UNIVERSIDAD DEL AZUAY

Facultad de Ciencia y Tecnología

Escuela de Ingeniería en Alimentos

Estudio quimioinformático del perfil aromático de muestras de miel de abejas de diferente origen botánico

Trabajo previo a la obtención del grado

académico de Ingeniero en Alimentos

Autor(as):

Astudillo Serrano Romina Victoria

Gómez Marín Daniela Alejandra

Director:

Dr. Cristian Rojas Villa

Cuenca – Ecuador

2024

DEDICATORIA

Quiero dedicar este trabajo de titulación a mi familia. Especialmente a mis padres, Víctor Hugo Astudillo Sánchez y María Isabel Serrano Fonseca, quienes siempre han sido la inspiración de mi vida y me han brindado fortaleza y motivación. También a mis abuelitos, que siempre me han dado su apoyo incondicional y han sido fundamentales en este camino.

Romina Astudillo Serrano

Este trabajo está dedicado a mis papás, Liliam Marín y Tarquino Gómez, por su incondicional apoyo.

Con todo el amor del mundo para Paula, por ser mi principal fuente de inspiración.

Daniela Gómez Marín

AGRADECIMIENTOS

Agradezco a Dios, porque sin Él nada es posible. Él me ha otorgado la sabiduría, inteligencia y fortaleza necesarias para terminar mi carrera. Quiero agradecer a mis padres, por ser siempre mi guía, por su amor incondicional y por todo su esfuerzo. A mis abuelitos, por ser una parte fundamental de mi trayectoria universitaria; su sabiduría y conocimiento han sido una gran fuente de motivación. A mis hermanos les agradezco su paciencia y motivación, especialmente a mi hermana, quien ha sido mi compañera en esta etapa de mi vida y nunca me dejó sola.

También quiero expresar mi gratitud a mis compañeros, quienes me acompañaron en esta etapa, por las valiosas experiencias y aprendizajes compartidos. En especial, a mi amiga y compañera Daniela, con quien siempre nos apoyamos y motivamos a lo largo de la carrera.

Agradezco profundamente a mis profesores por las enseñanzas y conocimientos adquiridos, los cuales han sido fundamentales en mi formación profesional. En particular, quiero destacar al Dr. Cristian Rojas Villa y al Dr. Piercosimo Tripaldi, quienes me guiaron durante este trabajo, compartiendo su sabiduría y experiencia.

Romina Astudillo Serrano

Quiero empezar agradeciendo a Dios por guiarme en cada paso que he alcanzado mi vida, a mis padres por sus incansables esfuerzos. A mis hermanas, Lily y Cristina, por su apoyo y amor incondicional a pesar de la distancia. A mis sobrinos Darina y Pablo, porque a pesar de su corta edad han sido mi fortaleza durante este camino. A mi compañero de vida, Steven, por su paciencia y su aliento en cada momento.

Gracias a mi compañera Romina por permitirme realizar este trabajo juntas. Al Doctor Cristian Rojas Villa y al Doctor Piercosimo Tripaldi, por compartir conmigo sus conocimientos para la elaboración de este trabajo. Finalmente, deseo expresar mi gratitud de manera especial al Doctor Diego Suarez, por ser un excelente profesor y amigo, por su infinita paciencia y sus interminables consejos.

Sepan que los recuerdo siempre con un cariño sincero.

Daniela Gómez Marín

Deseamos expresar nuestros más sinceros agradecimientos al Doctor Cristian Rojas Villa. Gracias de corazón por guiarnos durante la elaboración de este trabajo, por compartir sus conocimientos y confiar en nosotras, por su apoyo constante y sincero en todo momento. Nos sentimos orgullosas y afortunadas de poder compartir este proyecto con una persona tan sabia que ha aportado mucho a la ciencia. Gracias por hacer esto posible y por permitirnos conseguir nuestros objetivos de la mejor manera.

Romina y Daniela.

RESUMEN

En el presente trabajo se realizó un modelado quimioinformático del perfil aromático de 40 muestras de miel de abejas de diferente origen botánico mediante las relaciones cuantitativas estructura propiedad (QSPR). Para el desarrollo del modelo se utilizó una base de datos de 151 compuestos orgánicos volátiles con sus índices de retención (*I*). Este parámetro fisicoquímico se cuantificó en un cromatógrafo de gases Agilent 6890 acoplado al detector de un espectrómetro de masas cuadrupolo Agilent 5973. La separación se realizó en una columna capilar de polietilenglicol HP-Innowax. Cada compuesto fue representado por 4146 descriptores moleculares independientes de la conformación, los cuales fueron analizados mediante el algoritmo de Wootton, Sergent y Phan-Tan-Luu para la reducción no supervisada. Posteriormente, la base de datos se dividió en grupos de calibración y predicción con 105 y 46 compuestos, respectivamente. Las moléculas de calibración se usaron para la selección supervisada de descriptores mediante los algoritmos genéticos (GAs) acoplados con la regresión lineal múltiple (MLR). Este proceso permitió seleccionar un modelo óptimo con 8 descriptores moleculares analizado mediante el coeficiente de determinación (R^2) y el error cuadrático medio (*RMSE*). El modelo *in silico* presenta buena calidad en calibración (R^2 = 0.893 y *RMSEC* = 127.40) y predicción (R^2 = 0.908 y *RMSEC* = 118.23). Complementariamente, la estabilidad del modelo fue analizada mediante técnicas de validación cruzada, mientras que la ausencia de correlación casual se cuantificó mediante la aleatorización-Y. Así, también se definió el dominio de aplicabilidad.

Palabras claves: *Compuestos orgánicos volátiles, Descriptores moleculares, Índice de retención, Miel de abejas, QSPR.*

ABSTRACT

In this study, cheminformatic modeling of the aromatic profile of 40 honeybees from different botanical origins was carried out by means of quantitative structure property relationships (QSPR). For the development of the model, a database of 151 volatile organic compounds with their retention indices (*I*) was used. This physicochemical parameter was quantified using an Agilent 6890 gas chromatograph coupled to the detector of an Agilent 5973 quadrupole mass spectrometer. Chromatographic separation was carried out using an HP-Innowax polyethylene glycol capillary column. Each compound was represented by 4146 conformation-independent molecular descriptors, which were analyzed using the Wootton, Sergent and Phan-Tan-Luu algorithm for unsupervised reduction. Subsequently, the database was split into calibration and prediction sets of 105 and 46 compounds, respectively. The calibration molecules were used for the supervised selection of descriptors using genetic algorithms (GAs) coupled to multiple linear regression (MLR). This process allowed the selection of an optimal model with 8 molecular descriptors, which was analyzed by the coefficient of determination (R^2) and the root mean squared error (*RMSE*). The *in silico* model exhibited good quality in calibration (R^2 = 0.893 and *RMSEC* = 127.40) and prediction (R^2 = 0.908 and *RMSEP* = 118.23). In addition, the stability of the model was analyzed using cross-validation techniques, while the absence of change correlation was diagnosed by Y-randomization along with applicability domain assessment.

Keywords: *Honeybees, Volatile organic compounds, Retention Index, Molecular descriptors, QSPR.*

INDICE DE CONTENIDOS

INDICE DE TABLAS

ÍNDICE DE FIGURAS

INTRODUCCIÓN

La relación cuantitativa estructura/propiedad, reconocida como QSPR por su abreviatura en inglés, ha tenido un alcance fundamental en el área de la quimioinformática debido a su versatilidad. Comprender la correlación que mantiene la estructura molecular con las propiedades fisicoquímicas de los compuestos supone una etapa importante en el campo de la industria alimentaria, pues permite construir una relación matemática entre la estructura química de las moléculas y sus características fisicoquímicas. Así, dichos análisis se vuelven relevantes debido a su aptitud de prever las propiedades de nuevas sustancias, lo que se complementa con el uso de las innovadoras herramientas computacionales (Todeschini & Consonni, 2009).

Hoy en día existe un aumento de disponibilidad de datos, lo que da lugar a revelar patrones que sirven como guía en la investigación facilitando la toma de decisiones en la industria alimentaria. Estos estudios tratan de comprender las relaciones de las estructuras químicas de los compuestos orgánicos volátiles (VOCs) teniendo en cuenta que son los que contribuyen al perfil aromático de los alimentos. Se puede obtener los índices de retención de los compuestos a partir de una técnica que integra el muestreo, la concentración y la extracción de los VOCs; la microextracción en fase sólida (SPME), la cual, se utiliza previo al análisis mediante cromatografía de gases acoplada a la espectrometría de masas (GS-MS), de manera que se analiza de forma eficaz los VOCs en muestras de alimentos. Así pues, la combinación de la capacidad de separar los compuestos con la cromatografía de gases y la perceptibilidad y especificidad de la espectrometría de masas es ideal para el análisis de mezclas complejas (Soria, Martínez-Castro, & Sanz, 2008).

En este proyecto de titulación, se buscará analizar el índice de retención el cual se identifica como un factor relevante ya que, a partir de este valor podremos identificar los VOCs. Como se mencionó anteriormente, el perfil aromático de los alimentos está fuertemente vinculado con su composición volátil, de esta manera, se reconoce como un factor relevante al momento de decisión del consumidor. Los compuestos analizados se obtuvieron de un amplio conjunto de datos obtenidos de 40 variedades distintas de miel, tomados de la base de datos propuesto en el artículo "*SPME followed by GC–MS: a powerful technique for qualitative analysis of honey volatiles"* (Soria, Sanz, & Martínez-Castro, 2009), donde utilizó la técnica GC-MS. Los análisis de cromatografía

de gases acoplada a espectrometría de masas (GC-MS) se realizaron en un cromatógrafo de gases Agilent 6890 acoplado a un detector cuadrupolo de espectrómetro de masas Agilent 5973. Se analizaron los índices de retención y se planteó estrategias para visualizar las moléculas de VOCs presentes en la miel, de manera que se facilite la construcción de modelos informáticos predictivos con aplicaciones importantes (Acevedo, Carrasco, & Basterrechea, 2010; Soria et al., 2009).

