UNIVERSIDAD DEL AZUAY

Facultad de Ciencia y Tecnología

Ingeniería en Alimentos

Estudio de la percepción del umbral de olor mediante relaciones cuantitativas estructura-propiedad

Trabajo de graduación previo a la obtención del título de:

INGENIERA EN ALIMENTOS

Autor:

María Elisa Pacheco Jaramillo

Director:

Dr. Cristian Rojas Villa

Co-Director:

Dr. Piercosimo Tripaldi

CUENCA-ECUADOR

2019

DEDICATORIA

El presente trabajo de titulación lo dedico a mis padres: Marcelo y Liuba que con su esfuerzo, dedicación y enseñanzas me permitieron alcanzar esta meta; de igual manera a mis hermanos Marcela y Emilio, que con su motivación incondicional pude culminar toda esta travesía.

AGRADECIMIENTOS

Agradezco infinitamente a los profesores de la carrera de Ingeniería en Alimentos, por las enseñanzas y conocimientos impartidos durante todos estos años, en especial a Dr. Cristian Rojas Villa y al Dr. Piercosimo Tripaldi, que con su paciencia y dedicación han transmitido nuevos e innovadores conocimientos sobre mí.

A Lenin por estar a mi lado en todo momento, motivándome y ayudándome en lo que sea posible, gracias por todas tus enseñanzas y apoyo durante la carrera, me doy cuenta de que soy muy afortunada por tener a una persona tan maravillosa como tú a mi lado, además has sido un pilar fundamental en mi vida y en la culminación de esta carrera.

A mis tías y a mi abuelita que han sido un eje fundamental en el transcurso de mi vida, y han sabido guiar mi camino en los buenos y malos momentos.

A todas mis amigas del “Barrio Chino”, gracias por todo chicas, sin su apoyo y amistad incondicional desde el colegio hasta el último ciclo de universidad, no hubiera podido cumplir este objetivo. A mis amigos y compañeros de la carrera de Ingeniería en Alimentos que, con su gran apoyo y amistad, conseguimos finalizar esta etapa de nuestras vidas. A mis amigas Karla y Rafa, gracias por su amistad de todos estos años, siempre pendientes en todos los aspectos de mi vida, gracias por todo ese apoyo en estos cuatro años de universidad. A mis amigas Doménica Victoria y Katty gracias por todos los buenos momentos que pasamos en el laboratorio.

A la Universidad del Azuay, a toda su directiva, por formar parte de mi proceso de formación académica.

Estudio de la percepción del umbral de olor mediante relaciones cuantitativas estructura-propiedad

RESUMEN

Este trabajo desarrolló un modelo predictivo *in silico* para el umbral de olor de 176 compuestos orgánicos volátiles (VOCs) basado en una relación cuantitativa estructura-propiedad (QSPR). Las moléculas se optimizaron mediante el método semiempírico PM3 para calcular 5440 descriptores moleculares en el programa alvaDesc. Inicialmente, se usó el método V-WSP para la reducción de descriptores. Posteriormente, el conjunto de datos se dividió en grupos de calibración (70%) y predicción (30%). Los compuestos se dividieron en moléculas de alto y bajo poder odorante para modelarlos mediante el método de clasificación *k*NN (*k*-vecinos más cercanos) acoplado con los algoritmos genéticos (GAs). Se obtuvo un modelo con 2 descriptores y 5 vecinos cercanos, utilizando tasa de aciertos en calibración (NER_{cal} = 0.75), validación cruzada (NER_{cv} = 0.75) y predicción (NER_{pred} = 0.78). El modelo se desarrolló con los principios de la Organización para la Cooperación y el Desarrollo Económico (OCDE) para hacerlo aplicable.

Palabras Clave: Umbral de olor, VOCs, QSPR, *k*NN

Dr. Cristian Roias Villa

Director de tesis

Ing. María Fernanda Rosales

Coordinadora de la Escuela Ingeniería en Alimentos

María Elisa Pacheco Jaramillo

Autora

Analysis of odor threshold perception through a quantitative structure-property relationship

ABSTRACT

A predictive *in silico* model for the odor threshold of 176 volatile organic compounds (VOCs) was developed based on a quantitative structure-property relationship (QSPR). All the molecules were optimized by means of the semiempirical PM3 method to calculate 5440 molecular descriptors using the alvaDesc software. Initially, the V-WSP method was used to reduce the initial pool of descriptors. Subsequently, the dataset was divided into training (70%) and validation (30%) sets. Compounds were divided into high and low odorants in order to be modeled by means of the *k*NN (*k*-Nearest Neighbors) classifier coupled with the genetic algorithms (GAs). The optimal model (two descriptors and five neighbors) was developed using the non-error rate in calibration (NER_{cal} = 0.75), cross-validation (NER_{cv} = 0.75) and prediction (NER_{pred} = 0.78). To make it applicable, this QSPR model was developed in compliance with the principles of the Organisation for Economic Cooperation and Development (OECD).

Keywords: Odor Threshold, VOCs, QSPR, *k*NN

Dr. Cristian Roias Villa
Thesis Director

Ing. María Fernanda Rosales
Food Engineering Faculty Coordinator

María Elisa Pacheco Jaramillo
Author

UNIVERSIDAD DEL AZUAY
Dpto. Idiomas

Translated by
Ing. Paúl Arpi

ÍNDICE DE CONTENIDOS

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

María Elisa Pacheco Jaramillo
Trabajo de Titulación
Dr. Cristian Rojas Villa
Dr. Piercosimo Tripaldi
Octubre, 2019

Estudio de la percepción del umbral de olor mediante relaciones cuantitativas estructura-propiedad

INTRODUCCIÓN

Umbral de olor

En términos generales, se define al umbral de olor como la concentración más baja de gas o vapor de un material que puede detectarse por su olor. Si un compuesto no es muy potente y se necesita una gran cantidad de él para que se detecte, se concluye que el compuesto está presente en una concentración muy baja; por lo que no es probable que contribuya a un aroma característico. Sin embargo, si el compuesto es potente y se necesita una concentración muy pequeña para su detección, este se encuentra en una concentración muy alta y por consiguiente es muy probable que contribuya a un aroma característico (Sharon S. Murnane, Alex H. Lehocky, & Patrick D. Owens, 1989).

Identificación del Umbral del Olor

Umbral de Detección o Reconocimiento

El umbral de reconocimiento es la concentración más baja de olor que provoca una respuesta sensorial en el ser humano. Los valores de umbral no son parámetros fisiológicos fijos o constantes físicas, sino puntos estadísticos que representan el mejor valor de estimación a partir de un grupo de respuestas individuales (Sharon S. Murnane et al., 1989).

Detección del olor

Las mediciones de intensidad de olor son difíciles de realizar, debido al hecho de que no se disponen de instrumentos analíticos específicos, con la excepción de la nariz electrónica, la cual puede variar su efectividad a causa de la humedad del ambiente y por la cantidad de sensores que lo conforman. Por este motivo, hay que tener en cuenta que el valor del umbral de olor se define por diversos factores, tales como temperatura, presión, ausencia /presencia de otras sustancias. Por esta razón, una de las técnicas utilizadas es el método de dilución hasta el umbral (Dilution-to-threshold method), en el que se diluye una muestra de un compuesto aromático con aire inodoro en varios niveles y la serie de dilución se presenta en orden ascendente de concentración de olor. De un nivel a otro, la dilución disminuye y la

cantidad del compuesto volátil aumenta. Los primeros niveles incluyen la muestra diluida con una gran cantidad de aire inodoro para que la evaluación pueda comenzar por debajo del umbral de detección. Finalmente, el valor del umbral de olor se puede expresar en concentraciones de agua, aire, entre otros solutos. Algunas sustancias pueden detectarse cuando su concentración es de solo unos pocos miligramos por cada 1000 toneladas. También se toma en cuenta que el valor del umbral de olor de un compuesto está influenciado por el medio (Powers, 2004). Existen otros métodos para la detección de olor, entre los que se encuentran: método de elección forzada, olfatometría y la nariz electrónica.

