UNIVERSIDAD NACIONAL DE CÓRDOBA

MAESTRÍA EN ESTADÍSTICA APLICADA

FACULTAD DE CIENCIAS ECONÓMICAS

FACULTAD DE CIENCIAS AGROPECUARIAS

FACULTAD DE MATEMÁTICA, ASTRONOMÍA, FÍSICA Y COMPUTACIÓN

APLICACIÓN DEL APRENDIZAJE AUTOMÁTICO PARA LA PREDICCIÓN DEL GUSTO DE MOLÉCULAS ORGÁNICAS

Tesista:
Cristian Xavier Rojas Villa

Directores:
Fernando García
Davide Ballabio

2021

El presente trabajo de Tesis se desarrolló en el Laboratorio de Quimiometría y QSAR de la Facultad de Ciencia y Tecnología de la Universidad del Azuay (Cuenca–Ecuador).

Se presenta en consideración de las autoridades de la Facultad de Ciencias Económicas, Facultad de Ciencias Agropecuarias y Facultad de Matemática, Astronomía, Física y Computación de la Universidad Nacional de Córdoba para acceder al grado académico de Magíster en Estadística Aplicada.

AGRADECIMIENTOS

Deseo expresar mi agradecimiento a todas las personas que hicieron posible el desarrollo de la presente tesis de maestría:

A mis directores Fernando García y Davide Ballabio, por aceptar la dirección del trabajo de tesis y por su invaluable enseñanza y aporte tanto en lo científico como en lo personal.

A mis amigos del programa de Maestría: Rommel, Viviana, Ricardo y Jorge; con quienes compartí inolvidables momentos de estudio y reuniones.

A mis amigos del Grupo de Investigación en Quimiometría y QSAR de la Universidad del Azuay, mi lugar de trabajo, de forma particular a Diego Suárez y Piercosimo Tripaldi por la invaluable revisión del manuscrito y sus valiosas sugerencias para mejorar la calidad técnica del mismo. De igual forma a Karen Pacheco, Elisa Pacheco y Mateo Mendoza, quienes durante su estancia en el grupo de investigación colaboraron activamente en la búsqueda de la información, la digitalización de la base de datos y el diseño de las estructuras moleculares.

Finalmente, mi mayor gratitud va para mi esposa Fer, quien es un apoyo constante en el desarrollo de la ciencia; al igual que toda mi familia.

Esta tesis está dedicada con profundo cariño

a toda mi familia

A la memoria de mi padre José Tomás

Un buen científico debe tener la imaginación de un niño, la determinación de un joven, la racionalidad de un adulto y la experiencia de un anciano. La dificultad radica en tener todas estas cualidades al mismo tiempo

Roberto Todeschini en Molecular Descriptors for Chemoinformatics: Wiley–VCH (2009)

ÍNDICE

SIGLAS

3D–MoRSE	3D–Molecule Representation of Structures based on Electron diffraction Representación molecular 3D de estructuras basadas en la difracción de electrones
AAIs	Amino Acid indices Índices de aminoácidos
Acc	Accuracy Exactitud
AD	Applicability Domain Dominio de aplicabilidad
ADMET	Absorption, Distribution, Metabolism, Excretion, and Toxicity Absorción, distribución, metabolismo, excreción y toxicidad
BNN	Binned Nearest Neighbors Vecinos más cercanos agrupados
CART	Classification And Regression Trees Árboles de clasificación y regresión
CATS	Chemically Advanced Template Search Búsqueda de plantillas químicamente avanzadas
ChEBI	Chemical Entities of Biological Interest Entidades químicas de interés biológico
CML	Chemical Markup Language Lenguaje de marcado químico
CRD	Cysteine–Rich Domain Dominio rico en cisteína
DA	Discriminant Analysis Análisis discriminante
DNN	Deep Neuron network Red neuronal profunda
ECFPs	Extended Connectivity FingerPrints Huellas dactilares moleculares de conectividad ampliada
EMCC	Extended Matthew correlation coefficient Coeficiente de correlación de Matthew extendido
EPA	Environmental Protection Agency Agencia de Protección Ambiental
FN	False Negative Falso negativo
FP	False Positive Falso positivo

FPs	Fingerprints
	Huellas dactilares moleculares
GAs	Genetic Algorithms
	Algoritmos genéticos
GA–SAR	Genetic Algorithm utilizing Self–Assessment–Report
	Algoritmo genético utilizando un informe de autoevaluación
GBM	Gradient Boosting Cachine
	Máquina de impulso de gradiente
GETAWAY	GEometry, Topology, and Atom–Weights AssemblY
	Descriptores de ensamblado de pesos de átomos, geometría y topología
GMP	Guanosine MonoPhosphate
	5'–guanilato disódico
GPCR	G Protein–Coupled Receptor
	Receptores acoplados a proteínas G
IMP	Inosine MonoPhosphate
	5'–inosinato disódico
*k*NN	*k*–Nearest Neighbors
	k–vecinos más cercanos
LDA	Linear Discriminant Analysis
	Análisis discriminante lineal
LLA	Linear Learning Machine
	Máquina de aprendizaje lineal
LMO	Leave–Many–Out
	Dejar–varios–fuera
LOO	Leave–One–Out
	Dejar–uno–fuera
LVs	Latent Variables
	Variables latentes
MACCS	Molecular ACCess System
	Sistema de acceso molecular
MCC	Matthew correlation coefficient
	Coeficiente de correlación de Matthew
MDS	MultiDimensional Scaling
	Escalado multidimensional
ML	Machine Learning
	Aprendizaje automático
MM	Molecular Mechanics
	Mecánica molecular
MQN	Molecular Quantum Numbers
	Números cuánticos moleculares
MSG	MonoSodium Glutamate
	glutamato monosódico
N3	*N*–Nearest Neighbors
	N–vecinos más cercanos
NER	Non–Error–Rate & Balanced Accuracy