CAPÍTULO I

MARCO TEÓRICO

1.1 Generalidades de la miel de abejas

En conformidad a la Organización de las Naciones Unidas para la Alimentación y la Agricultura (Codex Alimentarius 12-1981), se define a la miel de abejas como un producto natural, que se reconoce por su sabor dulce y es elaborada por las abejas *Apis Mellifera* partiendo del néctar de las flores o de exudaciones de fuentes extra florales (insectos succionadores de plantas). Estas secreciones son recolectadas y posteriormente transformadas en mezclas con sustancias específicas propias de las abejas. A continuación, las deshidratan, almacenan y finalmente las depositan en panales para su posterior maduración (Alimentarius, 1981).

La miel de abejas forma parte de una de las formas más antiguas de alimentación, puesto que se considera como uno de los primeros productos en ser aprovechados por el hombre para nutrirse. Dentro de su composición, se destaca la presencia de carbohidratos como glucosa y fructosa en mayor cantidad, además, se pueden encontrar otras sustancias en pequeñas cantidades como minerales, proteínas, vitaminas, ácidos orgánicos, ácidos fenólicos y otros fitoquímicos (Bertoncelj, Doberšek, Jamnik, & Golob, 2007). A la miel de abejas se le han asociado ciertas características beneficiosas para el ser humano, por ejemplo, es una fuente rica de antioxidantes naturales, los mismos que tienen la capacidad de disminuir el riesgo de enfermedades crónico degenerativas. Se ha demostrado que las propiedades antioxidantes de la miel están asociadas a su contenido de fenoles totales, y otras sustancias tales como el ácido ascórbico, la catalasa, la peroxidasa. La cantidad presente de cada compuesto varía según el origen botánico de la miel. Además, la miel es una sustancia idónea para evitar las reacciones de oxidación que tienden degenerar la vida útil del alimento, por ejemplo, el pardeamiento enzimático en frutas y verduras, la oxidación lipídica de la carne (Bertoncelj et al., 2007).

El aspecto físico de la miel, en cuanto al color, puede variar desde un color pardo oscuro hasta no tener ningún color. Además, la miel puede ser fluida o viscosa, o por el contrario cristaliza total o parcialmente. Sus características sensoriales gustativas pueden variar dependiendo de la planta de la que proviene (Alimentarius, 1981). En general, la calidad de este producto se puede ver alterada dependiendo de algunos

factores como las fuentes florales, las condiciones climáticas y las condiciones geográficas (Vazhacharickal, 2021).

Los indicadores más importantes que determinan la calidad de la miel de abejas son la ausencia de posibles contaminantes y la frescura de la misma. Este alimento debe carecer de metales pesados en cantidades significativas que suponen un peligro para la salud de quien la consume. Además, la miel comercial no debe contener ingredientes adicionales, ni debe haber empezado su fermentación. No se deberá someter este producto a calentamiento y a su vez se deberá evitar tratamientos químicos que facilitan su cristalización (Alimentarius, 1981).

Por otra parte, la frescura se determinará a partir del hidroximetilfurfural (HMF) y la actividad enzimática; diastasa. El HMF se trata de un aldehído cíclico que se produce por la descomposición de la fructosa por la presencia de ácidos, permite medir la frescura de la miel debido a que sus niveles aumentan por la exposición a altas temperaturas. El contenido máximo debería ser de 40 mg/kg y 80 mg/kg para mieles de origen tropical. En cuanto a su actividad enzimática, la diastasa es una enzima que se encuentra presente de manera natural en las mieles frescas, es termolábil y sus niveles disminuyen durante su período de almacenamiento o por el calentamiento, lo que indica el nivel de antigüedad de la miel de abejas (Suescún & Vit, 2008).

La miel de abejas puede clasificarse en dos categorías:

1. Monoflorales: predomina el néctar de una sola especie. Es decir, la miel contiene más del 45% de polen de un tipo de flor.

2. Multiflorales: se componen del néctar de varias especies diferentes, y su proporción varía de acuerdo al origen (Suescún & Vit, 2008).

1.2 Compuestos orgánicos volátiles de la miel

Los constituyentes volátiles de la miel de abejas están directamente relacionados con su aroma, el cual define la calidad del producto y tiene gran relevancia en la elección de los consumidores del tipo de miel. La calidad de las diversas mieles depende de su contenido de azúcares, de su aroma, densidad, color, y de sus características físicas y químicas; las cuales están estrechamente relacionadas con el tipo de plantas o flores que se utilicen. Las condiciones climáticas también afectan a la composición y propiedades

de la miel, generalmente las mieles de color claro tienen un aroma suave, mientras que las oscuras presentan un aroma más fuerte (Soria et al., 2009).

El aroma de la miel de abejas, por su parte, se debe al contenido de algunas sustancias odoríferas volátiles que se encuentran en pequeñas cantidades y se juntan. El reconocimiento de los VOCs representa un desafío en este campo, ya que contener o soltar estas sustancias es un proceso importante en la ciencia de los alimentos (Soria et al., 2009). Es crucial elaborar miel con características distintas que puedan caracterizar y validar su origen. Se reconoce a la miel como un alimento altamente consumido por sus características sensoriales y aromáticas, gracias a sus componentes químicos de bajo peso molecular. Además, se sabe que estos compuestos volátiles tienen origen en diversas familias químicas y dependen de la exudación de las plantas y de su néctar, pero también se pueden formar al momento de procesar y almacenar la miel (Cuevas-Glory, Ortiz-Vázquez, Centurión-Yah, Alea, & Sauri-Duch, 2008). Por lo tanto, identificar y reconocer la proveniencia floral de la miel de abejas es importante, puesto que de esto dependen las características sensoriales del producto final que buscan los consumidores al momento de comprarla. Se puede determinar la procedencia botánica de la miel realizando un estudio de los compuestos orgánicos volátiles (VOCs) por cromatografía de gases acoplada a espectrometría de masas (GC-MS) (Barra, Ponce-Díaz, & Venegas-Gallegos, 2010).

En resumen, es importante identificar y reconocer los compuestos orgánicos volátiles de este producto, para diferenciar su procedencia floral, pues existen varios tipos de miel que tienen perfiles de compuestos orgánicos volátiles (VOCs) diferentes. La identificación y el análisis de los VOCs también puede ser un método de garantía de calidad, de esta manera se puede verificar que la miel no ha sido adulterada o posee algún tipo de contaminación. Existen diferentes métodos y técnicas quimiométricas o de análisis de los VOCs que brindan una manera viable de analizar la procedencia botánica de este producto, especialmente si la procedencia es incierta (Baroni et al., 2006).

Asimismo, es importante saber que la composición de la miel ayuda a comprender su sabor, aroma y posibles beneficios sobre la salud. Por otro lado, la miel de abejas también es importante en el desarrollo de otros productos los cuales pueden tener fines terapéuticos. En consecuencia, un buen estudio y análisis de la miel contribuirá a

complacer las necesidades y preferencias de los consumidores, con la posibilidad una alta gama de propiedades sensoriales con diferentes variedades para mayor satisfacción del producto. Por lo tanto, los consumidores tienen a disposición una amplia variedad de miel dependiendo sus propiedades terapéuticas, antimicrobianos y antioxidantes que requieran (Baroni et al., 2006).

El estudio de los componentes orgánicos volátiles (VOCs) de la miel ayuda a reconocer e identificar los componentes de la misma, de manera que se determinará su origen floral. Este análisis se puede desarrollar mediante técnicas como la microextracción en fase sólida (SPME) antes de su reconocimiento por cromatografía de gases acoplada a la espectrometría de masas (GC-MS), la cual es una técnica comúnmente utilizada para el análisis de VOCs en muestras de miel. La microextracción en fase sólida se reconoce como un método sencillo que da lugar a la extracción y desunión de los VOCs de la matriz de la miel, seguida de su identificación y cuantificación mediante espectrometría de masas (Baroni et al., 2006).

La microextracción en fase sólida es, además, una técnica accesible y adaptable que mezcla el muestreo, la extracción y concentración de los compuestos volátiles antes de su análisis, es una herramienta que se ha utilizado ampliamente y es reconocida en la caracterización del aroma de la miel. Existen varios mecanismos de aplicación y extracción con diferente polaridad en el campo de la ciencia de alimentos; el mecanismo de extracción del recubrimiento que tiene el fin de recuperar los VOCs de la miel y la polaridad tienen gran influencia en este proceso, estos pueden tener diferentes resultados de dispersión y sensibilidad, en el estudio propuesto por (Soria et al., 2009), se utilizaron varias fibras de diferente polaridad para extraer compuestos volátiles de 40 muestras de miel antes de su análisis por GC-MS. Su objetivo fue explorar la capacidad de esta técnica para caracterizar la compleja composición volátil de mieles uniflorales comerciales y mezclas de mieles.