Método de elección forzada

Para detectar el umbral de olor, los miembros entrenados del panel reciben muestras de compuestos orgánicos volátiles (VOCs: Volatile Organic Chemicals) entre muestras limpias (inoloras). Los evaluadores sensoriales deben identificar la presencia de un olor. El umbral de detección es el nivel en el que un panelista puede diferenciar entre el odorante diluido y la muestra limpia (Sharon S. Murnane et al., 1989).

Olfatometría

Los olfatómetros son instrumentos que, utilizados con un sujeto humano, detectan y miden los olores ambientales. Un operador controla la entrega de la muestra mientras el sujeto de prueba inhala a través del puerto de rastreo del equipo *The AS'CENT International Olfactometer*® (Figura 1) para detectar la presencia de olor (Sharon S. Murnane et al., 1989).

Figura 1. Olfatómetro The AS'CENT International Olfactometer® (Powers, 2004)

Nariz Electrónica

En los últimos años se ha utilizado la nariz electrónica como una alternativa para la detección del umbral de olor en las industrias de alimentos, bebidas y perfumes. Esta ha sido desarrollada para imitar el sentido del olfato humano y se utiliza para el desarrollo de productos y el control de calidad (Sharon S. Murnane et al., 1989).

Este equipo consta de un conjunto de sensores que detectan los compuestos químicos que los humanos perciben como olores y los registra como resultados numéricos. El instrumento generará un patrón diferente de respuesta para diferentes tipos de muestras. Las narices electrónicas disponibles comercialmente tienen 32, 64 o 128 sensores. Cada sensor tiene una respuesta característica individual y algunos de ellos se superponen y son sensibles a sustancias químicas similares, al igual que los receptores en la nariz humana. Un solo sensor responde parcialmente a una amplia gama de productos químicos y es más sensible a una estrecha gama de compuestos. Por otro lado, múltiples sensores en un solo instrumento proporcionan una respuesta confiable a una gran cantidad de VOCs, ya que contienen sensores sensibles para compuestos específicos. La selección del sensor es crítica, tanto desde el punto de vista de la sensibilidad a los compuestos que contribuyen olores desagradables (mal olor) como a la respuesta y durabilidad de los sensores en ambientes húmedos (Powers, 2004).

Modelos para detectar el umbral de olor

Últimamente se han desarrollado y refinado técnicas de modelado para determinar el umbral de olor; para ello, se han propuesto ecuaciones que correlacionan los umbrales de olor para así ayudar a estimar los umbrales en ausencia de mediciones de olor reales. Se determinaron los coeficientes de correlación por encima de 0.7 hasta 0.9, para todos estos modelos incluyendo así varias clases de compuestos orgánicos volátiles. Uno de los modelos utilizados para detectar el umbral de olor de diferentes VOCs se basa en las fases de gas condensado, este método tiene diferentes variables independientes como son la refractividad molar en exceso de soluto, la dipolaridad/polarizabilidad del soluto, la acidez y basicidad del enlace de hidrógeno y coeficiente de partición gas/hexadecano. Todas estas variables independientes se obtienen a partir de datos experimentales. En general, los modelos para estimar los umbrales de detección han sido significativos y las investigaciones recientes continúan aumentando la validez de los mismos (Sharon S. Murnane et al., 1989).

Relaciones Cuantitativas Estructura-Propiedad

Las relaciones cuantitativas estructura-propiedad (QSPR: quantitative structure-property relationship) son modelos matemáticos asistidos por computadora que relacionan propiedades fisicoquímicas de los compuestos con su estructura molecular. Los estudios QSPR se realizan sobre la base de la correlación entre los valores experimentales y los descriptores que se derivan de la estructura molecular de los compuestos respectivos. Los modelos QSPR se usan para predecir las propiedades de otros compuestos moleculares para los cuales no se ha medido experimentalmente dicha propiedad (Roy, Kar, & Das, 2015). El enfoque QSPR ha sido ampliamente utilizado por varios grupos de investigación para la predicción del umbral de olor (Edwards, Anker, & Jurs, 1991),(Xu, Luan, Liu, & Gao, 2011),(Pal, Mitra, & Roy, 2013),(Pal, Mitra, & Roy, 2014),(Toropov, Toropova, Cappellini, Benfenati, & Davoli, 2016),(Ojha & Roy, 2018),(Belhassan, Chtita, Lakhlifi, & Bouachrine,

2019). En consecuencia, los modelos QSPR constituyen una alternativa válida para la detección de potenciales compuestos orgánicos volátiles a partir de grandes bases de datos, lo que reduce la medición experimental de este universo grande de componentes aromáticos.

Objetivos de una relación cuantitativa estructura-propiedad

Los modelos QSPR tienen como finalidad correlacionar las propiedades estructurales, fisicoquímicas y la actividad biológica de diferentes compuestos, mediante varios enfoques, desarrollando un modelo matemático con estas propiedades. Este modelado utiliza los datos disponibles y permite predecir un gran número de compuestos. Esto proporciona una oportunidad para que esta técnica sea utilizada en varios campos (Roy et al., 2015). Entre los objetivos principales se encuentran:

1. Predicción de la actividad/propiedad/toxicidad a partir de las respuestas de compuestos químicos estructuralmente similares.
2. Reducción y reemplazo de experimentación en laboratorio, especialmente en animales.
3. Búsqueda virtual en bibliotecas de datos.
4. Mecanismo de acción de los descriptores moleculares incluidos en un modelo; es decir, la naturaleza de la información descriptiva codificada por los mismos.
5. Categorización de datos.
6. Optimización de nuevas estructuras moleculares con propiedades deseadas.
7. Refinamiento estructural de moléculas objetivo sintéticas.

Componentes y etapas de modelo QSPR

Los pasos principales en los estudios de QSPR incluyen (Bassaganya-Riera, 2018):

1. Selección de un conjunto de datos y cálculo de los descriptores moleculares
2. Selección de variables
3. Construcción del modelo
4. Validación y predicción de los datos
5. Interpretación datos

Principios del modelado QSPR

Para garantizar que el modelo QSPR sea confiable, la Organización para la Cooperación Económica y el Desarrollo OECD (Organisation for Economic Cooperation and Development) ha establecido cinco principios básicos que se deben cumplir (Organisation for Economic Co-operation and Development, 2014):

1. Propiedad definida: se refiere a cualquier efecto fisicoquímico, biológico o ambiental que pueda medirse y modelarse.
2. Algoritmo inequívoco: el propósito de este principio es garantizar la transparencia en la descripción del algoritmo del modelo.

3. Definición del dominio de aplicabilidad: expresa el hecho de que los modelos QSPR son modelos reduccionistas que se asocian con limitaciones en términos de los tipos de estructuras químicas, propiedades fisicoquímicas y mecanismos de acción para los cuales los modelos pueden generar predicciones confiables.
4. Medidas apropiadas de ajuste, robustez y predictividad.
5. Interpretación del mecanismo de acción (si es posible): la intención de este principio brindar una explicación de cómo los descriptores de un modelo se relacionan con la propiedad.

