificial. Los hallazgos obtenidos no solo amplían nuestro conocimiento sobre los compuestos organotiofosforados y su toxicidad, sino que también establecen una base sólida para la creación de estrategias más efectivas en el diseño de compuestos con menor toxicidad y mayor seguridad en futuras investigaciones y aplicaciones.

6 REFERENCIAS

[1] Wang H, Hu B, Gao Z, Zhang F, Wang J. Emerging role of graphene oxide as sorbent for pesticides adsorption: Experimental observations analyzed by molecular modeling. Journal of Materials Science & Technology 2021;63:192–202. https://doi.org/10.1016/j.jmst.2020.02.033.

[2] Butkovskyi A, Jing Y, Bergheim H, Lazar D, Gulyaeva K, Odenmarck SR, et al. Retention and distribution of pesticides in planted filter microcosms designed for treatment of agricultural surface runoff. Science of The Total Environment 2021;778:146114. https://doi.org/10.1016/j.scitotenv.2021.146114.

[3] Yang L, Wang Y, Hao W, Chang J, Pan Y, Li J, et al. Modeling pesticides toxicity to Sheepshead minnow using QSAR. Ecotoxicology and Environmental Safety 2020;193:110352. https://doi.org/10.1016/j.ecoenv.2020.110352.

[4] Yang L, Wang Y, Chang J, Pan Y, Wei R, Li J, et al. QSAR modeling the toxicity of pesticides against Americamysis bahia. Chemosphere 2020;258:127217. https://doi.org/10.1016/j.chemosphere.2020.127217.

[5] Manjarres-López DP, Andrades MS, Sánchez-González S, Rodríguez-Cruz MS, Sánchez-Martín MJ, Herrero-Hernández E. Assessment of pesticide residues in waters and soils of a vineyard region and its temporal evolution. Environmental Pollution 2021;284:117463. https://doi.org/10.1016/j.envpol.2021.117463.

[6] Zdravković M, Antović A, Veselinović JB, Sokolović D, Veselinović AM. QSPR in forensic analysis – The prediction of retention time of pesticide residues based on the Monte Carlo method. Talanta 2018;178:656–62. https://doi.org/10.1016/j.talanta.2017.09.064.

[7] Li Y, Miao R, Khanna M. Neonicotinoids and decline in bird biodiversity in the United States. Nat Sustain 2020;3:1027–35. https://doi.org/10.1038/s41893-020-0582-x.

[8] Hassaan MA, El Nemr A. Pesticides pollution: Classifications, human health impact, extraction and treatment techniques. The Egyptian Journal of Aquatic Research 2020;46:207–20. https://doi.org/10.1016/j.ejar.2020.08.007.

[9] Organización Mundial de la Salud, Programa Internacional de Seguridad de las Sustancias Químicas. Clasificación recomendada por la OMS de los plaguicidas por el peligro que presentan y directrices para la clasificación 2019. Ginebra: Organización Mundial de la Salud; 2020.

[10] Park E, Lee J, Lee J, Lee J, Lee HS, Shin Y, et al. Method for the simultaneous analysis of 300 pesticide residues in hair by LC-MS/MS and GC-MS/MS, and its

application to biomonitoring of agricultural workers. Chemosphere 2021;277:130215. https://doi.org/10.1016/j.chemosphere.2021.130215.

[11] Rahman M, Hoque MdS, Bhowmik S, Ferdousi S, Kabiraz MP, van Brakel ML. Monitoring of pesticide residues from fish feed, fish and vegetables in Bangladesh by GC-MS using the QuEChERS method. Heliyon 2021;7:e06390. https://doi.org/10.1016/j.heliyon.2021.e06390.

[12] Ko E, Choi M, Shin S. Bottom-line mechanism of organochlorine pesticides on mitochondria dysfunction linked with type 2 diabetes. Journal of Hazardous Materials 2020;393:122400. https://doi.org/10.1016/j.jhazmat.2020.122400.

[13] Hernández AF, Bennekou SH, Hart A, Mohimont L, Wolterink G. Mechanisms underlying disruptive effects of pesticides on the thyroid function. Current Opinion in Toxicology 2020;19:34–41. https://doi.org/10.1016/j.cotox.2019.10.003.