1.3 Cromatografía de gases acoplada a espectrometría de masas

La cromatografía de gases (GC) es un método que da lugar a la separación de los componentes de una mezcla en función de su volatilidad. La espectrometría de masas (MS), por otro lado, es una técnica que ioniza los compuestos separados y mide su relación masa-carga, brinda información sobre el peso molecular y estructura de los

compuestos (Rojas-Monroy, 2005). La combinación de estas dos técnicas (GC-MS) se convierte en una herramienta importante que permite analizar de manera simultánea todos los compuestos químicos presentes en una muestra; compuestos volátiles o semivolátiles (Stashenko & Martínez, 2010). El proceso se lleva a cabo con una vaporización inicial del prototipo inyectándole en un cromatógrafo de gases en el cual se separan los constituyentes individuales. Una vez que los componentes se han separado se ionizan y se fragmentan en el espectrómetro de masas. El resultado será un espectro de masas reproducible mediante el cual se pueden identificar los compuestos que se encuentran en la muestra. Por lo tanto, se permite realizar el reconocimiento y la cuantificación de varios componentes dentro de la industria alimentaria y diferentes campos (Stashenko & Martínez, 2010). En la Figura 1 se puede observar un ejemplo de un equipo de cromatografía de gases acoplado a un espectrómetro de masas.

En la técnica GC-MS se obtiene el índice de retención (*I*) de un compuesto, el cual es un valor obtenido mediante interpolación (normalmente logarítmica), donde se muestra la relación del tiempo de retención ajustado de un compuesto especifico con los tiempos de retención de muestras que eluyen antes y después del pico objetivo. Para su cálculo se consideran *n*-alcanos como patrones y se deduce la interpolación logarítmica (Schomburg, 1990). Por ejemplo; en el caso del *n*-undecano ($nC_{11}H_{24}$), el índice de retención es $I = 1100$.

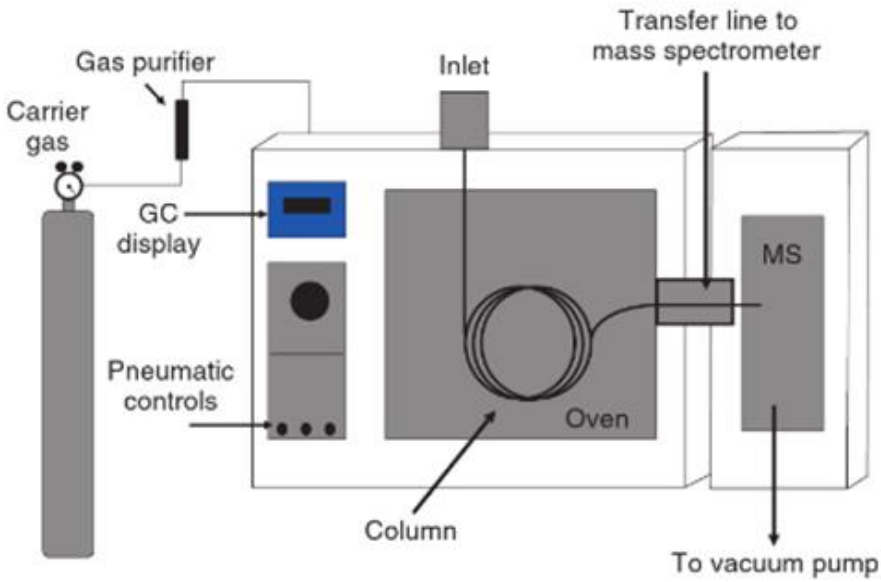

Figura 1 Diagrama de un cromatógrafo de gases acoplado a un espectrómetro de masas (Sparkman, Penton, & Kitson, 2011)

1.4 Relaciones cuantitativas estructura-propiedad (QSPR)

La teoría de las relaciones cuantitativas estructura-actividad/propiedad (QSAR/QSPR) (Dearden, 2016; Hansch & Leo, 1995; Todeschini & Consonni, 2009) o también conocida como estudios *in silico*, tiene origen hace muchos años a partir de trabajos que han impulsado el interés de predecir propiedades, particularmente en el campo del diseño de fármacos y la química medicinal. No obstante, la teoría QSAR/QSPR ha evolucionado con el paso del tiempo hasta llegar a ser crucial en diversas diciplinas como las ciencias medicinales, farmacológicas, químicas de los alimentos. La metodología QSAR se emplea para predecir la relación actividad y propiedad de las moléculas en estudio, las mismas que permitirán conseguir deducciones sobre la información del fenómeno de interés. Hoy en día, varias organizaciones internacionales han reconocido el desarrollo de modelos predictivos QSAR como una herramienta eficaz para la exploración de la información contenida en las estructuras moleculares utilizando técnicas racionales fundamentadas (Todeschini, Consonni, Ballabio, & Grisoni, 2020).

El modelado QSAR implica datos cuantitativos y el uso de técnicas quimiométricas. Los datos cuantitativos de los compuestos químicos se consiguen de dos fuentes (Roy, Kar, & Das, 2015b):

1. Obtención de datos experimentales sobre la actividad o propiedad del fenómeno que es de interés en el estudio.
2. Recopilación de datos químicos representados por medio de descriptores moleculares.

Para llevar a cabo el desarrollo del modelo QSAR es de gran importancia organizar los datos de los compuestos con respuestas conocidas y representar sus estructuras químicas para el cálculo de descriptores moleculares. Una vez que los datos se encuentran ordenados y con la información completa, se eliminan compuestos duplicados y no relevantes, para posteriormente dividir la base de datos en grupos de entrenamiento y validación para poder realizar el desarrollo y evaluación el modelo. De esta manera se podrá realizar la validación y predicción del modelo con métodos internos y externos, definiendo su dominio de aplicabilidad. Después de realizar este proceso se podrán interpretar los datos usando información de los descriptores moleculares para explorar

mecanismos de acción y crear compuestos químicos novedosos que tengan una actividad o propiedad especifica deseada.

En el caso de los VOCs, se modela el índice de retención como propiedad experimental mediante variables que representan la información de la configuración química en un formato específico contenida en un grupo de moléculas. Los descriptores moleculares tienen la capacidad de deducir características específicas de los compuestos, por ejemplo, tipo de átomos, grupos funcionales o la geometría molecular, entre otros (Rojas & Duchowicz, 2021).

Además, un modelo QSAR busca predecir una propiedad de interés de los compuestos, reducir de experimentos de laboratorio y optimizar nuevos blancos moleculares. Se debe definir claramente la propiedad a modelar, la transparencia en el algoritmo del modelo y la delimitación del dominio de aplicabilidad; para de esta manera garantizar la confiabilidad del modelo. Este tipo de estudios da lugar a un menor gasto de dinero y ayuda al proceso de desarrollo de nuevos blancos moleculares (Goode-Romero, Aguayo-Ortiz, & Domínguez, 2019; Rojas & Duchowicz, 2021; Roy, Kar, & Das, 2015a).

CAPÍTULO II

MATERIALES Y MÉTODOS

2.1 Desarrollo de la base de datos de índices de retención de los compuestos orgánicos volátiles de la miel

Para realizar un modelo QSPR, el primer requisito es crear una base de datos filtrada y curada, para de esta manera asegurar la confiabilidad en la investigación. La base de datos contendrá los datos experimentales de los índices de retención *(I)* de los compuestos orgánicos volátiles (VOCs) presentes en la miel de abejas se obtuvo del artículo científico titulado "*SPME followed by GC–MS: a powerful technique for qualitative analysis of honey volatiles*" (Soria et al., 2009). Además, se tomó información adicional de moléculas relacionadas con el perfil aromático de la miel de la siguiente literatura especializada: "*Honey Volatiles as a Fingerprint for Botanical Origin—A Review on their Occurrence on Monofloral Honeys*" (Machado, Miguel, Vilas-Boas, & Figueiredo, 2020) asimismo, se tomó información de "*Systematic Review of the Characteristic Markers in Honey of Various Botanical, Geographic, and Entomological Origins*" (Wang et al., 2022).

Se recopiló la información del índice de retención experimental de cada compuesto obtenido mediante la técnica SPME (microextracción en fase sólida) en 2 fibras de diferente polaridad y mecanismo de extracción: poliacrilato (PA) y carboxeno/polimetilsiloxano (CAR/PDMS). Las fibras SPME pasaron por un proceso de desorción a una temperatura de 250 °C durante 2 minutos. El análisis de cromatografía de gases acoplada a espectrometría de masas (GC-MS) fue realizado en un cromatógrafo de gases Agilent 6890 adaptado a un detector cuadrupolo de espectrometría de masas Agilent 5973. La separación cromatográfica se realizó en la columna capilar de polietilenglicol HP-Innowax (50 m de largo × 0.20 mm de espesor interno y grosor de película de 0.20 µm). La temperatura en el horno del cromatógrafo se programó a 45°C durante 2 min para posteriormente subir hasta 190°C a una razón de 4°C/min. Se mantuvo la temperatura máxima durante 50 minutos. Se empleó helio como gas de arrastre a una velocidad de alrededor de 1 mL/min. Los espectros de masas fueron grabados utilizando el método de impacto electrónico (EI) a una energía de 70 electronvoltios (eV), cubriendo un rango de 35 a 450 unidades de masa atómica por

carga (m/z). La temperatura de la interfaz y la fuente se mantuvo a 280°C y 230°C, respectivamente. Todos los análisis se realizaron por duplicado. (Soria et al., 2009). Para distinguir cada compuesto, se extrajo el número de registro del servicio de resúmenes químicos CAS, la especificación de introducción lineal molecular simplificada (SMILES) y el número de identificación PubChem CID de las quimiotecas: PubChem (Kim et al., 2023) y FooDB (Wishart, 2014). En la fase inicial de filtrado, se descartaron compuestos ambiguos para el modelado, los mismos que se mostraron en la base de datos con su fórmula bruta y no con su nombre químico. Además, se eliminaron compuestos no identificados los cuales se mostraron como "Not IdentiWed".