Modelos QSPR relacionados al umbral de olor

En la literatura existen reportados varios trabajos relacionados al modelado *in silico* del umbral de olor con modelos. Edwards y colaboradores (1991) realizaron un análisis QSPR para el umbral de olor de compuestos activos constituidos por alcoholes y pirazinas. En este estudio se concluyó que el método de regresión lineal múltiple (MLR: multiple linear regression) puede utilizarse satisfactoriamente para modelar esta propiedad.

Varios años después, Xu y colaboradores (2011) desarrollaron dos modelos para predecir el umbral de olor de alcoholes alifáticos. El primer modelo es basado en regresión lineal múltiple (MLR), mientras que el segundo se basa en las redes neuronales de función de base radial (RBFNN: radial basis function neural networks). Estos autores concluyeron que el modelo RBFNN muestra una mejor capacidad predictiva.

Por otra parte, Pal y colaboradores (2013) realizaron un modelo predictivo del umbral de olor para una serie de 74 pirazinas, utilizando el algoritmo de aproximación de la función genética (GFA: Genetic function approximation algorithm). El modelo desarrollado se usó para diseñar 27 derivados de la pirazina con valores predichos aceptables de umbral de olor (OT: odor threshold). Al año siguiente, estos mismos autores (Pal et al., 2014) establecieron una relación entre los datos de umbral de olor de 53 alcoholes alifáticos y sus atributos estructurales utilizando los índices del átomo topoquímico ampliado (ETA: Extended Topochemical Atom). El modelo desarrollado sugiere que la ramificación molecular y los parámetros electrónicos influyen significativamente en la potencia del olor de las moléculas, por lo que se puede usar para la predicción de esta propiedad en otros alcoholes alifáticos.

Más adelante, Toropov y colaboradores (2016) utilizaron el formato SMILES (Simplified molecular-input line-entry system) de 1259 compuestos orgánicos volátiles para desarrollar un modelo QSPR en el programa CORAL (CORrelation And Logic). Los datos fueron divididos tres veces en grupos de calibración (80%) y validación (20%) y se utilizó la media aritmética de los diferentes grupos para obtener los valores de umbral de olor.

Recientemente, Ojha y Roy (2018) establecieron un modelado QSPR para el OT de 85 compuestos aromáticos presentes en diferentes tipos de vinos. Como representación de la estructura molecular se utilizaron descriptores 2D y 3D, con los cuales se aplicaron diferentes estrategias de selección de variables para obtener un modelo de regresión de mínimos cuadrados parciales (PLS). El modelo *in silico* fue ampliamente validado en términos de la aceptabilidad y predictividad. Finalmente, Belhassan y colaboradores (2019) utilizaron el umbral de olor de 78 derivados de la pirazina para desarrollar un análisis 2D-QSPR basado en MLR. El modelo obtenido proporcionó una descripción detallada de los aspectos moleculares que subyacen a la propiedad de los umbrales del olor de los compuestos estudiados.

CAPÍTULO I

METODOLOGÍA

Materiales y Métodos

Generación de la base de datos

Los datos numéricos sobre el umbral de olor se tomaron del libro *"Compilation of odor and taste threshold values data"* (Stahl, 1973), la base de datos está compuesta de 1238 compuestos orgánicos volátiles para los cuales se reporta el umbral de olor y sabor. Las unidades de los umbrales son diferentes en algunos casos; por ejemplo, se reportan en partes por millón (ppm), miligramos por litro (mg/L), partes por billón (ppb), gramos por mililitro (g/mL), entre otras. Posteriormente, se procedió al filtrado de la base de datos a través del programa KNIME (Berthold et al., 2008), programando nodos que permitan realizar las operaciones que se detallan en la Figura 2 y se describen a continuación:

1. Separar entre los compuestos medidos en agua o aire. De esta forma se obtienen 627 moléculas medidas en agua.
2. Separar las moléculas para las cuales se reporta el umbral de olor de las que se reporta el umbral de sabor (taste threshold). Se obtienen así 348 compuestos con umbral de olor.
3. Filtrar esta vez por el tipo de umbral de olor: detección, reconocimiento, no específico y otros. De esta manera se obtiene 313 compuestos orgánicos volátiles en la categoría de detección como tipo de umbral de olor.
4. Estandarizar las unidades de medida del umbral de olor a partes por millón. De esta forma, se convirtieron las unidades de partes por billón de 85 moléculas y otras unidades distintas (g/mL, mg/L, molal) de 5 compuestos.
5. Finalmente, a partir de los 237 VOCs se identificaron moléculas duplicadas por nombre o número de registro CAS, para fusionarlas y obtener 176 moléculas que serán utilizadas para el desarrollo del modelo QSPR.

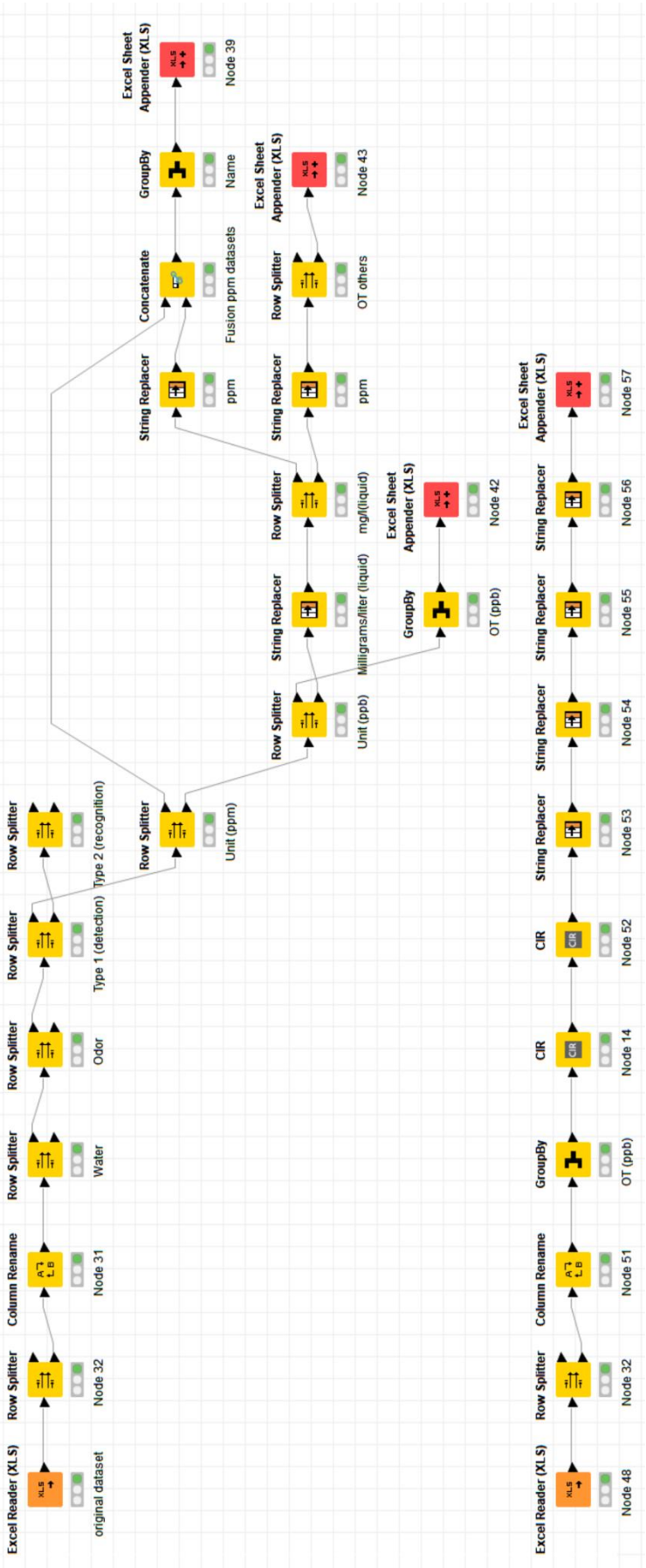

Figura 2. Diagrama de flujo KNIME para el curado de la base de datos de umbral de olor

Representación de la estructura molecular

Los compuestos orgánicos volátiles fueron representados en el programa HyperChem (Hypercube Inc.), donde inicialmente se optimizaron mediante los campos de fuerza de la mecánica molecular (MM+) y posteriormente se refinaron sus coordenadas mediante el método semiempírico PM3. En ambos casos, se usó el algoritmo del gradiente conjugado en la versión Polak-Ribiere y las geometrías se consideraron optimizadas cuando la desviación estándar del vector gradiente es menor a 0.01 kcal×(Å×mol)$^{-1}$.