	Tasa de aciertos & Exactitud balanceada
OECD	Organisation for Economic Co–operation and Development
	Organización para la Cooperación Económica y el Desarrollo
OOB	Out–Of–Bag
	Fuera de la bolsa
PCA	Principal Component Analysis
	Análisis de componentes principales
PCs	Principal Components
	Componentes principales
PFPs	Path Fingerprints
	Huellas dactilares moleculares de trayecto
PLSDA	Partial Least Squares Discriminant Analysis
	Análisis discriminante de mínimos cuadrados parciales
PPPs	Potential Pharmacophore Points
	Potenciales sitios farmacóforos
Pr	Precision
	Precisión
QDA	Quadratic Discriminant Analysis
	Análisis discriminante cuadrático
QSAR	Quantitative Structure–Activity Relationship
	Relaciones cuantitativas estructura–actividad
QSPR	Quantitative structure property relationships
	Relaciones cuantitativas estructura–propiedad
RDF	Radial Distribution Function
	Función de distribución radial
REACH	Registration, Evaluation, Authorisation and Restriction of Chemicals
	Registro, evaluación, autorización y restricción de sustancias químicas
RF	Random forests
	Bosques aleatorios
RLR	Ridge Logistic Regression
	Regresión logística ridge
RS	Relative Sweetness
	Dulzor relativo
SIMCA	Soft Independent Modelling by Class Analogy
	Modelado suave independiente por analogía de clases
SMIfp	SMILES fingerprint
	Huellas dactilares de conteo del sistema de especificación de introducción lineal molecular simplificada
Sn	Sensitivity
	Sensibilidad
SNE	Stochastic Neighbor Embedding
	Incrustación de vecinos estocásticos
Sp	Specificity
	Especificidad
SVM	Support Vector Machine

	Máquinas de soporte vectorial
T1R2	Type 1 Receptor 2
	Receptor 2 tipo 1
T1R3	Type 1 Receptor 3
	Receptor 3 tipo 1
TMD	Heptahelical TransMembrane Domain
	Dominio transmembrana heptahelical
TN	True negative
	Verdadero negativo
TP	True Positive
	Verdadero positivo
TRCs	Taste Receptor Cells
	Células receptoras de gusto
t–SNE	t–distributed Stochastic Neighbor Embedding
	Incrustación de vecinos estocásticos distribuidos en t
VFTD	Venus FlyTrap Domain
	Dominio extracelular amino–terminal tipo Venus atrapamoscas
V–WSP	Variable reduction based on Wootton, Sergent, Phan–Tan–Luu's algorithm
	Método de reducción de variables basado en el algoritmo de Wootton, Sergent y Phan–Tan–Luu
WHALES	Weighted Holistic Atom Localization and Entity Shape
	Descriptores de localización atómica holística ponderada y de forma de entidad
WHIM	Weighted Holistic Invariant Molecular descriptors
	Descriptores moleculares invariantes holísticos ponderados

INTRODUCCIÓN

El estudio de la percepción del gusto se ha convertido en un tema de interés en diversas disciplinas científicas, entre ellas la química (quimioinformática), bioquímica (bioinformática), farmacología, sensometría, entre las más importantes. Adicionalmente, la interacción que existe entre estos campos ha provocado que el conocimiento sobre este tema haya tenido avances notables en los últimos años. Se considera a la química del gusto como una línea de investigación relativamente nueva en la química de los alimentos. En la actualidad existen cinco gustos básicos, también denominados gustos mediados por receptores, que son el dulce, amargo, umami, salado y ácido. Las sustancias responsables del gusto constituyen un conjunto extremadamente amplio y diverso de compuestos químicos que tienen la capacidad de estimular los receptores del sabor, o los nervios específicos de tal forma que se produzca la sensación de un gusto particular. Sin embargo, durante la síntesis y elucidación de un nuevo blanco molecular, sutiles modificaciones en la estructura química permiten el cambio de un gusto a otro, la pérdida de la percepción del mismo (insípido) o la presencia de diversos sabores (multigusto). Es así que la disponibilidad de compuestos químicos (ligandos) de gusto se haya incrementado drásticamente y se encuentren dispersos en diversas fuentes bibliográficas. De forma complementaria, en la actualidad existe mayor conocimiento sobre la naturaleza de los receptores que perciben cada uno de los gustos y la forma en que interaccionan con los ligandos (modelo estructural de llave–cerradura) para generar una respuesta en el cerebro que se traduce en la sensación gustativa.

A inicios de los años sesenta Corwin Hansch y Toshio Fujita publican las primeras investigaciones sobre las relaciones cuantitativas estructura–

actividad/propiedad (QSAR/QSPR)[1]. Estos trabajos brindaron un fuerte estímulo para utilizar esta metodología para la predicción de nuevas actividades o propiedades de diversos tipos de moléculas. Particularmente, en el año 1980, Lemont B. Kier publicó el que posiblemente se considera el primer estudio QSAR relacionado a la discriminación de compuestos dulces y amargos, utilizando diversas aldoximas. A partir de este estudio, el interés en la discriminación de los compuestos dulces y amargos se ha incrementado a lo largo de la historia. Esto se refleja en las trece investigaciones publicadas solo en la última década. En estos trabajos se presentan bases de datos cada vez más extensas, con las cuales se aplican diversas máquinas del aprendizaje automático para generar modelos de clasificación predictivos.

Con estos antecedentes, la motivación de la presente tesis de maestría es utilizar la relación entre la estructura y la actividad de los compuestos químicos para el desarrollo de modelos computacionales gustativos más útiles y eficaces. Para este propósito, se compilará una base de datos extensa de la información que se encuentra reportada en diversas fuentes bibliográficas. Seguidamente, se verificará y filtrará la información de tal forma de obtener una base de datos validada para aplicar el aprendizaje no supervisado con el propósito de definir el espacio químico del gusto. Posteriormente, se utilizarán diversas estrategias del aprendizaje supervisado (clasificación) para proponer modelos que permitan realizar predicciones confiables del gusto de nuevas moléculas.