Posteriormente, el filtrado y el curado para las estructuras moleculares se llevó a cabo en el programa alvaMolecule (Alvascience, 2023b), usando los SMILES canónicos. Este proceso brindó la información necesaria para fusionar los compuestos que presentan la misma representación molecular SMILES, pero tienen diferente configuración, es decir, los compuestos que son isómeros. Para estos compuestos se calculó el índice de retención promedio para el modelado computacional.

2.2 Cálculo de descriptores moleculares

Los descriptores moleculares son representaciones matemáticas de una molécula, es decir, son algoritmos que convierten la información química de una molécula en un número, usando métodos lógicos y matemáticos (Todeschini & Consonni, 2009). Cada descriptor obtenido abarca una parte especifica de la información química presente en la estructura molecular. Se sabe que los descriptores moleculares desempeñan un papel crucial cuando se busca obtener de manera numérica y cuantitativa información de la estructura de una sustancia. En la actualidad se ha logrado establecer diversas conexiones entre la estructura molecular y las propiedades de los compuestos químicos gracias a la dedicación por parte de los científicos en el desarrollo de descriptores moleculares, y a su vez, gracias al progreso de la quimioinformática. El cálculo de los descriptores moleculares de los VOCs filtrados y curados se realizó en el programa alvaDesc (Alvascience, 2023a). Para el cálculo de los descriptores se usó la representación SMILES. Seguidamente, se aplicó una reducción de variables (descriptores) al eliminar descriptores constantes, casi constantes y con valores faltantes. Asimismo, se calcularon las claves moleculares del sistema de acceso molecular MACCS (Durant, Leland, Henry, & Nourse, 2002; Mauri, Consonni, &

Todeschini, 2017; O'Donnell, 2008), las mismas que se conforman por dos conjuntos de fragmentos basados en una representación 2D en formato MDL. Estas han sido creadas y optimizadas con el objetivo de realizar una búsqueda de subestructuras químicas. Es importante mencionar que la tabla de 166 claves moleculares contiene criterios de entrada para cada elemento de la tabla periódica y permite distinguir entre diversos compuestos orgánicos. Si una base de datos en estudio contiene en su mayoría otros tipos de compuestos (fragmentos), es adecuado el uso de un conjunto diverso de fragmentos.

2.3 Reducción no supervisada de descriptores

El aprendizaje no supervisado de reducción de variables, consiste en algoritmos que no toman en cuenta la respuesta experimental. También se los conoce como técnicas de exploración de datos que se emplean para descartar aquellas variables que muestran ruido, redundancia, correlación inusual y multicolinealidad. En otras palabras, elige las variables que sean más representativas y que conserven la máxima cantidad de información. En el presente trabajo se utilizó el método de reducción V-WSP, el cual se basa en un algoritmo propuesto por Wootton, Sergent y Phan-Tan-Luu (WSP). Este método selecciona un subconjunto representativo de descriptores de manera que se presente un valor mínimo de correlación en el espacio multidimensional. El algoritmo inicia con la selección de una variable inicial y un valor límite establecido de correlación, luego se calcula el coeficiente de correlación lineal de Pearson entre la variable inicial y todos los demás descriptores. Posteriormente se eliminan las variables cuyo valor de correlación absoluto sea mayor o igual al umbral de correlación establecido. Este proceso se repite hasta que no existan descriptores para ser seleccionados (Ballabio et al., 2014).

2.4 Desarrollo del modelo QSPR

Para el desarrollo del modelo QSPR se empleó una técnica de regresión de mínimos cuadrados ordinarios (OLS) acoplado con los algoritmos genéticos (GAs). El método OLS es comúnmente utilizado para modelar una variable aleatoria continua. Con el uso de esta técnica es posible establecer relaciones lineales entre cada una de las variables independientes y una variable dependiente. Aquí se estima los coeficientes de regresión reduciendo la suma de los cuadrados de los residuos entre los valores observados y los valores calculados. Esta estimación es equivalente a la obtenida a partir del

procedimiento de máxima verosimilitud (Rencher & Schaalje, 2008; Rojas & Duchowicz, 2021). Por otra parte, los algoritmos genéticos (Kubat, 2017; Leardi, 2020; Leardi & Gonzalez, 1998) se utilizan como método de selección supervisada de descriptores moleculares. Este método se basa en la teoría de la evolución de Darwin, de manera que trabaja a partir de una población de "cromosomas" que ha sido creada de manera casual. Se define a un cromosoma como un vector binario de *p* bits (número total de variables); donde bits que tienen un valor de 1, representan la presencia de descriptores en el modelo, mientras que los bits con valor cero indican que esos descriptores no se encuentran en el modelo. Durante los GAs, el número de cromosomas se mantiene constante. Es posible crear nuevos cromosomas por medio de dos operaciones principalmente:

1. Reproducción (crossover): se basa en la reproducción a partir de dos padres seleccionados de la población con el fin generar descendencia. Este proceso se consigue de manera aleatoria o enfocado en los individuos más convenientes. Los cromosomas de los hijos combinan el material genético de los padres (bits 0 o 1) según un método probabilístico.
2. Mutación: se trata de una operación en donde los cromosomas pueden alterarse o mutar. La mutación es una operación que evita que los individuos queden atrapados en un espacio local. La posibilidad de que ocurra una mutación es menor frente a la probabilidad de que ocurra una reproducción, puesto que es importante impedir que la población se desvíe demasiado del área ideal probable.

2.5 Validación del modelo QSPR

Con el objetivo de analizar la representatividad de la estructura-propiedad de la base de datos de VOCs, es crucial separar la base de datos en dos conjuntos: calibración y predicción. Esta división se realizará mediante el método de subconjuntos balanceados (BSM) (Rojas, Duchowicz, Tripaldi, & Pis Diez, 2015a, 2015b) el cual divide una base de datos considerando un análisis de conglomerados *k*-medias. De esta manera se garantiza que el grupo de calibración sea significativo frente al grupo de predicción en términos de la estructura-propiedad. El conjunto de calibración se utilizará para la calibración de los modelos y la selección supervisada de descriptores: Por otra parte, el conjunto de predicción se utilizará para medir la capacidad predictiva del modelo

óptimo. El BSM toma en consideración la propiedad experimental y los descriptores bidimensionales.

Adicionalmente, el modelo será sometido a protocolos de validación interna, entre los cuales se citan dejar-uno-fuera (LOO), el cual es un método que permite estimar *n* modelos en los que se descarta una molécula a la vez (*n – 1*) para luego realizar su predicción (Arlot & Celisse, 2010). También se aplicará el método de dejar-varios-fuera (LMO), el cual integra una alteración de tamaño mayor en el conjunto de calibración para conseguir una mejor valoración de la capacidad predictiva del modelo (Baumann & Stiefl, 2004). Los datos se dividen en *k* grupos de validación a partir de dos procedimientos: bloques continuos y ventanas venecianas (Ballabio & Consonni, 2013) en la cual cada *k* grupo abarca *n*/*k* elementos. Otro enfoque LOO es el método Montecarlo, en el cual en múltiples interacciones se excluye un porcentaje de moléculas de forma aleatoria. En cada iteración se calibra el modelo con los compuestos del subconjunto de entrenamiento, para posteriormente estimar las respuestas de las moléculas de evaluación (Krakowska, Custers, Deconinck, & Daszykowski, 2016). Por otro lado, el método de validación cruzada Bootstrap utiliza el muestreo aleatorio con remplazo, en esta técnica los datos del subconjunto de entrenamiento estarán formados por objetos duplicados de manera que se igualen con la cantidad de elementos que fueron separados para construir el conjunto de validación. En otras palabras, en este método se mantiene constante el tamaño de la base de datos (Efron, 1979, 1982, 1987). Finalmente, la aleatorización-Y se aplica para estimar la existencia de correlación casual en el modelo. Aquí, se transforma de forma casual los componentes del vector respuesta, de manera que no se encuentre similitud con las asignaciones iniciales; se espera que los modelos calibrados tengan una calidad inferior en comparación con el modelo desarrollado (Lindgren, Hansen, Karcher, Sjöström, & Eriksson, 1996).

2.6 Grado de contribución de los descriptores moleculares

Para evaluar cuánto influyen los descriptores moleculares en el modelo, es importante examinar los coeficientes estandarizados de la regresión lineal múltiple. Es decir, mientras mayor valor posee el coeficiente de un determinado descriptor, mayor relevancia tendrá en el modelo para estimar los índices de retención, y viceversa. Se espera determinar el significado de los descriptores de manera que se relacione la propiedad experimental, es decir, el conocimiento del modo de funcionamiento de los

descriptores ayudará a describir el índice de retención de las VOCs; sin embargo, hay que tener en cuenta que a veces la complejidad matemática dificulta la obtención de dicha interpretación (Rojas & Duchowicz, 2021).

2.7 Dominio de aplicabilidad

El dominio de aplicabilidad (AD) (Jaworska, Nikolova-Jeliazkova, & Aldenberg, 2005) se describe como un área dentro del espacio químico, determinada por los descriptores moleculares del modelo y la respuesta experimental. En este estudio se aplicó el valor de influencia (*h*) o leverage (Rencher & Schaalje, 2008), que evalúa la distancia de cada objeto con respecto al centro del modelo. Las moléculas con valores altos de influencia son las que tienen mayor distancia, por lo que constituyen una predicción poco confiable. Este valor se determina de la siguiente manera:

$$\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$$

donde **H** representa una matriz en la que los elementos diagonales (h_{ii}) muestran los valores de influencia de cada compuesto. Es decir, representa la contribución del *i*-ésimo compuesto en la estimación de su respuesta. Para el conjunto de calibración, estos valores están entre 0 y 1, a diferencia del conjunto de predicción que solo tienen un límite inferior de cero. El dominio de aplicabilidad AD establece un límite crítico (h^*), calculado como tres veces el valor promedio de las valoraciones de influencia del conjunto de calibración:

$$h^* = 3\frac{p'}{n}$$

donde p' (d + 1) representa la cantidad de parámetros del modelo y n es la cantidad de compuestos en el conjunto de calibración.