Mecánica molecular

La mecánica molecular (MM+) es una herramienta para la generación de estructuras de las moléculas, en la que los átomos se representan mediante partículas (esferas) dotadas de masa y carga, y los enlaces mediante resortes. Este método construye una expresión para la energía potencial en función de las posiciones atómicas en el eje cartesiano XYZ, mediante el uso las constantes de fuerza de alargamiento del enlace y la introducción de términos que permiten considerar interacciones entre los átomos no enlazados. El método MM+ minimiza (optimiza) la energía potencial para distintos confórmeros mediante la modificación de las posiciones atómicas (Cuevas & Cortés, 2003).

Método semiempírico PM3

El método semiempírico PM3 (Parametric Method Number 3) se utiliza para cálculos cuánticos de la estructura electrónica. Se basa en el método semiempírico AM1 (Austin Model 1), pero con la diferencia que toma algunas integrales como parámetros a optimizar y utiliza funciones gaussianas como orbitales atómicos. El método PM3 ha sido parametrizado para la mayoría de elementos presentes en la tabla periódica (Cuevas & Cortés, 2003).

Cálculo de descriptores moleculares

La representación optimizada de cada compuesto se utilizó para el cálculo de los descriptores moleculares a través del programa alvaDesc (Alvascience Srl, 2019).Los descriptores moleculares son el resultado final de un procedimiento lógico y matemático que transforma la información química codificada dentro de una representación simbólica de una molécula en un número útil o también es el resultado de un experimento estandarizado (Todeschini & Consonni, 2009).

Reducción no supervisada de descriptores moleculares

Los métodos de reducción no supervisados de variables seleccionan un subconjunto de variables capaces de preservar la información representativa que poseen los descriptores moleculares. Uno de estos métodos es el V-WSP, que es una modificación del algoritmo propuesto por Wootton, Sergent y Phan-Tan-Luu (WSP), que selecciona un subconjunto de

variables de manera que se encuentren a una distancia mínima de correlación fija de cada variable en el espacio multidimensional (Ballabio et al., 2014). Dada una matriz de datos con *n* filas y *p* columnas, el algoritmo V-WSP realiza los siguientes cálculos:

1. Elegir un descriptor inicial "*j*" y un umbral de correlación (*thr*).
2. Calcular los coeficientes de correlación lineal de Pearson (*R*) entre el *j*-ésimo descriptor y todos los demás *d* descriptores.
3. Eliminar los descriptores con valor absoluto de $R_{dj} \geq thr.$
4. Se fija el descriptor "*j*" y se selecciona entre los restantes descriptores aquel que tenga la correlación absoluta más alta con el mismo.
5. Repetir los pasos 2, 3 y 4 hasta que no existan descriptores para ser seleccionadas.

Selección supervisada de descriptores moleculares

Para la selección supervisada de los descriptores se recurrió a los algoritmos genéticos (GAs) (R. Leardi, 2009; Riccardo Leardi & Gonzalez, 1998), los cuales se basan en la teoría de la evolución de Darwin, es decir se selecciona individuos al azar de la población actual para ser padres y los usa para producir los hijos para la próxima generación de tal forma que, durante generaciones sucesivas, la población "evoluciona" hacia una solución óptima. Los GAs se aplican para resolver una variedad de problemas de optimización que no son adecuados para los algoritmos de optimización estándar, en los que la función objetivo es discontinua, no diferenciable, estocástica o altamente no lineal. Conjuntamente el algoritmo genético utiliza tres tipos principales de reglas en cada paso para crear la siguiente generación a partir de la población actual:

- Selección: seleccionan a los individuos, llamados padres, que contribuyen a la población en la próxima generación.
- Reproducción (crossover): combinan a dos padres para formar hijos para la próxima generación pueden ser codificadas por una secuencia de 0s y 1s.
- Mutación: aplican cambios aleatorios a los padres individuales para formar hijos.

Cuando los GAs se aplican a bases de datos que contienen grandes cantidades de variables, se puede correr el riesgo de que los modelos se sobreajusten, para ello, se realiza un número pequeño de corridas independientes a partir de varias poblaciones iniciales registrando la frecuencia de selección de variables. Entonces, el modelo se construye adicionando la variable más frecuente de cada población y agregando las demás variables en función de su frecuencia de selección.

Métodos de regresión

Estos métodos pueden tratar fácilmente matrices de datos muy grandes, extrayendo la parte relevante de la información y produciendo modelos confiables pero complejos. Tienen como objetivo predecir una variable continua "Y" a partir de una o varias variables explicativas "X"

además de describir la estructura común subyacente de las variables. Se utilizaron los métodos de regresión de mínimos cuadrados ordinarios (OLS: ordinary least squares) y mínimos cuadrados parciales (PLS: Partial Least Squares).

OLS

Es un método para estimar los parámetros desconocidos en un modelo de regresión lineal para variables de respuestas continuas que se han registrado en una escala de intervalo. Este método minimiza la suma de los cuadrados entre las respuestas observadas en el conjunto de datos y las respuestas predichas por la aproximación lineal (Hutcheson, 2011).

PLS

El método de mínimos cuadrados parciales es un método de regresión lineal que tiene algunas similitudes con el análisis de componentes principales (PCA: Principal Component Analysis). Este método trabaja con proyecciones de los objetos en un nuevo espacio definido por combinaciones lineales de las variables originales denominadas variables latentes (LVs) y que contemporáneamente presenten la máxima correlación con la respuesta (Wold, Sjöström, & Eriksson, 2001).

Métodos de Clasificación

Método de los *k* vecinos más cercanos

El método de los *k*-vecinos más cercanos (*k*NN: *k*-Nearest Neighbors) (Cover & Hart, 1967) es un método no lineal y robusto que clasifica un objeto según la mayoría de sus *k* vecinos más cercanos en el espacio de datos. Para ello, se calcula la distancia del elemento con respecto a los demás, y ordena dichas distancias de menor a mayor para ir seleccionando el grupo al que pertenece. Este grupo será, por tanto, el de mayor frecuencia con menor distancia.

Al aplicar *k*NN, se debe buscar el valor óptimo de *k*, una opción para seleccionar *k* es mediante procedimientos de validación cruzada, es decir, probando un conjunto de valores *k*; luego, el valor de *k* que da el error de clasificación más bajo en validación cruzada puede seleccionarse como el óptimo. Debido a estas características, se ha sugerido que *k*NN es un método comparativo estándar para técnicas de clasificación. Debido a que se debe dicotomizar la variable umbral de olor y encontrar un valor óptimo de umbral, se utilizó el método Simplex para optimizar la separación de las clases. Los resultados se expresan en tasa de aciertos (NER: Non Error Rate), sensibilidad (*Sn:* sensitivity), y especificidad (*Sp*: specificity).