[14] Darwiche W, Delanaud S, Dupont S, Ghamlouch H, Ramadan W, Joumaa W, et al. Impact of prenatal and postnatal exposure to the pesticide chlorpyrifos on the contraction of rat ileal muscle strips: involvement of an inducible nitric oxide synthase-dependent pathway. Neurogastroenterol Motil 2017;29:e12918. https://doi.org/10.1111/nmo.12918.

[15] Leung MCK, Meyer JN. Mitochondria as a target of organophosphate and carbamate pesticides: Revisiting common mechanisms of action with new approach methodologies. Reproductive Toxicology 2019;89:83–92. https://doi.org/10.1016/j.reprotox.2019.07.007.

[16] Alfonso M, Durán R, Fajardo D, Justo L, Faro LRF. Mechanisms of action of paraoxon, an organophosphorus pesticide, on in vivo dopamine release in conscious and freely moving rats. Neurochemistry International 2019;124:130–40. https://doi.org/10.1016/j.neuint.2019.01.001.

[17] Ventura C, Zappia CD, Lasagna M, Pavicic W, Richard S, Bolzan AD, et al. Effects of the pesticide chlorpyrifos on breast cancer disease. Implication of epigenetic mechanisms. The Journal of Steroid Biochemistry and Molecular Biology 2019;186:96–104. https://doi.org/10.1016/j.jsbmb.2018.09.021.

[18] Li C, Xie Y, Guo Y, Cheng Y, Yu H, Qian H, et al. Effects of ozone-microbubble treatment on the removal of residual pesticides and the adsorption mechanism of pesticides onto the apple matrix. Food Control 2021;120:107548. https://doi.org/10.1016/j.foodcont.2020.107548.

[19] Li Z. Spatiotemporal pattern models for bioaccumulation of pesticides in common herbaceous and woody plants. Journal of Environmental Management 2020;276:111334. https://doi.org/10.1016/j.jenvman.2020.111334.

[20] Etzel RA, Forthal DN, Hill RH, Demby A. Fatal parathion poisoning in Sierra Leone n.d.:5.

[21] Das GP, Jamil K, Rahman MF. Effect of Four Organophosphorus Compounds on Human Blood Acetylcholinesterase: In Vitro Studies. Toxicology Mechanisms and Methods 2006;16:455–9. https://doi.org/10.1080/15376520600719281.

[22] Ranjan A, Chauhan A, Jindal T. In-silico and in-vitro evaluation of human acetylcholinesterase inhibition by organophosphates. Environmental Toxicology and Pharmacology 2018;57:131–40. https://doi.org/10.1016/j.etap.2017.12.014.

[23] Niraj RRK, Saini V, Kumar A. QSAR analyses of organophosphates for insecticidal activity and its in-silico validation using molecular docking study. Environmental Toxicology and Pharmacology 2015;40:886–94. https://doi.org/10.1016/j.etap.2015.09.021.

[24] Sahub C, Tuntulani T, Nhujak T, Tomapatanaget B. Effective biosensor based on graphene quantum dots via enzymatic reaction for directly photoluminescence detection of organophosphate pesticide. Sensors and Actuators B: Chemical 2018;258:88–97. https://doi.org/10.1016/j.snb.2017.11.072.

[25] Zvinavashe E, Du T, Griff T, Berg HHJ van den, Soffers AEMF, Vervoort J, et al. Quantitative structure-activity relationship modeling of the toxicity of organothiophosphate pesticides to Daphnia magna and Cyprinus carpio. Chemosphere 2009;75:1531–8. https://doi.org/10.1016/j.chemosphere.2009.01.081.

[26] Villaverde JJ, Sandín-España P, Alonso-Prados JL, Lamsabhi AM, Alcamí M. Pesticide byproducts formation: Theoretical study of the protonation of alloxydim degradation products. Computational and Theoretical Chemistry 2018;1143:9–19. https://doi.org/10.1016/j.comptc.2018.08.006.

[27] Vaz WF, D'Oliveira GDC, Perez CN, Neves BJ, Napolitano HB. Machine learning prediction of the potential pesticide applicability of three dihydroquinoline derivatives: Syntheses, crystal structures and physical properties. Journal of Molecular Structure 2020;1206:127732. https://doi.org/10.1016/j.molstruc.2020.127732.