El documento inicia con una breve introducción para contextualizar la problemática inherente a la predicción del gusto de las diversas estructuras moleculares. En el Capítulo 1 se presenta una revisión de la teoría relacionada a la química del gusto, en la que se incluyen detalles de las diversas bases de datos y definiciones del espacio químico reportados en la literatura. A continuación, el Capítulo 2 brinda la teoría necesaria para entender la naturaleza de las relaciones cuantitativas estructura–actividad, la forma en que se representan los compuestos químicos para obtener los conjuntos de variables que las describen. En el Capítulo 3 se describe las principales técnicas del aprendizaje automático, haciendo énfasis particular en los métodos que se han utilizado en las aplicaciones. En el Capítulo 4 se presenta la base de datos, con la cual se ha definido y analizado el espacio químico del gusto; para posteriormente calibrar los modelos de clasificación

[1] En esta tesis se utilizan siglas en idioma inglés, las que se definen la primera vez que se introducen y se encuentran detalladas en inglés y español en la sección de SIGLAS al inicio del documento.

que permitan predecir el gusto de las moléculas estudiadas. Finalmente, en la sección de discusión final y perspectivas futuras se presenta un análisis crítico de los logros alcanzados durante el trabajo de investigación, los aportes que se derivan del mismo y las futuras líneas de investigación.

Objetivo General

Aplicar técnicas de aprendizaje automático para la predicción del gusto de diversos tipos de moléculas.

Objetivos Específicos

1. Compilar una base de datos de gustos de diversos tipos de moléculas a partir de múltiples fuentes bibliográficas.
2. Curar las estructuras moleculares y filtrar la base de datos para identificar compuestos erróneos, duplicados o moléculas ambiguas.
3. Aplicar el aprendizaje no supervisado de la incrustación de vecinos estocásticos distribuidos en t para la definición del espacio químico del gusto.
4. Desarrollar y comparar modelos predictivos para los gustos básicos basados en diferentes algoritmos del aprendizaje automático: *k*–vecinos más cercanos (*k*NN), *N*–vecinos más cercanos (N3), vecinos más cercanos agrupados (BNN), bosques aleatorios (RF) y el análisis discriminante de mínimos cuadrados parciales (PLSDA).

Capítulo 1

QUÍMICA DEL GUSTO

1.1. Aspectos generales sobre el gusto

La química del gusto es un área de interés en la ciencia de los alimentos, pues son varios los investigadores alrededor del mundo que se han interesado en desarrollar estudios para entender los mecanismos por medio de los cuales se perciben los distintos gustos o sabores [1]. Se define al gusto como la combinación de sensaciones químicas percibidas por los receptores moleculares (membranas biológicas) de la lengua, que interaccionan con compuestos solubles que poseen diferentes propiedades osmóticas, endotérmicas y exotérmicas [2]. Los gustos se categorizan en cinco grupos básicos: dulce, amargo, umami, salado y ácido; sin embargo, existen diversas sustancias que no presentan un único sabor, sino más bien una mezcla compleja de sensaciones de estos gustos básicos e incluso otros como picante, astringente o quemante [1–3].

La percepción del gusto varía de persona a persona y depende de diversos factores internos, tales como la concentración del compuesto de sabor, psicología y anatomía del humano, interacción con otras moléculas (potenciadores o inhibidores), entre los más importantes [3]. Durante la ingesta y masticación de los alimentos, los compuestos químicos de sabor ingresan en los poros de las papilas gustativas de la lengua, donde interaccionan con las células receptoras de gusto (TRCs), las que contienen quimiorreceptores que detectan de forma específica alguno de los cinco gustos básicos. En consecuencia, cuando se produce la interacción de un ligando (compuesto químico) con un receptor específico (modelo estructural de llave–cerradura), se recibe una señal química en la médula del cerebro [2]. Debido a que el gusto se detecta mediante receptores específicos, se ha propuesto el uso del término «gusto mediado por receptores» [4,5].

1.2. Definición de los diversos gustos

1.2.1. Dulce

El gusto dulce es posiblemente el más significativo para las personas, debido a que produce una sensación gustativa agradable en la mayoría de alimentos y fármacos [1,6]. De entre los múltiples edulcorantes caracterizados hasta el día de hoy, la sacarosa (CAS 57–50–1 y PubChem CID 5988) es la más importante. De hecho, este azúcar se emplea como estándar para cuantificar el dulzor de otros edulcorantes, debido a que evoca un gusto dulce limpio y no produce retrogustos, incluso en altas concentraciones [1]. De esta manera, se define al dulzor relativo como la relación entre la concentración de una solución del estándar Sacarosa y la concentración de otro endulzante; es decir, a la Sacarosa se le asigna un valor de dulzor de 1 (o 100) y el dulzor del edulcorante será comparado con respecto a ella [7,8].

El receptor del gusto dulce tiene la capacidad de reconocer moléculas que pertenecen a clases de compuestos muy variados, por ejemplo, azúcares, aminoácidos, péptidos, proteínas y otras clases de moléculas orgánicas [4]. Este receptor está formado por un heterodímero de dos subunidades relacionadas con la secuencia que pertenece a la familia de receptores acoplados a proteínas G (GPCR) de clase C. El receptor 2 tipo 1 (T1R2) y receptor 3 tipo 1 (T1R3) están formados por tres dominios estructurales (Figura 1.1): 1) dominio extracelular amino–terminal tipo «Venus atrapamoscas» (VFTD), 2) dominio transmembrana heptahelical (TMD) y 3) dominio rico en cisteína (CRD) que conecta los dos primeros dominios. Los lóbulos T1R2 y T1R3 cambian su disposición de forma flexible para crear una conformación «abierta» o «cerrada» que permite el reconocimiento de diversos tipos de edulcorantes [2].