Método simplex

El método de optimización simplex (Turina, 1986) está basado en el principio de movimiento de paso a paso hacia el objetivo, estableciendo cambios simultáneo de varias variables. Este método evalúa la función objetivo en la solución y con esto permite determinar si esta solución

es óptima o no; en caso de no ser óptima el algoritmo recorre los vértices del polígono de soluciones posibles realizando un proceso iterativo hasta obtener el valor que maximiza o minimiza la función objetivo.

$$X_i = X_{1,c} + \alpha(X_{1,c} - X_{1,1})$$

(Ec.1)

Donde:

X_i= Coordenada del umbral

$X_{1,c}$= *X* centroide de los dos mejores resultados

α = Factor de Expansión (0,5)

$X_{1,1}$= Coordenada del umbral desechado

Validación del Modelo

Para validar el modelo se construyó un conjunto de calibración y predicción, esta división se debe hacer de tal forma que se mantenga una relación estructura-propiedad en los dos conjuntos. Es decir, que las moléculas del grupo de calibración sean representativas del grupo de predicción (Veerasamy et al., 2011). Es de esperar que un modelo QSPR se desempeñe mejor para moléculas homogéneas, ya que funcionará interpolando en lugar de hacer extrapolaciones de la propiedad. Así, la base de datos se dividió en dos grupos:

- Grupo de calibración (training set): conjunto de datos utilizados para ajustar el modelo, estos datos reales se utilizan para entrenar el modelo, es decir, el modelo ve y aprende de los mismos.
- Grupo de predicción (test set): el conjunto de predicción se utiliza para evaluar la capacidad predictiva de un modelo, por lo que no deben ser considerados durante la calibración del mismo. El mérito de un modelo QSPR radica en su capacidad predictiva más que en su capacidad de ajuste.

La partición se realizó de forma aleatoria y proporcional a la numerosidad de las clases con un algoritmo programado en MATLAB (The MathWorks Inc.).

Grado de contribución y mecanismo de acción de los descriptores moleculares

Debido a que el método de clasificación *k*NN no proporciona coeficientes para cuantificar la contribución de cada descriptor, la interpretación de los mismos se llevó a cabo en función de la definición de cada descriptor molecular y como estos se relacionan con el modelo de umbral de olor. Proporcionar una interpretación de los mecanismos de acción de los descriptores de un modelo QSPR no siempre es posible, debido principalmente a la complejidad en la definición del descriptor; es decir, las teorías detrás de ellos en algunos casos son abstractas. Esta interpretación del mecanismo de acción contribuye al conocimiento de cómo los descriptores moleculares describen el fenómeno del umbral de olor (Roy et al., 2015).

Dominio de aplicabilidad del modelo

El dominio de aplicabilidad (AD: Applicability Domain) se define como una región teórica en el espacio químico construido por los descriptores del modelo y la respuesta modelada. El dominio de aplicabilidad juega un papel crucial para estimar la incertidumbre en la predicción de un determinado compuesto basado en qué tan similar es con respecto a los compuestos empleados para construir el modelo QSPR (Roy et al., 2015). Se puede identificar un compuesto fuera del dominio de aplicabilidad cuando la distancia de este compuesto con respecto al centro del modelo (conjunto de datos de calibración) excede un valor de umbral definido.

En los métodos de clasificación, uno de los enfoques para definir el AD del modelo es el basado en la similitud *k*NN (Sahigara, Ballabio, Todeschini, & Consonni, 2013; Sahigara et al., 2012). En este método se utilizan las distancias promedio de las moléculas del grupo de predicción con respecto a los *k* vecinos más cercanos del grupo de calibración. Posteriormente, se comparan dichas distancias con un umbral predefinido de la siguiente manera: si la distancia promedio es menor al umbral, dicha molécula estará dentro del AD del modelo y su predicción será confiable, caso contrario su predicción constituirá una extrapolación del modelo.

CAPÍTULO II

RESULTADOS

Generación base de datos

Los datos de umbral de olor y las moléculas se recopilaron del libro *"Compilation of odor and taste threshold values data"* (Stahl, 1973). De los cuales las moléculas con las que se trabajó fueron 176. Para cada molécula, se verificó el número de registro (CAS: Chemical Abstracts Service) y la notación lineal de cadena SMILES canónicos en las bibliotecas digitales PubChem (Kim et al., 2015), NIST (National Institute of Standards and Technology) (Linstrom & Mallard, 2001), ChemSpider (Pence & Williams, 2010) y CIR (Chemical Identifier Resolver) (NCI/CADD Group, 2019). En la Figura 3 se detalla a través de un diagrama de flujo como se realizó el curado de la información.