El valor de influencia para cualquier molécula del conjunto de predicción se determina con la siguiente formula:

$$h_{ii} = x_i\,(\boldsymbol{X}^T\boldsymbol{X})^{-1}x_i^T$$

Únicamente las moléculas situadas dentro de esta área teórica ($h_{ii} < h^*$) se consideran predicciones fiables o interpolaciones del modelo; de lo contrario, las predicciones se categorizan como extrapolaciones del modelo (poco confiables).

CAPÍTULO III

RESULTADOS Y DISCUSIONES

3.1 Desarrollo de la base de datos

La base de datos inicial estuvo constituida por 193 compuestos orgánicos volátiles (VOCs) y sus índices de retención obtenidos mediante la técnica GC-MS en una columna de polietilenglicol HP-Innowax a partir de 40 muestras de miel de abejas (Soria et al., 2009). Posteriormente, usando las quimiotecas PubChem (Kim et al., 2023) y FoodDB (Wishart, 2014), se llevó a cabo un comprobación del nombre químico, a partir de la cual se obtuvieron el número de registro CAS *(Chemical Abstracts Service)*, la notación lineal de la cadena SMILES *(Simplified Molecular Input Line Entry Sistem)* y el número de identificación PubChem CID. A continuación, se identificaron y eliminaron 25 compuestos ambiguos: 17 presentaban únicamente la fórmula bruta y 8 estaban declarados como "*no identificados*". De esta manera, se retuvieron 168 compuestos en la base de datos.

Debido a que el propósito de este estudio es calibrar un modelo independiente conformación, se utilizó la notación SMILES canónico para el curado de las moléculas en el programa alvaMolecule (Alvascience, 2023b). Aquí se identificaron compuestos que tienen la misma fórmula química bidimensional, es decir, los isómeros. Bajo este criterio, se fusionaron 25 compuestos y se utilizó el *I* promedio para los mismos. Como resultado, se obtuvo es una base de datos filtrada y curada de 151 compuestos orgánicos volátiles para el desarrollo del modelo QSPR.

3.2 Cálculo de los descriptores moleculares

Usando la notación lineal de cadena SMILES de los compuestos orgánicos volátiles, se calcularon 4146 descriptores moleculares independientes de la conformación utilizando el programa alvaDesc (Alvascience, 2023a). Estos descriptores se agruparon en 21 bloques. Para reducir el tamaño del conjunto inicial de descriptores, se eliminaron descriptores con valores casi constantes, constantes y que tengan valores faltantes. Adicionalmente, se calcularon 166 claves estructurales MACCS.

3.3 Reducción no supervisada de descriptores moleculares.

Se utilizó la técnica de reducción de variables V-WSP utilizando el programa MATLAB (The MathWorks Inc., 2022) el mismo que permite seleccionar un subconjunto significativo de variables de un determinado grupo de datos, de manera que encuentra una correlación baja entre las variables. Se seleccionó un umbral de correlación de 0.95 (thr = 0.95). Para llevar a cabo la reducción, se contrastó el coeficiente absoluto de correlación de Pearson (R_{ij}) con el umbral previamente seleccionado. Este procedimiento permitió reducir la base de datos a 407 descriptores para la etapa del desarrollo del modelo.

3.4 Desarrollo del modelo QSPR

Para mantener una relación estructura-propiedad, la partición de la base de datos se realizó mediante el método de subconjuntos balanceados (BSM) (Rojas & Duchowicz, 2021). De esta forma, la base de datos de 151 moléculas se dividió en conjuntos de calibración y predicción con 105 y 46 compuestos respectivamente. El grupo de calibración permitió realizar la calibración del modelo, en tanto que el grupo de predicción se empleó para evaluar la capacidad del modelo para realizar nuevas predicciones. Para seleccionar los mejores descriptores del modelo se aplicaron los algoritmos genéticos (GAs) en dos etapas:

1) Se corrió el método de los algoritmos genéticos para cada bloque de descriptores moleculares de manera separada y se seleccionaron los mejores descriptores por cada bloque.
2) Posteriormente, se fusionaron los mejores descriptores obtenidos en cada bloque (407 variables) y se aplicó nuevamente los GAs de tal forma de obtener el mejor modelo QSPR.

Con el objetivo de seleccionar el modelo más adecuado, se tomó en consideración los resultados obtenidos tanto en el conjunto de calibración como en el de predicción, teniendo en cuenta el coeficiente de correlación (R^2) más alto durante la validación cruzada. Así, se obtuvo un modelo óptimo con 8 descriptores moleculares para la predicción del índice de retención de los VOCs:

$$I = -4301.91 + 5614.01\ ATSC1e + 20.97\ P\_VSA\_PPPA - 241.85\ B03[O\text{-}O] + 5908.61\ PDI - 88.88\ Eig09\_AEA(dm) + 264.33\ C\text{-}005 - 38.32\ T(O\text{-}O) - 643.61\ MACCSFP142$$

3.5 Validación del modelo

Para la validar del modelo QSPR se aplicaron diferentes criterios de validación cruzada: dejar-uno-fuera, ventanas venecianas (con 5 grupos), Bloques continuos (con 5 grupos), Monte Carlo con el 20% de exclusión y 1000 iteraciones, Boostrap con 1000 iteraciones. Asimismo, se consideraron los mismos parámetros para el conjunto de calibración y para el conjunto de predicción. Además, se aplicó el criterio de aleatorización-Y con mil iteraciones con el objetivo de verificar la estabilidad del modelo. Este enfoque indica que, al cambiar aleatoriamente la respuesta, se obtendrá un modelo con mínima capacidad predictiva. Los resultados del modelo se muestran en la Tabla 1.

Tabla 1 Resultados de la validación del modelo con los criterios de calibración, validación y predicción.

	R^2	*RMSE*
Calibración	0.893	127.395
Validación cruzada		
Ventanas venecianas	0.855	148.108
Bloques continuos	0.847	152.454
Monte Carlo	0.856	148.633
Bootstrap	0.834	159.245
Dejar-uno-afuera	0.867	141.822
Predicción	0.908	118.228
Aleatorización-Y	0.077	391.055

Se puede observar en la Figura 2 una tendencia lineal y creciente en torno a la línea de ajuste perfecto de los valores de los índices de retención experimental con relación a los índices de retención predichos por el modelo. Este comportamiento indica que el ajuste del modelo entre los puntos experimentales y predichos para el grupo de calibración y predicción es adecuado y conveniente.

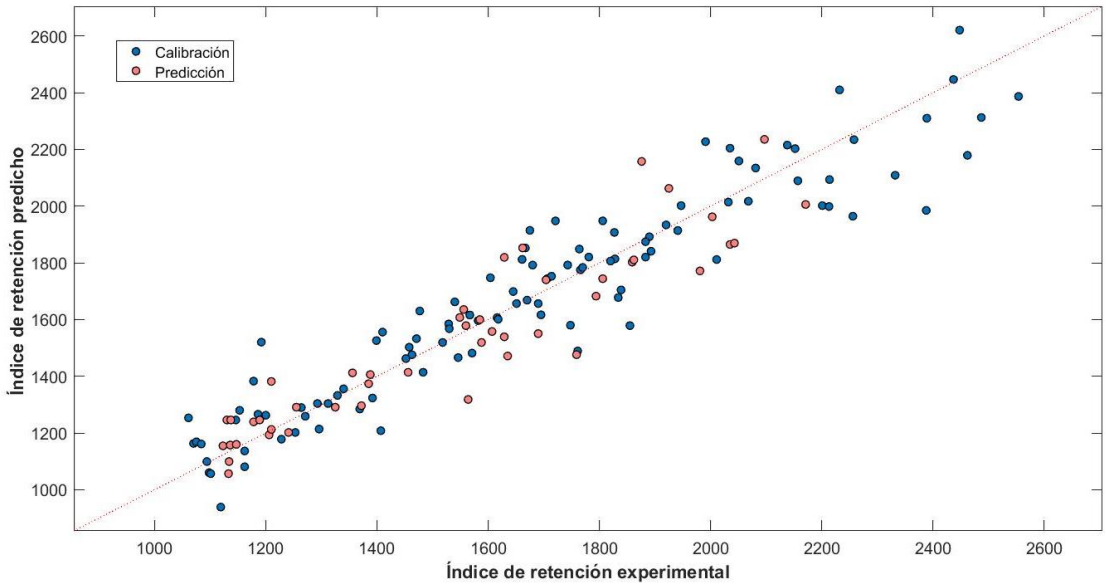

Figura 2 Índices de retención experimental versus índices de retención predichos del modelo *in silico* para los compuestos orgánicos volátiles de la miel de abejas.

Complementariamente, la Figura 3 muestra los residuos estandarizados versus los índices de retención predichos. Aquí se puede observar que los residuos están guiados por un modelo aleatorio alrededor del valor cero, lo que indica que el modelo es apropiado para modelar los índices de retención de los compuestos orgánicos volátiles de la miel de abejas. Al definir un intervalo de ± 2 RMSC (error cuadrático medio de calibración) se identificaron tres compuestos que presentan valores altos de residuo y que se consideran datos atípicos: 2,3-dihidro-benzofurano, 1-(2-furanil)-etanona y 1-fenil-pentan-2-ol.