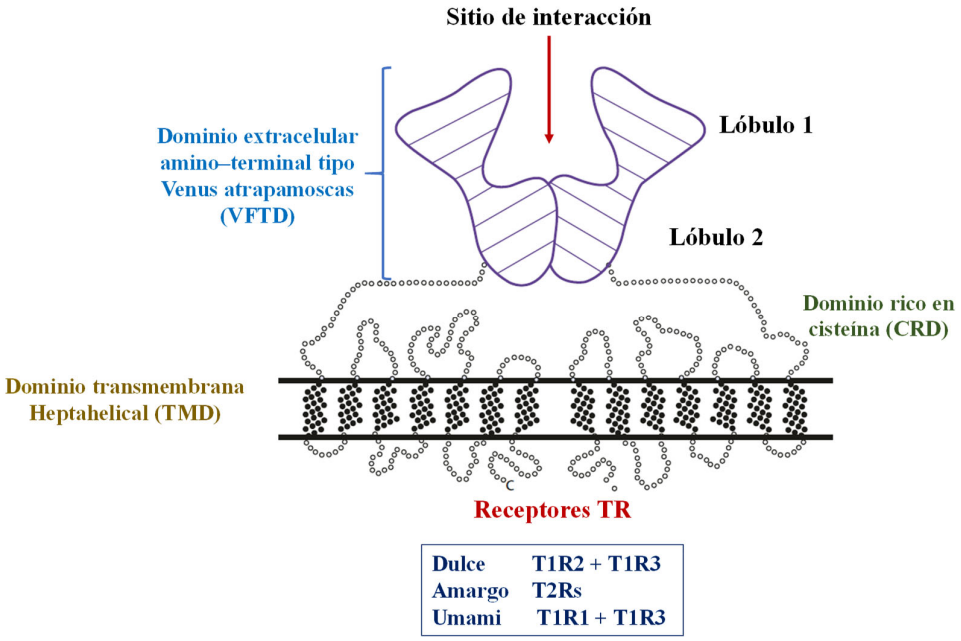

Figura 1.1 Representación esquemática de los quimiorreceptores del gusto dulce, amargo y umami.

1.2.2. Amargo

El amargor es un gusto identificado principalmente en los alcaloides, por ejemplo la quinina (CAS 130–95–0 y PubChem CID 3034034), que es un aditivo ampliamente usado en la industria de los alimentos [8]. Por este motivo, este gusto ha estado vinculado a que el hombre se proteja contra el consumo de alcaloides perjudiciales (venenos). No obstante, en ciertos alimentos, tales como cerveza, té, café, chocolate, aceitunas y otros, el amargor se percibe como un gusto placentero y ayuda a definir la identidad de los mismos [1,8]. Asimismo, existen fitonutrientes que se encuentran en frutas y verduras, así como distintos fármacos elaborados a partir de diversas hierbas que son beneficiosas para el cuerpo humano [9]. Los compuestos amargos son reconocidos por la familia de receptores tipo 2 (T2Rs), los cuales comprenden alrededor de 35 miembros altamente diversos (Figura 1.1). Estas proteínas contienen entre 300 y 330 aminoácidos con un dominio N–terminal extracelular corto (sin una estructura tipo «Venus atrapamoscas»). La mayoría de los receptores T2R se expresan en las mismas TRCs, que funcionan como sensores específicamente ajustados para sustancias químicas amargas [2].

El gusto amargo está estrechamente relacionado a la forma en cómo se percibe el dulzor, por esta razón algunas moléculas desencadenan tanto sensaciones dulces como amargas [1].

1.2.3. Umami

Umami es un término japonés con el que se conoce al gusto más recientemente reconocido y que se define también como «delicioso» o «sabroso» [10]. Se encuentra particularmente relacionado con L–aminoácidos, como por ejemplo el glutamato monosódico (MSG) (CAS 142–47–2 y PubChem CID 23672308), así como otros compuestos denominados potenciadores de sabor [2,11]. De hecho, el glutamato monosódico presenta un efecto sinérgico con los nucleótidos 5'–monofosfato o inosinato 5'–monofosfato, aunque estos compuestos solos presenten un sabor umami débil [2,12]. Los heterodiméricos T1R1 y T1R3 (Figura 1.1) se ensamblan para formar el receptor del gusto umami. Así, un nucleótido interacciona con el correspondiente receptor en tres puntos, dos de los cuales son electrofílicos (A y B) que se acoplan con los dos oxígenos de fosforilo y el oxígeno del carbono 6, respectivamente; mientras que el sitio X interacciona con el sustituyente del carbono 2, particularmente cuando éste se encuentra deslocalizado [2]. La mayoría de estos compuestos se dividen en dos grupos: 1) L–α–aminoácidos, representados por el glutamato monosódico y 2) 5'–ribonucleótidos y sus derivados, representados por el 5'–inosinato disódico (IMP) (CAS 4691–65–0 y PubChem CID 135414245) o el 5'–guanilato disódico (GMP) (CAS 5550–12–9 y PubChem CID 135414246) [10].

1.2.4. Ácido

El gusto ácido, también conocido como agrio, se percibe en las papilas gustativas de la lengua mediante canales iónicos sensibles a los protones H^+ que provienen de los ácidos, sales ácidas y otros compuestos que, al ser diluidos en agua, generan el ion hidronio (H_3O^+). Este gusto también puede ser inducido mediante el paso de corriente eléctrica en la lengua, lo cual probablemente genera iones por hidrólisis del agua o compuestos ácidos [2]. Adicionalmente, los ácidos no disociados también juegan un rol importante en la percepción de este gusto; por ejemplo, algunos ácidos débiles (cítrico, succínico, málico o láctico) que se encuentran naturalmente en los alimentos tienen un sabor ácido más intenso que el ácido clorhídrico (CAS 7647–01–0

y PubChem CID 313) al mismo pH [12]. Por otro lado, otros compuestos ácidos, tales como el oxalato ácido de potasio (CAS 127–95–7 y PubChem CID 23662386) o ácido protocatecuico (CAS 99–50–3 y PubChem CID 72) producen contemporáneamente los gustos ácido y amargo [2]. De forma análoga al amargor, el gusto ácido activa una alarma en el cerebro, debido a que algunas sustancias nocivas poseen este gusto.