[28] Villaverde JJ, Sevilla-Morán B, López-Goti C, Alonso-Prados JL, Sandín-España P. QSAR/QSPR models based on quantum chemistry for risk assessment of pesticides according to current European legislation. SAR and QSAR in Environmental Research 2020;31:49–72. https://doi.org/10.1080/1062936X.2019.1692368.

[29] Moreira-Filho JT, Braga RC, Lemos JM, Alves VM, Borba JVVB, Costa WS, et al. BeeToxAI: An artificial intelligence-based web app to assess acute toxicity of chemicals to honey bees. Artificial Intelligence in the Life Sciences 2021;1:100013. https://doi.org/10.1016/j.ailsci.2021.100013.

[30] Contreras I, Vehi J. Artificial Intelligence for Diabetes Management and Decision Support: Literature Review. J Med Internet Res 2018;20:e10775. https://doi.org/10.2196/10775.

[31] Vijayan RSK, Kihlberg J, Cross JB, Poongavanam V. Enhancing preclinical drug discovery with artificial intelligence. Drug Discovery Today 2022;27:967–84. https://doi.org/10.1016/j.drudis.2021.11.023.

[32] Loyola-Gonzalez O. Black-Box vs. White-Box: Understanding Their Advantages and Weaknesses From a Practical Point of View. IEEE Access 2019;7:154096–113. https://doi.org/10.1109/ACCESS.2019.2949286.

[33] Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell 2019;1:206–15. https://doi.org/10.1038/s42256-019-0048-x.

[34] Breiman L. Random Forest. Machine Learning 2001;45:5–32. https://doi.org/10.1023/A:1010933404324.

[35] Ali J, Khan R, Ahmad N, Maqsood I. Random Forests and Decision Trees 2012;9.

[36] Biau G. Analysis of a Random Forests Model 2010. https://doi.org/10.48550/ARXIV.1005.0208.

[37] Tibshirani R. Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society Series B (Methodological) 1996;58:267–88.

[38] Melkumova LE, Shatskikh SYa. Comparing Ridge and LASSO estimators for data analysis. Procedia Engineering 2017;201:746–55. https://doi.org/10.1016/j.proeng.2017.09.615.

[39] Hoerl AE, Kennard RW. Ridge Regression: Biased Estimation for Nonorthogonal Problems n.d.

[40] McDonald GC. Ridge regression. WIREs Comp Stat 2009;1:93–100. https://doi.org/10.1002/wics.14.

[41] Zou H, Hastie T. Regularization and variable selection via the elastic net. J Royal Statistical Soc B 2005;67:301–20. https://doi.org/10.1111/j.1467-9868.2005.00503.x.

[42] Cortes C, Vapnik V. Support-vector networks. Mach Learn 1995;20:273–97. https://doi.org/10.1007/BF00994018.

[43] Brereton RG, Lloyd GR. Support Vector Machines for classification and regression. Analyst 2010;135:230–67. https://doi.org/10.1039/B918972F.

[44] Geerlings P, Chamorro E, Chattaraj PK, De Proft F, Gázquez JL, Liu S, et al. Conceptual density functional theory: status, prospects, issues. Theor Chem Acc 2020;139:36. https://doi.org/10.1007/s00214-020-2546-7.

[45] Franco-Pérez M, Polanco-Ramírez CA, Gázquez JL, Ayers PW, Vela A. Study of organic reactions using chemical reactivity descriptors derived through a temperature-dependent approach. Theor Chem Acc 2020;139:44. https://doi.org/10.1007/s00214-020-2557-4.

[46] Ramírez-Palma DI, García-Jacas CR, Carpio-Martínez P, Cortés-Guzmán F. Predicting reactive sites with quantum chemical topology: carbonyl additions in multicomponent reactions. Phys Chem Chem Phys 2020;22:9283–9. https://doi.org/10.1039/D0CP00300J.

[47] Bader RFW. Atoms in molecules: a quantum theory. Oxford [England] : New York: Clarendon Press ; Oxford University Press; 1994.