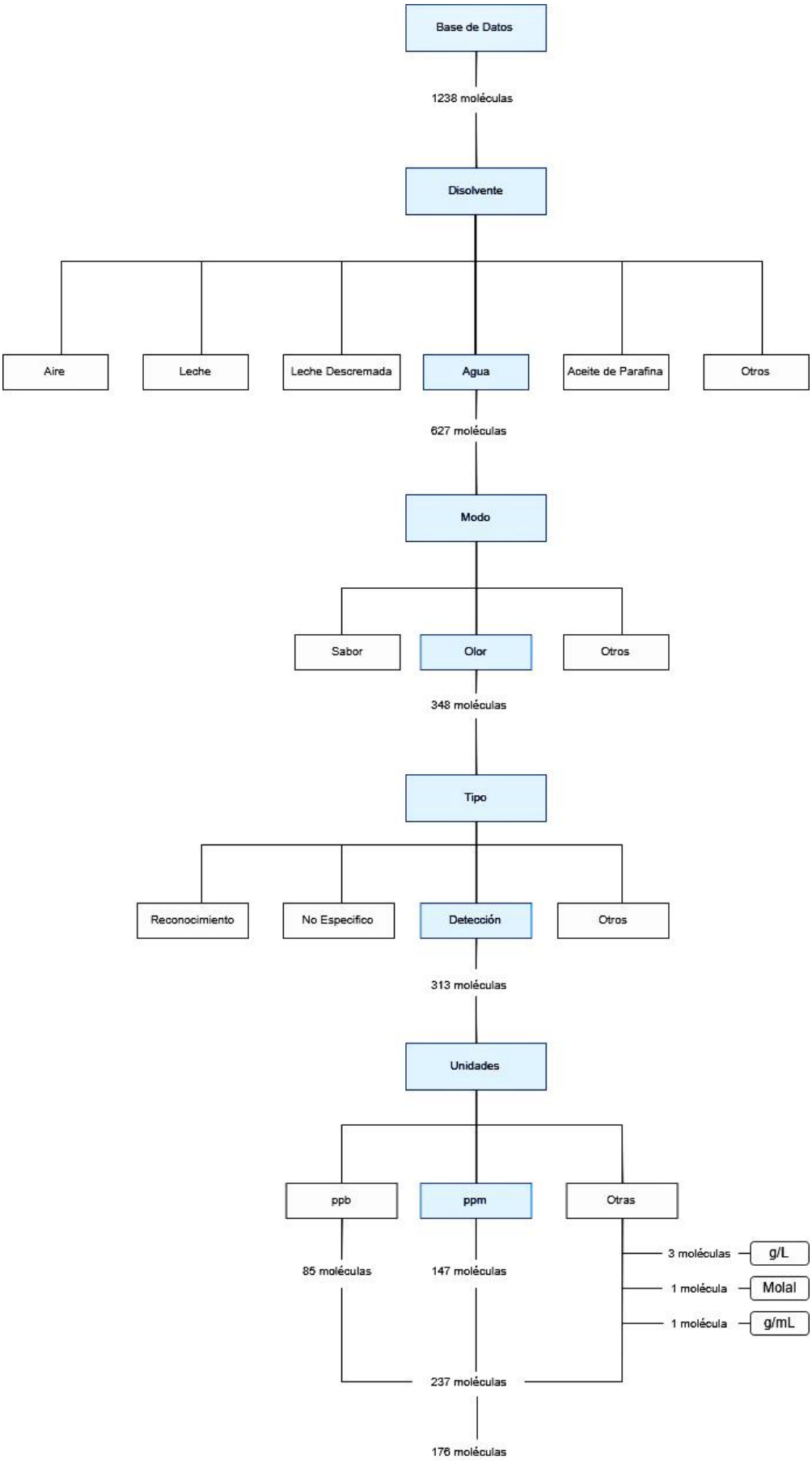

Figura 3. Diagrama de flujo de curado de la información

Representación de la estructura molecular y cálculo de descriptores moleculares

Las 176 moléculas fueron modeladas a través del programa HyperChem (Hypercube Inc.), donde las geometrías se optimizaron mediante los campos de fuerza de la mecánica molecular (MM+) y el método semiempírico PM3. Posteriormente, para cada compuesto orgánico volátil se calcularon 5274 descriptores moleculares en el programa alvaDesc (Alvascience Srl, 2019). En una primera etapa de filtración se excluyeron 1852 descriptores constantes, 2026 casi constantes y 150 con al menos un valor faltante. Adicionalmente, se calcularon 166 huellas dactilares moleculares MACCS (Molecular ACCess System). Estos grupos de descriptores y huellas MACCS se fusionaron para generar una base de datos de 3263 variables.

Reducción no supervisada de descriptores moleculares

Con los 3263 descriptores se utilizó el método V-WSP (Ballabio et al., 2014) implementado en MATLAB, con un umbral de correlación de 0.95 (*thr*=0.95). Este método seleccionó un subconjunto representativo de 1512 descriptores de tal forma que se encuentren a una distancia mínima de correlación en el espacio multidimensional. De esta manera es posible reducir el costo computacional durante el desarrollo del modelo.

Validación del Modelo

Para los modelos de regresión se realizó una división de la base de datos de manera aleatoria en grupos se calibración y predicción, con el 70% y 30 % de los compuestos, respectivamente. Para la validación de los modelos de clasificación, se consideró la misma proporción 70%/30%, con un algoritmo programado en MATLAB que mantenga la proporcionalidad entre las dos clases, compuestos de alto poder odorante son (Clase 1) y bajo poder odorante (Clase 2).

Selección Supervisada de Descriptores Moleculares

Modelos de regresión

En una primera etapa de modelado *in silico*, se usaron los datos de umbral de olor (OT) y el logaritmo del umbral de olor log (OT) del grupo de calibración para desarrollar modelos basados en mínimos cuadrados ordinarios y mínimos cuadrados parciales, acoplados con los algoritmos genéticos como estrategia de selección de variables. Durante la selección basada en los GAs, se optimizó el coeficiente de determinación en validación cruzada de ventanas venecianas. Posteriormente, se evaluó la calidad predictiva de cada modelo mediante la estimación de la propiedad para el conjunto de predicción. De esta manera se obtuvieron 6 modelos QSPR, cuyos parámetros se presentan en la Tabla 1.

Tabla 1. Resultados de los modelos de regresión QSPR para el umbral de olor

Modelo	d	LVs	R^2_{Cal}	RMSEC	R^2_{cv}	$RMSEC_{cv}$	R^2_{pred}	RMSEP
OT-OLS	2	*	0,801	120,88	0,715	144,62	0,299	226,71
OT-PLS	2	2	0,806	119,29	0,711	145,44	0,357	217,13
log(OT)-OLS	10	*	0,694	0,967	0,644	1,04	0,529	1,2
log(OT)-PLS	10	2	0,671	1,003	0,624	1,072	0,558	1,163
log(1/OT)-OLS	10	*	0,694	0,967	0,644	1,043	0,529	1,2
log(1/OT)-PLS	11	3	0,671	1,003	0,611	1,09	0,651	1,032

d: número de descriptores en el modelo; **LVs:** número de variables latentes; $\mathbf{R^2_{cal}}$ **:** coeficiente de determinación en calibración; **RMSEC:** error cuadrático medio de calibración ; $\mathbf{R^2_{cv}}$ **:** coeficiente de determinación en validación cruzada; $\mathbf{RMSEC_{cv}}$ **:** error cuadrático medio de validación cruzada; $\mathbf{R^2_{pred}}$ **:** coeficiente de determinación en predicción; **RMSEP:** error cuadrático medio de predicción.

Como se puede observar los resultados de R^2 de calibración son aceptables, sin embargo, los resultados en predicción son poco satisfactorios. Por tal motivo, se optó por desarrollar un modelo de clasificación mediante la discretización del log (OT).

Modelo de Clasificación

Para el desarrollo del modelo de clasificación *k*NN se usaron particiones del log (OT) para dos y tres clases. Se utilizó el algoritmo *k*NN acoplado a algoritmos genéticos, utilizando el autoescalado de los datos y la distancia euclidiana como tipo de medida.

En cuanto a la partición en tres clases, se aplicó el método de optimización Simplex en dos dimensiones, dividiendo los VOCs en clases de compuestos con alto poder odorante (Clase 1), compuestos con medio poder odorante (Clase 2) y compuestos con bajo potencial odorante (Clase 3); de tal forma que un valor bajo del log (OT) significa un odorante más potente (Ojha & Roy, 2018). Para este propósito se usaron los percentiles 33.3 y 66.6 como primer punto, para el segundo punto se optó por los percentiles 25 y 75, mientras que para el tercero se consideraron los percentiles 40 y 80. A partir de estos puntos iniciales se desarrollaron los modelos de clasificación *k*NN, buscando optimizar la tasa de aciertos NERcv. Los siguientes puntos se obtuvieron mediante el algoritmo de optimización simplex (Ec.1), buscando así la mejor solución (NER_{cv} máximo). Los resultados de la optimización Simplex se detallan en la Tabla 2.

Tabla 2.Resultados de la optimización Simplex para los modelos de clasificación *k*NN con tres clases

Modelo	*k*	d	NER_{cal}	NER_{cv}	NER_{pred}
1	3	21	0,67	0,68	0,55
2	8	8	0,61	0,65	0,63
3	6	7	0,57	0,53	0,61
4	3	5	0,56	0,53	0,59
5	3	3	0,57	0,55	0,57
6	2	8	0,60	0,61	0,55

Al observar que los resultados de predicción de la partición en tres clases no son altos, se procedió a utilizar la clasificación en dos clases en una sola dimensión del método Simplex para verificar si es posible mejorar los resultados de predicción. El procedimiento seguido es el mismo detallado previamente para las tres clases, con la diferencia de que ahora los compuestos de alto poder odorante son (Clase 1) y bajo poder odorante (Clase 2). Para este propósito en el primer punto se utilizó únicamente el percentil 33.33 y para el segundo punto el percentil 44. Los resultados se presentan en la Tabla 3.