1.2.5. Salado

El gusto salado es un estímulo producido por las sales solubles, particularmente aquellas de bajo peso molecular, por ejemplo, los cloruros de sodio, potasio y calcio [2]. Se ha establecido que son los cationes los responsables de este gusto, mientras que los aniones lo modifican. Así, los cationes Na^+ y Li^+ generan únicamente sabor salado, mientras que el K^+ (y otros cationes alcalinotérreos) produce una combinación salado–amargo. Por el contrario, los aniones inhiben el sabor salado generado por los cationes, siendo el anión Cl^- el menos inhibitorio de todos [1]. Por esta razón, el cloruro de sodio (NaCl) (CAS 7647–14–5 y PubChem CID 5234) es el único compuesto que brinda un gusto salado limpio e intenso. Por su parte, el cloruro de potasio (CAS 7447–40–7 y PubChem CID 4873) se puede considerar como un sustituyente del NaCl, aunque éste produce, además, un gusto amargo fuerte y retrogusto desagradable. Por el contrario, las sales que tienen alto peso molecular evocan sabor amargo en lugar de salado, por ejemplo, el cloruro de litio (CAS 7447–41–8 y PubChem CID 433294) y cloruro de amonio (CAS 12125–02–9 y PubChem CID 25517); sin embargo, su uso está limitado para consumo humano debido a cuestiones de seguridad (toxicidad) y retrogusto, respectivamente [2,12].

1.2.6. Multigusto

Además de los cinco gustos básicos, existen moléculas que generan una mezcla de sensaciones de los mismos, así como otros. Por ejemplo, el oxalato ácido de potasio (CAS 127–95–7 y PubChem CID 23662386) y el ácido protocatéquico (CAS 99–50–3 y PubChem CID 72) presentan el sabor ácido y amargo [2]; mientras que el acesulfamo de potasio (CAS 55589–62–3 y PubChem CID 11074431), la sacarina sódica (CAS 128–44–9 y PubChem CID 656582) y la hernandulcina (CAS 95602–94–1 y PubChem CID 125608)

poseen gusto dulce y amargo [8]. En consecuencia, estos compuestos se los puede denominar multigusto o multisabor.

1.2.7. Insípido

El término insípido se usa para definir la ausencia o pérdida de la sensación de los cinco gustos básicos: dulce, amargo, ácido, salado o umami [8].

1.3. Bases de datos de gustos

Los primeros trabajos relacionados a la discriminación entre los gustos dulce y amargo datan del año 1980 [13,14]. A partir de entonces ha existido un creciente interés en la generación de bases de datos y la utilización de diversas máquinas de aprendizaje para predecir el gusto de diversas familias de compuestos orgánicos.

1.3.1. SuperSweet

En el año 2011, se publica la base de datos SuperSweet [15], la cual brinda información de diversos carbohidratos, aminoácidos, edulcorantes artificiales, proteínas y otros compuestos; algunos de los cuales han sido identificados en la quimioteca PubChem [16]. Para cada molécula se reporta la estructura molecular, propiedades fisicoquímicas, dulzor relativo, valor calórico, índice glicémico, efecto terapéutico, metabolismo y la clase edulcorante. Esta base de datos implementa una interfaz gráfica para la búsqueda de similitud molecular, para lo cual se incluyen cuatro huellas dactilares (FP) diferentes (FP2, FP3, FP4 y MACCS) calculadas en el programa Open Babel. Para optimizar la búsqueda, se combinan las huellas dactilares FP2 y FP4. La huella dactilar FP4 se basa en un conjunto de patrones SMARTS y considera diversos grupos funcionales. Durante la búsqueda molecular se utiliza el coeficiente de similitud de Tanimoto para datos binarios. Se provee también un modelador 3D del receptor del dulzor y los sitios de unión para moléculas edulcorante pequeñas «docking», los cuales son útiles para el diseño de nuevos edulcorantes.

1.3.2. BitterDB

Un año más tarde, se publicó la base de datos BitterDB [17], la cual incluye aproximadamente 1000 compuestos. La mayoría de moléculas poseen

gusto amargo; sin embargo, otras declaran un gusto diferente al amargo (multigusto). Otras estructuras químicas que se incluyen han demostrado ser capaces de activar al menos un receptor humano de amargor. En esta base de datos se encuentra información relacionada a las principales propiedades moleculares, umbral de amargor, la notación lineal de cadena SMILES, número de registro CAS, entre las más importantes. De forma complementaria, BitterDB brinda un enlace a las quimiotecas de acceso libre PubChem [16] y ZINC [18], e incluye enlaces a las publicaciones relacionadas a las interacciones ligando–receptor, concentración efectiva para la activación del receptor, así como la concentración efectiva media (EC_{50}). Adicionalmente, se puede encontrar información sobre las mutaciones de los receptores y la forma en que influyen en su activación por compuestos amargos.

1.3.3. TasteDB

En un estudio posterior, se fusionaron SuperSweet y BitterDB para obtener la base de datos TasteDB [19] con 806 moléculas. Contemporáneamente, se consideraron 1760 compuestos orgánicos volátiles de las bases de datos SuperScent y Flavornet, para generar la base de datos FragranceDB. Seguidamente, se fusionaron TasteDB y FragranceDB para definir el espacio químico de 2517 compuestos mediante el análisis de componentes principales (PCA) como técnica de visualización mediante las dos primeras componentes. Para este propósito se calcularon 42 números cuánticos moleculares y un conjunto de huellas dactilares moleculares a partir de la notación lineal de cadena SMILES.