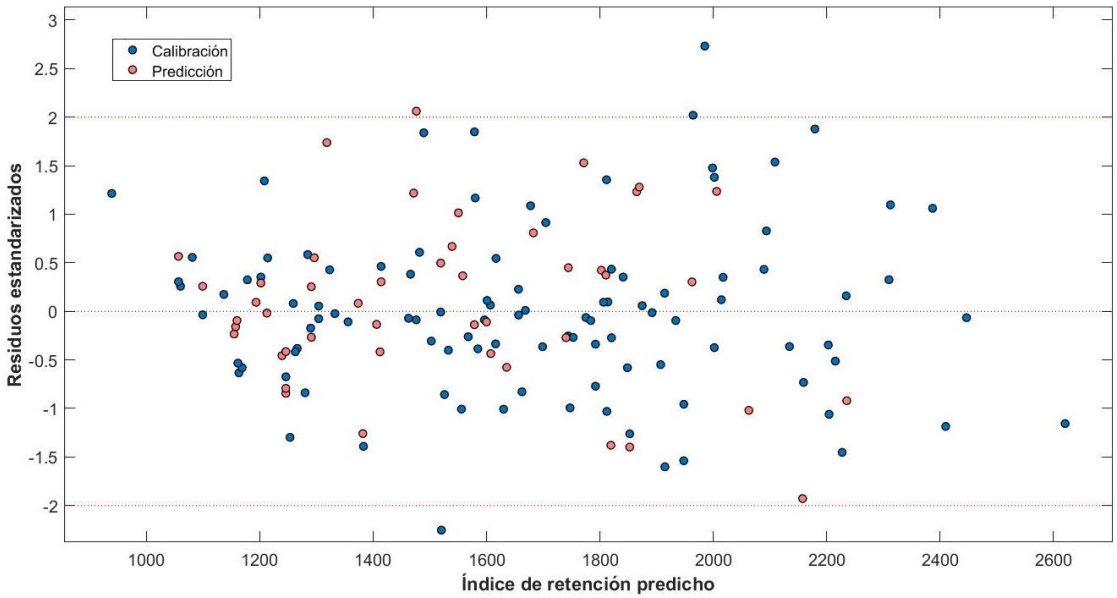

Figura 3 Índices de retención predichos versus residuos estandarizados del modelo *in silico* para los compuestos orgánicos volátiles de la miel de abejas.

3.6 Grado de contribución de los descriptores moleculares

Para analizar el nivel de contribución de los descriptores moleculares del modelo *in silico* para predecir el índice de retención (*I*), se calcularon los coeficientes estandarizados de regresión. La contribución de los descriptores es la siguiente: *P_VSA_PPP_A* (1.037) > *PDI* (0.942) > *T(O..O)* (0.342) > *C–005* (0.335) > *ATSC1e* (0.286) > *B03[O–O]* (0.234) > *MACCSFP142* (0.226) > *Eig09AEA(dm)* (0.162). Cada uno de los descriptores contienen información que explica características particulares de retención de los compuestos orgánicos volátiles presentes en la miel de abejas. En la Figura 4 se presenta gráficamente cada uno de los descriptores moleculares y su nivel de contribución para el modelo.

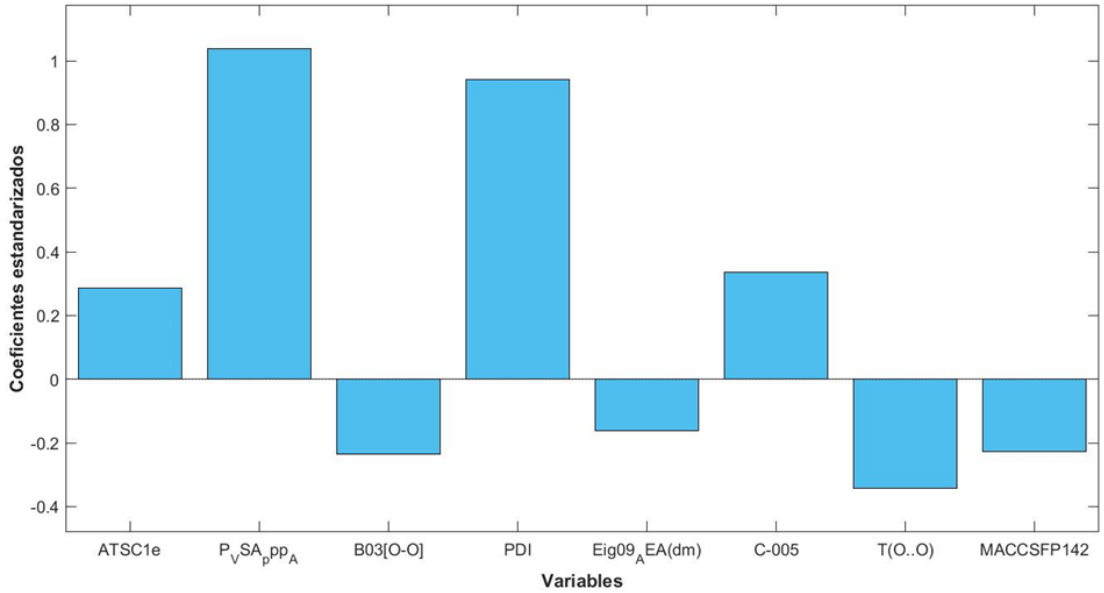

Figura 4 Representación gráfica del grado de contribución de los descriptores moleculares para los compuestos orgánicos volátiles presentes en muestras de miel de abejas.

Para obtener información de cada uno de los descriptores se utilizó el manual de usuario del programa alvaDesc (Alvascience, 2023a) y la información disponible en el libro "*Molecular Descriptors for Chemoinformatics*" (Todeschini & Consonni, 2009).

El descriptor potenciales puntos farmacóforos aceptor de enlaces de hidrogeno (*P_VSA_PPP_A)* pertenece al bloque de descriptores tipo P_VSA. Estos descriptores muestran la dimensión del área superficial de van der Waals (VSA) que tiene una propiedad *P* en un rango determinado. Estos descriptores pertenecen a una división de la superficie molecular limitada por los valores atómicos de la propiedad *P*. Por otra

parte, los potenciales puntos fármacos (PPP) son representaciones de cadenas tridimensionales de moléculas que codifican la información de una distancia geométrica entre todas las combinaciones posibles de puntos farmacóforos, por lo que resultan tener similitudes con los pares geométricos de propiedades de enlace. Los potenciales puntos farmacóforos son donantes y aceptores de enlaces de hidrógeno, sitios de interacción potencial de carga negativa, positiva y átomos hidrófobos. Todos los átomos de la molécula son analizados con el objetivo de identificar si pueden clasificarse potencialmente como uno de estos tipos. Los átomos pueden ser asignados a uno o más tipos de punto farmacóforo. Cada estructura molecular se reduce a sus puntos farmacóforos y se calculan las distancias entre todos esos puntos geométricos. A cada combinación se le asigna un número en la cadena de bits correspondiente a un rango de valores de distancias (Todeschini & Consonni, 2009).

Por otra parte, el descriptor índice de densidad de empaquetado (*PDI*) pertenece al bloque de propiedades moleculares. Este bloque engloba a descriptores moleculares híbridos que definen propiedades fisicoquímicas y biológicas, además de características específicas obtenidas a partir de modelos teóricos. *PDI* se encuentra dentro de los índices de forma Ciobutariu, los cuales se definen en términos de volumen de *Van der Waals* (V^{vdw}) y de la superficie molecular de *Van der Waals* (SA^{vdw}) (Ciubotariu et al., 2004). El índice de densidad de empaquetamiento se puede describir como:

$$R^{vdw} = \frac{V^{vdw}}{SA^{vdw}}$$

que representa una medida estérica/forma de una molécula.

Además, para moléculas acíclicas, se propusieron dos índices de globularidad. El primero se define como

$$G^{LOB} = \frac{R^{vdw}}{R^{s}}$$

donde R^s representa la relación entre el volumen y la superficie de una esfera equivalente, que rodea a la molécula, cuyo radio es igual a la mitad de dimensión más larga del paralelepípedo que envuelve la molécula. El segundo índice de globularidad se define como

$$G^{LEL} = \frac{V^{EL}}{V^{S}}$$

Donde V^{EL} es el volumen del elipsoide que rodea a toda la molécula y V^{s} es el volumen de una esfera con un radio igual a la mitad del eje elipsoidal más largo.

Las dos medidas de globularidad disminuyen con el crecimiento de las cadenas lineales y aumentan hacia la unidad cuando la molécula está muy ramificada o compactada. Este descriptor se define como un criterio de aromaticidad basado en la deslocalización de electrones. Este índice está estrechamente relacionado con los índices NICS y HOMA: cuantos mayores sean los índices PDI mayor es el valor absoluto de NICS y mayores son los valores de HOMA, reflejando así mayor aromaticidad (Todeschini & Consonni, 2009).

El descriptor suma de distancias topológicas entre átomos de oxígeno, *T(O..O)*, hace referencia al conjunto de distancias topológicas entre estos pares de átomos en el esqueleto químico. Este descriptor pertenece al bloque de pares de átomos topológicos ponderados, que se definen como la suma de las distancias topológicas entre los pares para diversos tipos de átomos:

$$T(u,v) = \frac{1}{2} \cdot \sum_{i=1}^{|V|} \sum_{j=1}^{|V|} \delta(i;u).\delta(j;v).d_{ij}$$

donde *u* y *v* hacen referencia a dos tipos de átomos diferentes, d_{ij} es la diferencia topológica entre el átomo i y el átomo j, y δ(i;u), δ(j;v) son dos funciones delta de Kronecker iguales a uno si el átomo i es de tipo u, el átomo j es de tipo v, respectivamente, caso contrario, será cero.