Tabla 3. Resultados de la optimización Simplex para los modelos de clasificación *k*NN con dos clases

Modelo	*k*	d	NER_{cal}	NER_{cv}	NER_{pred}
1	4	8	0,78	0,72	0,61
2	9	5	0,67	0,73	0,73
3	3	5	0,81	0,81	0,74
4	**5**	**2**	**0,75**	**0,75**	**0,78**

Los resultados presentados en la Tabla 3 indican que el modelo con el mayor poder predictivo es el número 4 (NER_{pred} = 0.78). Al ser baja la diferencia entre los resultados de calibración y predicción, se descarta la presencia de sobreajuste. En consecuencia, el modelo QSPR basado en clasificación *k*NN está constituido por dos descriptores moleculares y usa 5 vecinos (*k*=5) durante la clasificación. Los descriptores de este modelo se describen en la Tabla 4.

Tabla 4. Detalle de los descriptores moleculares incluidos en el modelo *k*NN

Nombre	Descripción	Bloque
SsOH	Suma de estados electrónicos de sOH	Índices del estado Electrotopológico por tipo de átomo
CATS2D_00_DD	CATS2D Donante-Donante a distancia topológica 00	Descriptores farmacóforos

Grado de contribución y mecanismo de acción de los descriptores moleculares

Como describimos anteriormente, el método de clasificación *k*NN no proporciona coeficientes para cuantificar la contribución de cada descriptor, por lo que la interpretación del descriptor se realizó sobre la base de la definición de cada descriptor molecular.

Los índices del estado electrotopológico por tipo de átomo (Atom-type E-state indices) (Hall & Kier, 2000) son descriptores moleculares que combinan información estructural sobre la accesibilidad electrónica asociada con cada tipo de átomo. Indica la presencia o ausencia de un tipo de átomo específico. También se calculan mediante la suma de los estados electrotopológicos de los átomos pertenecientes a un mismo tipo dentro de la molécula. Así, el descriptor SsOH indica la suma de grupos funcionales OH^- en la estructura química de los compuestos orgánicos volátiles.

Por otro lado, los descriptores de búsqueda de plantillas químicamente avanzadas (CATS: Chemically Advanced Template Search) (Fechner, Franke, Renner, Schneider, & Schneider, 2003; Renner, Fechner, & Schneider, 2006) se basan en la estructura bidimensional de una molécula, por lo que pertenecen a la categoría de descriptores de pares de átomos, además codifican la información topológica de una molécula. Estos están relacionados con la presencia de potenciales farmacóforos (PPP), donde un PPP es un tipo de átomo generalizado definido teniendo en cuenta aspectos fisicoquímicos. Los cinco PPP para el cálculo de descriptores CATS son: donante de enlaces de hidrógeno (D), aceptor de enlaces de hidrógeno (A), cargados positivamente o ionizables (P), con carga negativa o ionizable (N) y lipofílicos (L). De esta manera, el descriptor CATS2D_00_DD indica la presencia de donante de enlaces de hidrógeno (D) separados por una distancia topológica cero, es decir, átomos de oxígeno de un grupo OH o átomos de nitrógeno de un grupo NH o NH_2

Dominio de aplicabilidad del modelo

El dominio de aplicabilidad del modelo final, basado en la similitud de *k*NN, estableció un umbral definido con un valor 0.0645 lo que significa que la distancia promedio menor a este se encuentra dentro del dominio de aplicabilidad, por lo que su predicción es confiable. En contraste, solo 4 moléculas se encuentran fuera de este dominio al tener un valor mayor que

el umbral definido, las cuales describiremos a continuación en la Tabla 5, lo que significa que moléculas que excedían el umbral se consideraron como extrapolaciones del modelo.

Para encontrar el AD se utilizó el algoritmo basado en similitud *k*NN (Sahigara et al., 2013; Sahigara et al., 2012) ajustado con la similitud *k*NN, aplicando las siguientes opciones: auto escalado, distancia euclidiana, cinco vecinos y con validación cruzada de ventanas venecianas con 5 grupos.

Tabla 5. Moléculas fuera del dominio de aplicabilidad del modelo QSPR

Nombre	Distancia
3-metil-1-fenil-3-pentanol	0,0986
Ácido dicloroacético	0,0653
Pentionina (β,β-dimetil lantionina)	6,6636
Ácido propiónico	0,0665

CAPÍTULO III

DISCUSIÓN

El presente trabajo, desarrolló un modelo basado en las relaciones cuantitativas estructura-propiedad (QSPR) para el umbral de olor de 176 compuestos orgánicos volátiles (VOCs) medidos en agua y expresados en partes por millón (ppm). Al observar que los modelos de regresión otorgaban resultados muy bajos en predicción (Tabla 1), se optó por realizar una clasificación simplex en dos clases utilizando algoritmos genéticos (GAs) acoplados con el método de clasificación no paramétrico de los *k*-vecinos más cercanos (*k*NN). Entre estos cuatro modelos finales descritos en la Tabla 3, el modelo 4 mostró la mejor capacidad predictiva, utilizando dos descriptores y cinco vecinos cercanos, como lo demuestra el valor de NER_{pred}. En cuando al AD, el modelo dio como resultado 5 moléculas fuera del dominio que se consideraron como extrapolaciones (Tabla 5). Se debe tomar en cuenta que no se pueden comparar los resultados de modelos de regresión con el de clasificación desarrollado en esta investigación, debido a que toma diferentes parámetros para realizar el modelo final QSPR.

Tabla 6. Estudios sobre el umbral de olor

Referencia	Compuestos	Modelo	Sustrato	d	N_{cal}	N_{pred}	R^2_{cal}	S_{cal}	R^2_{cv}	S_{cv}	R^2_{pred}	S_{pred}
(Edwards et al., 1991)	Alcoholes	MLR	-	4	49	-	0.863	0.372	-	-	-	-
	Pirazinas		-	5	73	-	0.885	0.59	-	-	-	-
(Xu, Luan, Liu, & Gao, 2011)	Alcoholes Alifáticos	RBFNN	-	5	78	19	0.863	0.5231	-	-	0.76	0.5281
(Pal, Mitra, & Roy, 2013)	Derivados de pirazina	GFA-spline	Agua	2	52	22	0.783	-	0.729	-	0.822	-
(Pal et al., 2014)	Alcoholes Alifáticos	GFA-spline	Aire	3	42	11	0.809	-	0.778	-	0.813	-
(Toropov, Toropova, Cappellini, Benfenati, & Davoli, 2016)	Compuestos orgánicos volátiles heterogéneos	Regresión	Aire	4	967	290	0.62 ± 0.02	-	-	-	0.62±0.04	-

(Ojha & Roy, 2018)	Compuestos aromáticos del vino	PLS	Alcohol	5	64	21	0.841	0.726	0.812	-	0.923	-
(Belhassan, Chtita, Lakhlifi, & Bouachrine, 2019)	Derivados pirazina	MLR	Agua	4	57	21	0.682	-	0.606	-	0.672	-

Los modelos QSPR desarrollados en para el umbral de olor contemplan en su mayoría derivados de la pirazina, alcoholes alifáticos y diversos compuestos orgánicos volátiles. En la Tabla 6 se presentan los diversos modelos QSPR publicados en la literatura especializada. Es importante indicar que en todos estos estudios utilizaron bases de datos con pocos compuestos, lo que limita su generalización a otro tipo de VOCs. Por otro lado, los autores optan por utilizar modelos de regresión de mínimos cuadrados parciales (PLS), regresión lineal múltiple (MLR), algoritmo de aproximación de la función genética (GFA), redes neuronales de función de base radial (RBFNN), utilizando en su mayoría un máximo de 5 descriptores. En todos los modelos descritos se obtienen buenos resultados de R^2 en calibración y predicción, a excepción de los modelos de Toropov y colaboradores (2016) y Belhassan y colaboradores (2019) debido a los métodos utilizados, por ejemplo, el primero utiliza el software CORAL para su modelo de regresión además de poseer una base de datos grande, mientras que el segundo utiliza un modelo MLR, aunque sus resultados son buenos, comparados con los demás modelos son bajos.