[48] Cortesguzman F, Bader R. Complementarity of QTAIM and MO theory in the study of bonding in donor?acceptor complexes. Coordination Chemistry Reviews 2005;249:633–62. https://doi.org/10.1016/j.ccr.2004.08.022.

[49] Vijayaraj R, Subramanian V, Chattaraj PK. Comparison of Global Reactivity Descriptors Calculated Using Various Density Functionals: A QSAR Perspective. J Chem Theory Comput 2009;5:2744–53. https://doi.org/10.1021/ct900347f.

[50] Parr RG, Donnelly RA, Levy M, Palke WE. Electronegativity: The density functional viewpoint. The Journal of Chemical Physics 1978;68:3801–7. https://doi.org/10.1063/1.436185.

[51] Pearson RG. Absolute electronegativity and hardness: application to inorganic chemistry. Inorg Chem 1988;27:734–40. https://doi.org/10.1021/ic00277a030.

[52] Pearson RG. Absolute Electronegativity and Hardness: Application to Inorganic Chemistry n.d.:7.

[53] Gázquez JL. Perspectives on the Density Functional Theory of Chemical Reactivity. J Mex Chem Soc 2008:8.

[54] Lu T, Chen F. Multiwfn: A multifunctional wavefunction analyzer n.d.:13.

[55] Gázquez JL, Cedillo A, Vela A. Electrodonating and Electroaccepting Powers n.d.:5.

[56] Yadav P, Tandon H, Malik B, Chakraborty T. A new approach to compute atomic electrophilicity index in terms of Gordy's electronegativity. Journal of Chemical Research n.d.:6.

[57] Domingo LR, Perez P. The nucleophilicity N index in organic chemistry 2011:8.

[58] Martínez GM, Carlos C, Farnaz H-Z, Alain M-QR. Quantitative Electrophilicity Measures n.d.;34:13.

[59] Franco-Pérez M, Polanco-Ramírez CA, Gázquez JL, Ayers PW, Vela A. Study of organic reactions using chemical reactivity descriptors derived through a

temperature-dependent approach. Theor Chem Acc 2020;139:44. https://doi.org/10.1007/s00214-020-2557-4.

[60] Yang W, Parr RG. Hardness, softness, and the fukui function in the electronic theory of metals and catalysis. Proc Natl Acad Sci USA 1985;82:6723–6. https://doi.org/10.1073/pnas.82.20.6723.

[61] Roy RK, Krishnamurti S, Geerlings P, Pal S. Local Softness and Hardness Based Reactivity Descriptors for Predicting Intra- and Intermolecular Reactivity Sequences: Carbonyl Compounds. J Phys Chem A 1998;102:3746–55. https://doi.org/10.1021/jp973450v.

[62] Howard BE, Phillips J, Miller K, Tandon A, Mav D, Shah MR, et al. SWIFT-Review: a text-mining workbench for systematic review. Syst Rev 2016;5:87. https://doi.org/10.1186/s13643-016-0263-z.

[63] Lv T, Yan P, He W. On Massive JSON Data Model and Schema. J Phys: Conf Ser 2019;1302:022031. https://doi.org/10.1088/1742-6596/1302/2/022031.

[64] R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2021.

[65] Tomasulo P. ChemIDplus-Super Source for Chemical and Drug Information. Medical Reference Services Quarterly 2002;21:53–9. https://doi.org/10.1300/J115v21n01_04.

[66] Bourhis P, Reutter JL, Vrgoč D. JSON: Data model and query languages. Information Systems 2020;89:101478. https://doi.org/10.1016/j.is.2019.101478.

[67] Ooms J. The jsonlite Package: A Practical and Consistent Mapping Between JSON Data and R Objects. ArXiv:14032805 [StatCO] 2014.

[68] Fourches D, Muratov E, Tropsha A. Trust, But Verify: On the Importance of Chemical Structure Curation in Cheminformatics and QSAR Modeling Research. J Chem Inf Model 2010;50:1189–204. https://doi.org/10.1021/ci100176x.

[69] Grob S. Molinspiration Chemoinformatics software 2023.