El descriptor *C-005* pertenece al bloque de fragmentos centrados en el átomo (Ghose, Viswanadhan, & Wendoloski, 1998; Viswanadhan, Ghose, Revankar, & Robins, 1989), que realizan conteos de las distintas fracciones (muestras de átomos característicos) existentes en la estructura molecular. Aquí, cada parte es un átomo en la molécula definido por sus átomos vecinos. Específicamente, el descriptor *C-005*, perteneciente a las constantes atómicas hidrofóbicas de Ghose-Crippen, se describe como CH_3X con un valor de log *P* de -1.788 y la refractividad molar (MR) 3.015, en donde el modelo para el cálculo de log *P* está definido como:

$$\log P = \sum_k a_k \,.\, N_k$$

donde N_k es el número de apariciones del *k*-ésimo tipo de átomos, y a_k las constantes atómicas hidrofóbicas que miden la contribución lipofílica de los átomos en la molécula (Ghose & Crippen, 1986; Ghose, Pritchett, & Crippen, 1988; Ghose et al., 1998; Viswanadhan et al., 1989).

Por otra parte, el descriptor autocorrelación centrada de Broto-Moreau a desplazamiento 1 ponderada por la electronegatividad de Sanderson, *ATSC1e*, pertenece al bloque de autocorrelaciones 2D (Todeschini & Consonni, 2009). Estas autocorrelaciones definen como una propiedad en específico se divide a lo largo de la estructura molecular topológica. Son autocorrelaciones que consideran un grafo molecular completo (incluyen hidrógenos) ponderado por las propiedades fisicoquímicas establecidas con relación al valor del átomo de carbono, es decir, estas autocorrelaciones pueden verse como un caso especial de las matrices geodésicas de interacción, de estas también se pueden derivar otros tipos de descriptores y es la autocorrelación espacial más conocida. Se debe tomar en cuenta que las propiedades atómicas deben centrarse disminuyendo el valor medio de la propiedad en la molécula para obtener valores de autocorrelación adecuados. Se ha demostrado que si las propiedades están centradas, todos los descriptores de autocorrelación no están correlacionados (Hollas, 2002) demostró que

El descriptor presencia/ausencia de átomos de oxígeno a una distancia topológica 3 *B03[O–O]* son un caso particular de los descriptores de autocorrelación de tipo átomo (*ATAC*) (Carhart, Smith, & Venkataraghavan, 1985). Los *ATAC* son descriptores discretos que dan cuenta de la presencia/ausencia de pares de átomos, o conteo de sus ocurrencias (frecuencia), donde los pares de átomo se definen como cualquier par de átomos que tienen características particulares a una distancia topológica predefinida *K*. Se define como:

$$ATAC_k(u,v) = \frac{1}{2} \,.\, \sum_{i=1}^{|V|} \sum_{j=1}^{|V|} \delta(i;u).\delta(j;v).\delta(d_{ij};k)$$

La clave estructural *MACCSFP142* corresponde a la presencia de átomos de nitrógeno, representados mediante el SMILES de especificación de objetivo arbitraria SMARTS "[#7]". SMARTS es un lenguaje que le permite especificar subestructuras utilizando

reglas que son extensiones sencillas de SMILES. Utilizando SMARTS, se pueden realizar especificaciones de búsqueda de subestructuras flexibles y eficientes en términos de significatividad química. En los SMARTS se utilizan símbolos atómicos y de enlace para especificar un gráfico. Sin embargo, en SMARTS las etiquetas de los nodos y bordes del gráfico (sus "átomos" y "enlaces") se amplían para incluir operadores lógicos y símbolos atómicos y de enlace especiales; estos permiten que los átomos y enlaces SMARTS sean más generales (Daylight Chemical Information Systems, 2007).

Finalmente, el descriptor autovalor n. 9 de la matriz de adyacencia aumentada y ponderada por el momento dipolar, *Eig09_AEA(dm)*, pertenece al bloque de índices de adyacencia de borde. Los valores correspondientes a la matriz de adyacencia pueden utilizarse como descriptores moleculares (Todeschini & Consonni, 2009).

3.7 Dominio de aplicabilidad

El dominio de aplicabilidad permite determinar el espacio químico teórico en el cual el modelo calibrado desarrolla predicciones confiables de los índices de retención. Para el dominio de aplicabilidad se utilizó el cálculo del valor de influencia o leverage. Aquí se fijó un valor de influencia crítico $h^* = 0.2143$ (Figura 5), el cual indica que cualquier compuesto orgánico volátil del grupo de predicción que se encuentre por debajo del mismo forma parte del dominio de aplicabilidad, como resultado se puede determinar que la predicción de *I* es confiable. En este trabajo se identificó que únicamente una molécula perteneciente al grupo de predicción se encuentra fuera de la región del dominio de aplicabilidad (2-hidroxibenzoato de pentilo) por lo que su *I* predicho es una extrapolación del modelo y, por consiguiente, es poco confiable.

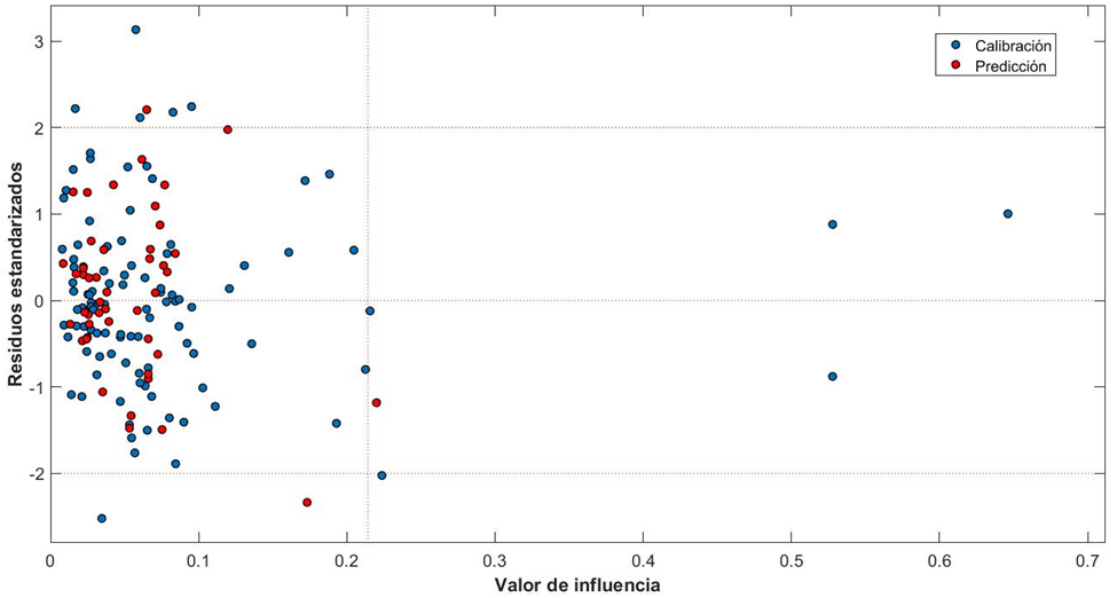

Figura 5 Gráfico de Williams que define el dominio de aplicabilidad del modelo *in silico* para la predicción del índice de retención de los VOCs.

CONCLUSIONES

En este trabajo de titulación se calibró un modelo quimioinformático considerando las relaciones cuantitativas estructura-propiedad para los índices de retención de los compuestos orgánicos volátiles presentes en miel de abeja de diferente origen botánico. La propiedad experimental fue medida utilizando la cromatografía de gases acoplada a espectrometría de masa. El modelo computacional consta de 8 descriptores moleculares independientes de la conformación obtenidos mediante el acoplamiento de la reducción no supervisada y selección supervisada de variables. El modelo muestra una capacidad de ajuste del 93% y capacidad predictiva del 90.8%. El trabajo permite concluir que un modelo independiente de la conformación es adecuado para la predicción de los índices de retención de compuestos orgánicos volátiles presentes en muestras de miel de abejas, por lo que constituye una base para próximas investigaciones en el área de la química computacional de los alimentos.

REFERENCIAS

Acevedo, J., Carrasco, R., & Basterrechea, M. (2010). Aplicación integrada de diferentes metodologías para la predicción del índice de retención en cromatografía gaseosa. *Revista Cubana de Química, 22*(1), 9-16.

Alimentarius, C. (1981). Codex Norma para la miel. *Codex stan*, 12-1981.

Alvascience. (2023a). alvaDesc (Software for Molecular Descriptors Calculation) version 2.0.16, https://www.alvascience.com.

Alvascience. (2023b). alvaMolecule (Software to View and Prepare Chemical Datasets) version 2.0.6, https://www.alvascience.com.

Arlot, S., & Celisse, A. (2010). A survey of cross-validation procedures for model selection.

Ballabio, D., & Consonni, V. (2013). Classification tools in chemistry. Part 1: linear models. PLS-DA. *Analytical Methods, 5*(16), 3790-3798.

Ballabio, D., Consonni, V., Mauri, A., Claeys-Bruno, M., Sergent, M., & Todeschini, R. (2014). A Novel Variable Reduction Method Adapted from Space-Filling Designs. *Chemometrics and Intelligent Laboratory Systems, 136*, 147-154.

Baroni, M. V., Nores, M. L., Díaz, M. D. P., Chiabrando, G. A., Fassano, J. P., Costa, C., & Wunderlin, D. A. (2006). Determination of volatile organic compound patterns characteristic of five unifloral honey by solid-phase microextraction− gas chromatography− mass spectrometry coupled to chemometrics. *Journal of Agricultural and Food Chemistry, 54*(19), 7235-7241.