Por otra parte, Edwards y colaboradores (1991) y Belhassan y colaboradores (2019) no presentan resultados del dominio de aplicabilidad de los modelos, lo cual constituye una limitante para su aplicación.

CONCLUSIÓN

Esta investigación ha desarrollado una relación cuantitativa predictiva para el umbral de olor de compuestos orgánicos volátiles (VOCs) de una base de datos de 176 moléculas, siguiendo los principios estipulados por la Organización para la Cooperación Económica y el Desarrollo (OECD). El método de optimización Simplex mostró un buen desempeño para determinar el umbral de separación de las clases a partir de una respuesta continua. Asimismo, los algoritmos genéticos (GAs) acoplados con el método de clasificación de los *k*-vecinos más cercanos (*k*NN) permitieron obtener un modelo con buena capacidad de ajuste, validación y predicción. En consecuencia, el modelo permite, por una parte, entender el mecanismo de acción de los descriptores moleculares para describir el comportamiento de esta propiedad, y por otra, predecir el poder odorante de nuevos compuestos no evaluados o no sintetizados.

REFERENCIAS BIBLIOGRÁFICAS

Alvascience Srl. (2019). alvaDesc (software for molecular descriptors calculation) version 1.0.12, https://www.alvascience.com.

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.

Bassaganya-Riera, J. (2018). *Accelerated Path to Cures*: Springer.

Belhassan, A., Chtita, S., Lakhlifi, T., & Bouachrine, M. (2019). 2D-QSPR Study of Olfactive Thresholds for Pyrazine Derivatives Using DFT and Statistical Methods. *Emerging Science Journal, 3*(3), 179-186.

Berthold, M. R., Cebron, N., Dill, F., Gabriel, T. R., Kötter, T., Meinl, T., . . . Wiswedel, B. (2008). KNIME: The Konstanz Information Miner. In C. Preisach, H. Burkhardt, L. Schmidt-Thieme & R. Decker (Eds.), *Data Analysis, Machine Learning and Applications* (pp. 319-326): Springer Berlin Heidelberg.

Cover, T. M., & Hart, P. E. (1967). Nearest neighbor pattern classification. *IEEE transactions on information theory, 13*(1), 21-27.

Cuevas, G., & Cortés, F. (2003). *Introducción a la química computacional*: Fondo de Cultura Económica.

Edwards, P. A., Anker, L. S., & Jurs, P. C. (1991). Quantitative structure-property relationship studies of the odor threshold of odor active compounds. *Chemical senses, 16*(5), 447-465.

Fechner, U., Franke, L., Renner, S., Schneider, P., & Schneider, G. (2003). Comparison of correlation vector methods for ligand-based similarity searching. *Journal of computer-aided molecular design, 17*(10), 687-698.

Hall, L. H., & Kier, L. B. (2000). The E-state as the basis for molecular structure space definition and structure similarity. *Journal of chemical information and computer sciences, 40*(3), 784-791.

Hutcheson, G. D. (2011). Ordinary least-squares regression. *L. Moutinho and GD Hutcheson, The SAGE dictionary of quantitative management research*, 224-228.

Hypercube Inc. HyperChem. http://www.hyper.com.

Kim, S., Thiessen, P. A., Bolton, E. E., Chen, J., Fu, G., Gindulyte, A., . . . Bryant, S. H. (2015). PubChem Substance and Compound databases. *Nucleic acids research, 44*(D1), D1202-D1213.

Leardi, R. (2009). Genetic Algorithms *Comprehensive Chemometrics* (pp. 631-653).

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.

Linstrom, P. J., & Mallard, W. G. (2001). *NIST Chemistry WebBook, NIST Standard Reference Database Number 69* Retrieved from http://webbook.nist.gov

NCI/CADD Group. (2019). Chemical Identifier Resolver, https://cactus.nci.nih.gov/chemical/structure.

Ojha, P. K., & Roy, K. (2018). Chemometric modeling of odor threshold property of diverse aroma components of wine. *RSC Advances, 8*(9), 4750-4760.

Organisation for Economic Co-operation and Development. (2014). Guidance Document on the Validation of (Quantitative)Structure-Activity Relationships [(Q)SAR] Models. OECD Series on Testing and Assessment, No. 69.

Pal, P., Mitra, I., & Roy, K. (2013). Predictive QSPR modelling for the olfactory threshold of a series of pyrazine derivatives. *Flavour and Fragrance Journal, 28*(2), 102-117.

Pal, P., Mitra, I., & Roy, K. (2014). QSPR modeling of odor threshold of aliphatic alcohols using extended topochemical atom (ETA) indices. *Croatica Chemica Acta, 87*(1), 29-37.

Pence, H. E., & Williams, A. (2010). ChemSpider: An Online Chemical Information Resource. *Journal of Chemical Education, 87*(11), 1123-1124.

Powers, W. (2004). The science of smell, part 3: odor detection and measurement. *Iowa State University Extension PM*.

Renner, S., Fechner, U., & Schneider, G. (2006). Alignment-free pharmacophore patterns-A correlation-vector approach. *Methods and Principles in Medicinal Chemistry, 32*, 49.

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

Sahigara, F., Ballabio, D., Todeschini, R., & Consonni, V. (2013). Defining a novel *k*-nearest neighbours approach to assess the applicability domain of a QSAR model for reliable predictions. *Journal of Cheminformatics, 5*(1), 27.

Sahigara, F., Mansouri, K., Ballabio, D., Mauri, A., Consonni, V., & Todeschini, R. (2012). Comparison of different approaches to define the applicability domain of QSAR models. *Molecules, 17*(5), 4791-4810.

Sharon S. Murnane, C., CSP, Alex H. Lehocky, M., CIH, & Patrick D. Owens, C., CSP. (1989). *Odor thresholds for chemicals with established occupational health standards*: Aiha.

Stahl, W. H. (1973). *Compilation of odor and taste threshold values data*.

The MathWorks Inc. MatLab, http://www.mathworks.com.

Todeschini, R., & Consonni, V. (2009). *Molecular Descriptors for Chemoinformatics* (Vol. 1). Weinheim: WILEY-VCH.

Toropov, A. A., Toropova, A. P., Cappellini, L., Benfenati, E., & Davoli, E. (2016). Odor threshold prediction by means of the Monte Carlo method. *Ecotoxicology and environmental safety, 133*, 390-394.

Turina, S. (1986). Optimization in Chromatographic Analysis. In R. E. Keiser (Ed.), *Planar chromatography Volume 1* (Vol. 1, pp. 15-46).

Veerasamy, R., Rajak, H., Jain, A., Sivadasan, S., Varghese, C. P., & Agrawal, R. K. (2011). Validation of QSAR models-strategies and importance. *International journal of drug design & discovery, 3*, 511-519.

Wold, S., Sjöström, M., & Eriksson, L. (2001). PLS-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems, 58*(2), 109-130.

Xu, X., Luan, F., Liu, H., & Gao, Y. (2011). QSPR models for the prediction of odor threshold of aliphatic alcohols. *Journal of Computational Science & Engineering, 2*, 69-78.