[70] Regulation (EC) No 1907/2006 of the European Parliament and of the Council of 18 December 2006 concerning the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH), establishing a European Chemicals Agency, amending Directive 1999/45/EC and repealing Council Regulation (EEC) No 793/93 and Commission Regulation (EC) No 1488/94 as well as Council Directive 76/769/EEC and Commission Directives 91/155/EEC, 93/67/EEC, 93/105/EC and 2000/21/EC (Text with EEA relevance). 2023.

[71] Frisch MJ, Trucks GW, Schlegel HB, Scuseria GE, Robb MA, Cheeseman JR, et al. Gaussian˜16 Revision C.01 2016.

[72] Chai J-D, Head-Gordon M. Long-range corrected hybrid density functionals with damped atom–atom dispersion corrections. Phys Chem Chem Phys 2008;10:6615. https://doi.org/10.1039/b810189b.

[73] Curtiss LA, Redfern PC, Rassolov V, Kedziora G, Pople JA. Extension of Gaussian-3 theory to molecules containing third-row atoms K, Ca, Ga–Kr. The Journal of Chemical Physics 2001;114:9287–95. https://doi.org/10.1063/1.1366337.

[74] de Castro EAS, de Oliveira DAB, Farias SAS, Gargano R, Martins JBL. Structure and electronic properties of azadirachtin. J Mol Model 2014;20:2084. https://doi.org/10.1007/s00894-014-2084-0.

[75] Deb DK, Sarkar B. Formation of Criegee intermediates and peroxy acids: a computational study of gas-phase 1,3-cycloaddition of ozone with catechol. Phys Chem Chem Phys 2019;21:14589–97. https://doi.org/10.1039/C9CP01312A.

[76] Mukherjee S, Thilagar P. Effect of alkyl substituents in BODIPYs: a comparative DFT computational investigation. RSC Adv 2015;5:2706–14. https://doi.org/10.1039/C4RA12071J.

[77] Hay PJ, Wadt WR. *Ab initio* effective core potentials for molecular calculations. Potentials for the transition metal atoms Sc to Hg. The Journal of Chemical Physics 1985;82:270–83. https://doi.org/10.1063/1.448799.

[78] Hourahine B, Aradi B, Blum V, Bonafé F, Buccheri A, Camacho C, et al. DFTB+, a software package for efficient approximate density functional theory based atomistic simulations. J Chem Phys 2020;152:124101. https://doi.org/10.1063/1.5143190.

[79] Todd A K. AIMAll 2019.

[80] Yeo I-K. A new family of power transformations to improve normality or symmetry. Biometrika 2000;87:954–9. https://doi.org/10.1093/biomet/87.4.954.

[81] Tian W, Zhang G, Zhang X, Dong Y. PCA weight and Johnson transformation based alarm threshold optimization in chemical processes. Chinese Journal of Chemical Engineering 2018;26:1653–61. https://doi.org/10.1016/j.cjche.2017.10.027.

[82] González-Estrada E, Cosmes W. Shapiro–Wilk test for skew normal distributions based on data transformations. Journal of Statistical Computation and Simulation 2019;89:3258–72. https://doi.org/10.1080/00949655.2019.1658763.

[83] De P, Kar S, Ambure P, Roy K. Prediction reliability of QSAR models: an overview of various validation tools. Arch Toxicol 2022;96:1279–95. https://doi.org/10.1007/s00204-022-03252-y.

[84] Darst BF, Malecki KC, Engelman CD. Using recursive feature elimination in random forest to account for correlated variables in high dimensional data. BMC Genet 2018;19:65. https://doi.org/10.1186/s12863-018-0633-8.