1.3.4. TastesDB

En el año 2017, se publicó la base de datos TastesDB [5] con 727 estructuras moleculares asociadas a los gustos dulce (435 moléculas), amargo (81 compuestos) e insípido (133 estructuras). En este trabajo se modeló la clase dulce, por lo que las clases amargo e insípido se fusionaron en la clase etiquetada como no dulce. Para cada estructura molecular se calcularon diversos descriptores moleculares y huellas dactilares moleculares de conectividad ampliada (ECFPs) independientes de la conformación. Con estas variables, se construyó una relación cuantitativa estructura–actividad (QSAR) basada en un sistema experto, que integra el aprendizaje no

supervisado (escalado multidimensional) y el aprendizaje supervisado (análisis discriminante de mínimos cuadrados parciales y el método de los *N*–vecinos más cercanos).

1.3.5. SweetenersDB

El mismo año 2017 se publicó la base de datos SweetenersDB [20] que contiene 316 edulcorantes con el respectivo dulzor relativo. Se calcularon diversos descriptores moleculares en el programa DRAGON, con los cuales se desarrollaron modelos QSAR mediante las técnicas de aprendizaje supervisado de bosques aleatorios (RF) y regresión de soporte vectorial (SVR). A continuación, se realizó un cribado virtual con un conjunto de moléculas naturales de la base de datos Supernatural II para diferenciar los compuestos dulces de aquellos que presentan amargor y toxicidad. Recientemente, en el año 2020, se publicó la segunda versión de SweetenersDB, con la que se calibraron diversos modelos de regresión basados en RF, SVM, AdaBoost Tree y los *k*–vecinos más cercanos (*k*NN). Para este efecto, se usó un conjunto de calibración de 252 compuestos y un conjunto de predicción de 64 moléculas, con los cuales se calcularon dos grupos de descriptores moleculares (Dragon y un modelo de acceso libre). Posteriormente, se realizó un cribado virtual de aproximadamente 4800 moléculas naturales para identificar tres edulcorantes potenciales que fueron analizados mediante acoplamiento molecular con el receptor T1R2/T1R3. También se construyó un espacio químico basado en la incrustación de vecinos estocásticos distribuidos en t (t–SNE).

1.4. Espacio químico del gusto

El espacio químico sirve para conceptualizar el número total de moléculas, reales o virtuales, como una analogía con el universo cosmológico en su inmensidad, es decir, las estrellas representan a los compuestos químicos que pueblan el espacio [21]. En consecuencia, se define el espacio químico como el conjunto de todas las estructuras moleculares posibles, descritas por un vector *N*–dimensional de descriptores moleculares que capturan la información química significativa de las mismas. La dimensión N está en el orden de magnitud de 10^2–10^4 [21,22]. Debido a que el espacio químico está definido en un espacio complejo multidimensional, la forma más intuitiva para analizarlo es mediante la reducción de su dimensionalidad mediante la

proyección de las similitudes/disimilitudes en un mapa bidimensional (2D) o tridimensional (3D) [22]. No obstante, es bien conocido que el espacio químico es extremadamente grande y en la actualidad simplemente una pequeña fracción de este universo de moléculas es conocido [21].

1.4.1. Espacio químico del gusto basado en el análisis de componentes principales

Una forma de analizar el espacio químico es mediante el análisis de componentes principales (PCA). Este enfoque se ha usado con las bases de datos TasteDB y FragranceDB, utilizando 42 números cuánticos moleculares (MQN) y 34 huellas dactilares de conteo del sistema de especificación de introducción lineal molecular simplificada (SMIfp) [19]. Se usó el gráfico de puntuaciones de las dos primeras componentes principales como una forma de visualizar el espacio químico multidimensional en un mapa general bidimensional. El análisis del espacio químico basado en los números cuánticos moleculares, evidencia el aumento de tamaño de los compuestos a lo largo de la primera componente PC1 (67.97 % de varianza), mientras que la segunda componente PC2 (15.54 % de varianza) permite separar a las moléculas por su rigidez estructural. Por otra parte, en el espacio basado en las huellas dactilares, la PC1 (66.9% de varianza) separa los compuestos de acuerdo al número de carbonos no aromáticos, mientras que la PC2 (18.97% de varianza) lo hace en función del número de átomos de carbono aromáticos. En términos generales, el espacio químico definido por los MQN y las SMIfp no reflejan ninguna distribución de propiedades de polaridad, debido a que esta propiedad se encuentra, por lo general, en la tercera componente PC3.

En otro estudio publicado en el año 2017, se definió el espacio químico para el gusto amargo mediante el PCA [23]. En la base de datos se recopilaron 2608 compuestos, de los cuales 691 son amargos (632 de BitterDB) y 1917 no amargos (flavores, dulces, insípidos). A partir de esta base de datos se extrajo el 70% de las moléculas de forma aleatoria y proporcional a la numerosidad de las clases (grupo de calibración). También se consideraron 41132 moléculas aleatorias tomadas de la quimioteca de entidades químicas de interés biológico (ChEBI). Las estructuras 2D se representaron mediante 12 descriptores fisicoquímicos: peso molecular, lipofilicidad, número de enlaces rotables, área de superficie polar, estados electrotopológicos, refractividad molecular, polarizabilidad molecular,

aceptor de enlaces de hidrógeno, donante de enlaces de hidrógeno, número de anillos, número de centros quirales y número de átomos pesados. El espacio químico para el grupo de calibración y las moléculas aleatorias se definió en término de la proyección de las dos primeras componentes principales. Las moléculas amargas se extienden ampliamente dentro del mapa químico; mientras que cada subconjunto no amargo cubre un subespacio distinto (el conjunto combinado cubre casi todo el dominio químico, aunque no distribuido uniformemente).