Barra, M. G., Ponce-Díaz, M. C., & Venegas-Gallegos, C. (2010). Volatile compounds in honey produced in the central valley of Ñuble province, Chile. *Chilean journal of agricultural research, 70*(1), 75-84.

Baumann, K., & Stiefl, N. (2004). Validation tools for variable subset regression. *Journal of Computer-Aided Molecular Design, 18*, 549-562.

Bertoncelj, J., Doberšek, U., Jamnik, M., & Golob, T. (2007). Evaluation of the phenolic content, antioxidant activity and colour of Slovenian honey. *Food chemistry, 105*(2), 822-828.

Carhart, R. E., Smith, D. H., & Venkataraghavan, R. (1985). Atom pairs as molecular features in structure-activity studies: definition and applications. *Journal of chemical information and computer sciences, 25*(2), 64-73.

Ciubotariu, D., Medeleanu, M., Vlaia, V., Olariu, T., Ciubotariu, C., Dragos, D., & Corina, S. (2004). Molecular van der Waals space and topological indices from the distance matrix. *Molecules, 9*(12), 1053-1078.

Cuevas-Glory, L. F., Ortiz-Vázquez, E., Centurión-Yah, A., Alea, J. A. P., & Sauri-Duch, E. (2008). Desarrollo de un método por microextracción en fase sólida para el análisis de la fracción volátil de la miel de abeja de Yucatán. *Técnica Pecuaria en México, 46*(4), 387-395.

Daylight Chemical Information Systems, I. (2007). SMARTS-A Language for Describing Molecular Patterns. In.

Dearden, J. C. (2016). The History and Development of Quantitative Structure-Activity Relationships (QSARs). *International Journal of Quantitative Structure-Property Relationships, 1*(1), 1-44.

Durant, J. L., Leland, B. A., Henry, D. R., & Nourse, J. G. (2002). Reoptimization of MDL keys for use in drug discovery. *Journal of chemical information and computer sciences, 42*(6), 1273-1280.

Efron, B. (1979). Bootstrap methods: another look at the jackknife annals of statistics 7: 1–26. *View Article PubMed/NCBI Google Scholar, 24*.

Efron, B. (1982). *The jackknife, the bootstrap and other resampling plans*: SIAM.

Efron, B. (1987). Better bootstrap confidence intervals. *Journal of the American statistical Association, 82*(397), 171-185.

Ghose, A. K., & Crippen, G. M. (1986). Atomic physicochemical parameters for three-dimensional structure-directed quantitative structure-activity relationships I. Partition coefficients as a measure of hydrophobicity. *Journal of Computational Chemistry, 7*(4), 565-577.

Ghose, A. K., Pritchett, A., & Crippen, G. M. (1988). Atomic physicochemical parameters for three dimensional structure directed quantitative structure-activity relationships III: Modeling hydrophobic interactions. *Journal of Computational Chemistry, 9*(1), 80-90.

Ghose, A. K., Viswanadhan, V. N., & Wendoloski, J. J. (1998). Prediction of hydrophobic (lipophilic) properties of small organic molecules using fragmental methods: an analysis of ALOGP and CLOGP methods. *The Journal of Physical Chemistry A, 102*(21), 3762-3772.

Goode-Romero, G., Aguayo-Ortiz, R., & Domínguez, L. (2019). Relaciones cuantitativas estructura-actividad/propiedad en dos dimensiones empleando el programa R. *Educación química, 30*(2), 27-40.

Hansch, C., & Leo, A. (1995). *Exploring QSAR.: Fundamentals and applications in chemistry and biology* (Vol. 1): American Chemical Society.

Hollas, B. (2002). Correlation properties of the autocorrelation descriptor for molecules. *MATCH Commun. Math. Comput. Chem, 45*, 27-33.

Jaworska, J., Nikolova-Jeliazkova, N., & Aldenberg, T. (2005). QSAR Applicability Domain Estimation by Projection of the Training Set Descriptor Space: A Review. *ATLA, 33*(5), 445–459.

Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., . . . Yu, B. (2023). PubChem 2023 update. *Nucleic acids research, 51*(D1), D1373-D1380.

Krakowska, B., Custers, D., Deconinck, E., & Daszykowski, M. (2016). The Monte Carlo validation framework for the discriminant partial least squares model extended with variable selection methods applied to authenticity studies of Viagra® based on chromatographic impurity profiles. *Analyst, 141*(3), 1060-1070.

Kubat, M. (2017). *An introduction to machine learning*: Springer.

Leardi, R. (2020). Genetic Algorithms in Chemistry. In S. Brown, R. Tauler, & B. Walczak (Eds.), *Comprehensive Chemometrics: Chemical and Biochemical Data Analysis* (Second ed., Vol. 1, pp. 617-634). The Netherlands: Elsevier.

Leardi, R., & Gonzalez, A. L. (1998). Genetic algorithms applied to feature selection in PLS regression: how and when to use them. *Chemometrics and Intelligent Laboratory Systems, 41*(2), 195-207.

Lindgren, F., Hansen, B., Karcher, W., Sjöström, M., & Eriksson, L. (1996). Model validation by permutation tests: applications to variable selection. *Journal of chemometrics, 10*(5-6), 521-532.

Machado, A. M., Miguel, M. G., Vilas-Boas, M., & Figueiredo, A. C. (2020). Honey volatiles as a fingerprint for botanical origin- A review on their occurrence on monofloral honeys. *Molecules, 25*(2), 374.

Mauri, A., Consonni, V., & Todeschini, R. (2017). Molecular descriptors. In *Handbook of computational chemistry* (pp. 2065-2093): Springer International Publishing.

O'Donnell, T. (2008). *Design and use of relational databases in chemistry*: CRC Press.

Rencher, A. C., & Schaalje, G. B. (2008). *Linear models in statistics*: John Wiley & Sons.

Rojas-Monroy, G. M. (2005). *Caracterización del aroma del café molido de Puerto Rico mediante la técnica de microextracción en fase sólida (spme) y cromatografía de gas acoplada a espectrometría de masas (gc/ms).*

Rojas, C., & Duchowicz, P. R. (2021). *Química Computacional de los Alimentos: Relaciones Cuantitativas Estructura-Actividad/Propiedad (QSAR/QSPR)*: Editorial Acribia, S.A.

Rojas, C., Duchowicz, P. R., Tripaldi, P., & Pis Diez, R. (2015a). QSPR Analysis for the Retention Index of Flavors and Fragrances on a OV-101 Column. *Chemometrics and Intelligent Laboratory Systems, 140*, 126-132.

Rojas, C., Duchowicz, P. R., Tripaldi, P., & Pis Diez, R. (2015b). Quantitative Structure–Property Relationship Analysis for the Retention Index of Fragrance-Like Compounds on a Polar Stationary Phase. *Journal of Chromatography A, 1422*, 277-288.

Roy, K., Kar, S., & Das, R. N. (2015a). *A primer on QSAR/QSPR modeling: fundamental concepts*: Springer.

Roy, K., Kar, S., & Das, R. N. (2015b). *Understanding the basics of QSAR for applications in pharmaceutical sciences and risk assessment*: Academic press.

Schomburg, G. (1990). *Gas chromatography: A practical course*: VCH.

Soria, A. C., Martínez-Castro, I., & Sanz, J. (2008). Some aspects of dynamic headspace analysis of volatile components in honey. *Food Research International, 41*(8), 838-848.

Soria, A. C., Sanz, J., & Martínez-Castro, I. (2009). SPME followed by GC–MS: a powerful technique for qualitative analysis of honey volatiles. *European Food Research and Technology, 228*, 579-590.

Sparkman, O. D., Penton, Z., & Kitson, F. G. (2011). *Gas chromatography and mass spectrometry: a practical guide*: Academic press.

Stashenko, E. E., & Martínez, J. R. (2010). Algunos aspectos prácticos para la identificación de analitos por cromatografía de gases acoplada a espectrometría de masas. *Scientia Chromatographica, 2*(1), 29-47.

Suescún, L., & Vit, P. (2008). Control de calidad de la miel de abejas producida como propuesta para un proyecto de servicio comunitario obligatorio. *Fuerza farmacéutica, 12*(1), 6-15.

The MathWorks Inc. (2022). MATLAB R2022b, http://www.mathworks.com.

Todeschini, R., & Consonni, V. (2009). *Molecular Descriptors for Chemoinformatics* (Second ed.). Germany: Wiley-VCH Verlag GmbH & Co.

Todeschini, R., Consonni, V., Ballabio, D., & Grisoni, F. (2020). 4.25-chemometrics for QSAR modeling. In *Comprehensive Chemometrics: Chemical and Biochemical Data Analysis, Second Edition: Four Volume Set* (Vol. 4, pp. 599-634): Elsevier.

Vazhacharickal, P. J. (2021). A review on health benefits and biological action of honey, propolis and royal jelly. *J. Med. Plants Stud, 9*(5), 1-13.

Viswanadhan, V. N., Ghose, A. K., Revankar, G. R., & Robins, R. K. (1989). Atomic physicochemical parameters for three dimensional structure directed quantitative structure-activity relationships. 4. Additional parameters for hydrophobic and dispersive interactions and their application for an automated superposition of certain naturally occurring nucleoside antibiotics. *Journal of chemical information and computer sciences, 29*(3), 163-172.

Wang, X., Chen, Y., Hu, Y., Zhou, J., Chen, L., & Lu, X. (2022). Systematic review of the characteristic markers in honey of various botanical, geographic, and entomological origins. *ACS Food Science & Technology, 2*(2), 206-220.

Wishart, D. (2014). FooDB: the food database. *FooDB version, 1*.