7 ANEXOS

Journal of Molecular Modeling (2023) 29:217
https://doi.org/10.1007/s00894-023-05630-4

ORIGINAL PAPER

Conceptual DFT, machine learning and molecular docking as tools for predicting LD_{50} toxicity of organothiophosphates

Uriel J. Rangel-Peña[1] · Luis A. Zárate-Hernández[1] · Rosa L. Camacho-Mendoza[1] · Carlos Z. Gómez-Castro[1] · Simplicio González-Montiel[1] · Miriam Pescador-Rojas[2] · Amilcar Meneses-Viveros[3] · Julián Cruz-Borbolla[1]

Received: 21 March 2023 / Accepted: 21 June 2023
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2023

Abstract

Context Several descriptors from conceptual density functional theory (cDFT) and the quantum theory of atoms in molecules (QTAIM) were utilized in Random Forest (RF), LASSO, Ridge, Elastic Net (EN), and Support Vector Machines (SVM) methods to predict the toxicity (LD_{50}) of sixty-two organothiophosphate compounds. The A-RF-G1 and A-RF-G2 models were obtained using the RF method, yielding statistically significant parameters with good performance, as indicated by R^2 values for the training set (R^2_{Train}) and R^2 values for the test set (R^2_{Test}), around 0.90.

Methods The molecular structure of all organothiophosphates was optimized via the range-separated hybrid functional ωB97XD with the 6–311 + + G** basis set. Seven hundred and eighty-seven descriptors have been processed using a variety of machine learning algorithms: RF LASSO, Ridge, EN and SVM to generate a predictive model. The properties were obtained with Multiwfn, AIMALL and VMD programs. Docking simulations were performed by using AutoDock 4.2 and LigPlot + programs. All the calculations in this work are carried out in Gaussian 16 program package.

Keywords Organothiophosphate · Toxicity · cDFT · QSTR · Artificial intelligence · Molecular docking

Introduction

The production of a great variety of crops can be affected by different organisms, including fungi, insects, animals as well as other plants, which have been classified as pests and their presence in the crops cause economic loss and problems in the food supply chain [1, 2]. The common solution to avoid, warn and reduce the damage due the presence of pests in gardens, agricultural land, and other areas has been the use of pesticides, which are recalcitrant chemical products able to control diseases, inhibit, and suppress pests [3]. The term pesticide includes an enormous number of chemical products that are used in both domestic and industrial environments and their use can change significantly their chemical and physical properties among others [4]. Pesticides can enter into the animals orally, nasally or via dermal absorption and subsequently are transported in the bloodstream to be metabolized and finally excreted through urine, sweat or absorbed in adipose tissue in their unmetabolized or metabolized form [5]. Due to the lipophilic properties of pesticides, they are transported along with lipids into the body, causing metabolic disorders such as obesity, dyslipidemia, insulin resistance, and thyroid function disorders [6, 7]. On the other hand, pre or postnatal exposure to pesticides can affect the development in adulthood due that compounds used in pesticides are available in breast milk [8]. The presence of low concentrations of organophosphates and organothiophosphates compounds in mice are viable to inhibit the enzymatic activity of acetylcholinesterase, causing oxidative stress that alters the metabolism of macronutrients in these mammals, generating malformations in them [9, 10].

Organothiophosphates contain a thiophosphoryl group (P = S) and constitute a broad class of widely used insecticides. The frequent use of organothiophosphates on

✉ Julián Cruz-Borbolla
jcruz@uaeh.edu.mx

[1] Area Académica de Química, Centro de Investigaciones Químicas, Universidad Autónoma del Estado de Hidalgo, Km. 4.5 Carretera Pachuca-Tulancingo, Ciudad del Conocimiento, C.P. 42184, Mineral de La Reforma, Hidalgo, México

[2] Escuela Superior de Cómputo, Instituto Politécnico Nacional, Mexico, México

[3] Departamento de Computación, CINVESTAV-IPN, Av. IPN 2508, Col. San Pedro Zacatenco, Ciudad de Mexico 07360, México

Published online: 28 June 2023

El Comité Nacional de la

Reunión Mexicana de Fisicoquímica Teórica

XX REUNION MEXICANA DE FISICOQUIMICA TEORICA

Otorga la presente

CONSTANCIA

A: Uriel Josafat Rangel Peña

Por su participación en la XX RMFQT con la ponencia oral: "DETERMINACIÓN DE MODELOS QSTR EMPLEANDO DFT E IA EN COMPUESTOS ORGANOTIOFOSFORADOS", ocurrida en Cuernavaca, Morelos del 17 al 19 de noviembre de 2022

Dra. Cercis Mórera Boado
UAEM
Comité Local de la RMFQT

Dr. Joaquín Barroso Flores
UNAM
Comité Nacional de la RMFQT