1.4.2. Espacio químico del gusto basado en el escalado multidimensional

En otro estudio se ha definido el espacio químico para compuestos dulces, amargos e insípidos de la base de datos TastesDB mediante el uso del escalado multidimensional (MDS) [5]. Se utilizaron 649 moléculas dulces y no dulces, las cuales fueron representadas mediante las huellas dactilares moleculares de conectividad ampliada (ECFPs) y las similitudes/disimilitudes se cuantificaron mediante la distancia para datos binarios de Jaccard–Tanimoto. El espacio químico definido por las dos primeras coordenadas permitió visualizar dos grupos consistentes de moléculas dulces y otro donde se encuentran superpuestas las moléculas dulces y no dulces. Posteriormente, se aplicaron técnicas del aprendizaje supervisado para discriminar las moléculas superpuestas en el tercer grupo.

1.4.3. Espacio químico del gusto basado en la incrustación de vecinos estocásticos distribuidos en t

En el año 2019, se publica BitterSweet [24] como una herramienta para clasificar compuestos amargos y dulces, en el que se consideró una base de datos curada de 918 amargos, 1510 no amargos, 1205 compuestos dulces y 1171 no dulces. Estos compuestos fueron representados por propiedades fisicoquímicas calculadas en el programa Canvas. Para el desarrollo del espacio químico se consideró un grupo de calibración de 2257 moléculas amargas–no amargas y 2205 compuestos dulces–no dulces; así como de otros compuestos seleccionados aleatoriamente de la quimioteca de entidades químicas de interés biológico (ChEBI). El espacio químico se obtuvo mediante la incrustación de vecinos estocásticos distribuidos en t (t–SNE),

el cual demuestra la diversidad molecular de los compuestos amargos, dulces, insípidos y no amargos, en comparación con las moléculas bioactivas aleatorias. Asimismo, el mapa bidimensional t–SNE muestra la distribución de las moléculas tomadas de diferentes fuentes bibliográficas, capturando de forma gradual subconjuntos del espacio químico general.

En un estudio más reciente, se han usado los 316 compuestos dulces de la base de datos SweetenersDB [20] para mapear el espacio químico mediante la t–SNE. Adicional a esta base de datos, se han utilizado 4796 compuestos naturales de la base de datos Super–Natural II y PhytoLab, así como tres compuestos testeados experimentalmente: arctiina (CAS 20362–31–6 y PubChem CID 100528), ginsenósido Rd (CAS 52705–93–8) y jujubosido A (CAS 55466–04–1 y PubChem CID 51346169). Cada estructura ha sido representada por descriptores moleculares independientes de la conformación calculados en los programas Dragon, RDKit, Mordred y ChemoPy. Estas variables han sido usadas para definir el mapa bidimensional t–SNE en el paquete Python, para lo cual se han utilizado los parámetros por defecto, es decir, perplejidad de 30, exageración de 12 y tasa de aprendizaje de 200 con 1000 iteraciones. El análisis del espacio químico indica que no hay una completa superposición de los compuestos naturales con las moléculas de SweetenersDB, lo cual sugiere que una buena parte del espacio químico de los productos naturales permanece sin explorar.

Capítulo 2

RELACIONES CUANTITATIVAS ESTRUCTURA–ACTIVIDAD

2.1. Principios de modelado QSAR

Las relaciones cuantitativas estructura–actividad/propiedad (QSAR/QSPR) son técnicas asistidas por computadora *in silico*, que nacieron en la década de 1960 con las investigaciones desarrolladas por Corwin Hansch y Toshio Fujita [8,25,26]. En esta metodología se busca predecir las actividades/propiedades de diversos compuestos químicos mediante modelos matemáticos predictivos, los cuales sirven para realizar nuevas estimaciones o entender el mecanismo de acción involucrado. En consecuencia, los estudios QSAR/QSPR no están aislados, sino que más bien complementan a las investigaciones teóricas y/o experimentales enfocadas en diseñar de forma racional nuevos compuestos y responder a interrogantes de carácter químico de los fenómenos involucrados [27]. Este hecho se refleja en que en la actualidad adquiere mayor importancia el uso de modelos QSAR/QSPR para fines regulatorios en diversos organismos internacionales, por ejemplo, la Agencia de Protección Ambiental (EPA) de los Estados Unidos y la legislación de registro, evaluación, autorización y restricción de sustancias químicas (REACH) de la Unión Europea [8,28].

2.1.1. Definición y formalismo

Las relaciones cuantitativas estructura–actividad/propiedad (QSAR/QSPR) se refieren al desarrollo de correlaciones matemáticas entre una respuesta, definida por una actividad o propiedad, y ciertos atributos químicos codificados en la estructura química de los compuestos, denominados descriptores moleculares. Sin embargo, existen otras respuestas específicas de las moléculas que se pueden estudiar con esta metodología,

por ejemplo, toxicidad, citotoxicidad, biodegradación, reactividad, dulzor relativo, entre otras. En consecuencia, se usa el término general QSAR para englobar a todos estos tipos específicos de características moleculares [8]. De esta manera, el formalismo básico de la teoría QSAR que permite predecir la actividad de las moléculas se puede simbolizar matemáticamente de acuerdo a la siguiente expresión:

$$Y = f\left(x_1, x_2, \ldots, x_d\right) \tag{2.1}$$

donde Y es la actividad biológica o la propiedad fisicoquímica de las moléculas, la que se relaciona mediante una función matemática (aprendizaje automático) con los d descriptores moleculares ($x_1, x_2, \ldots, x_d$).

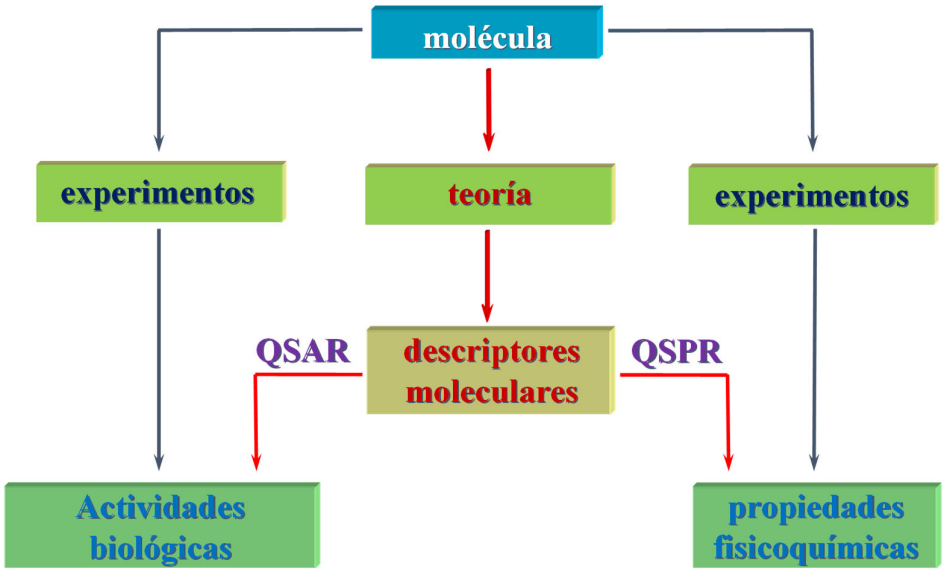

Figura 2.1 Diagrama de flujo para el desarrollo de relaciones cuantitativas estructura–actividad/propiedad

Las actividades biológicas y las propiedades fisicoquímicas son características intrínsecas de las moléculas que se obtienen mediante experimentos estandarizados. Por otra parte, la información contenida en los descriptores moleculares se obtiene de forma teórica mediante la aplicación de diversas teorías. La Figura 2.1 muestra la forma en que esta información experimental y teórica se complementan para desarrollar los modelos matemáticos QSAR/QSPR [28].

2.1.2. Objetivos del modelado QSAR

Cuando se realiza un estudio QSAR se busca principalmente desarrollar un modelo matemático predictivo de regresión, clasificación o híbrido, el cual estará acompañado de una interpretación del mecanismo de acción (información química involucrada) y del dominio de aplicabilidad (restricciones predictivas). Para el desarrollo del modelo se requiere de una colección de compuestos (quimioteca) que presenten el valor de respuesta de una actividad o propiedad de interés. El modelo validado permitirá realizar predicciones para un número mayor de estructuras moleculares. Otros objetivos que se alcanzan con los modelos QSAR son: reducir y reemplazar la experimentación de laboratorio usando animales, cribar virtual quimiotecas (públicas o privadas) para identificar moléculas con actividades/propiedades esperadas, optimizar la síntesis de compuestos con actividades/propiedades deseadas, identificar compuestos potencialmente peligrosos en las etapas iniciales del diseño de los mismos, predecir la toxicidad de los compuestos en seres humanos y especies ambientales y aplicar los modelos con fines regulatorios por parte de organismos gubernamentales [29,30], entre otras que dependerán de la naturaleza del problema en estudio.

2.1.3. Etapas en el desarrollo de un modelo QSAR

Para el desarrollo de un modelo QSAR se debe disponer de datos cuantitativos para analizarlos mediante las técnicas del aprendizaje automático más apropiadas. La información necesaria para los modelos QSAR se obtienen de dos fuentes fundamentales:

- medición experimental de la actividad/propiedad de interés mediante protocolos estandarizados.
- información química de las moléculas, codificada en los descriptores moleculares.

Debido a que la cantidad de información que se obtiene es grande (miles de descriptores moleculares), el uso de computadores con buena capacidad de cálculo es imperante para obtener los modelos en el menor tiempo posible. En términos generales, existen 4 etapas básicas que se deben considerar en los estudios QSAR, los cuales se esquematizan en la Figura 2.2 y se detallan a continuación [28,31].

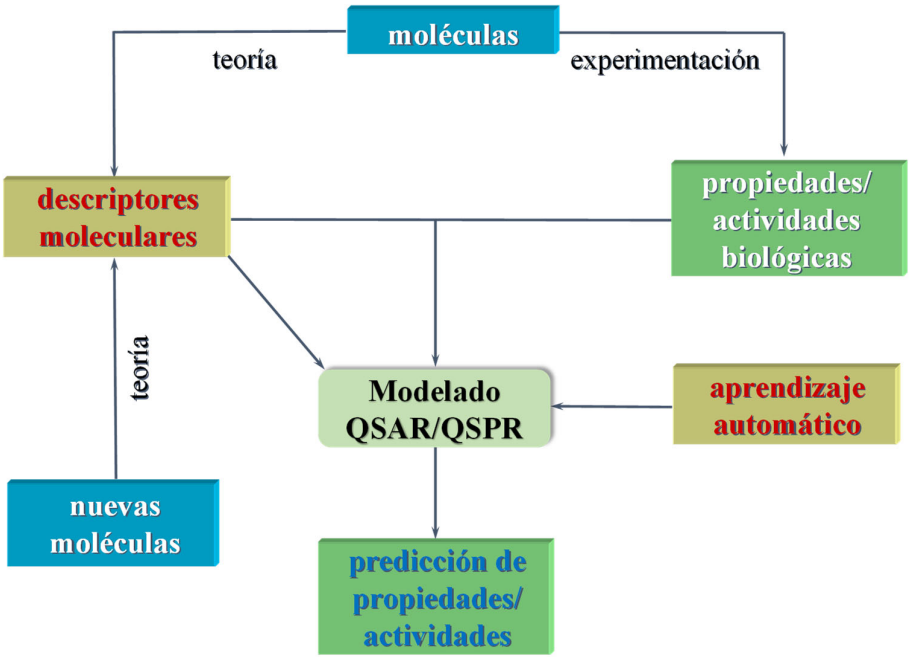

Figura 2.2 Etapas operativas de las relaciones cuantitativas estructura–actividad/propiedad

Preparación de los datos

Para iniciar el desarrollo de un modelo QSAR se debe contar con una colección de compuestos químicos con los valores numéricos de la actividad o propiedad (variables dependientes) a modelar. En algunos casos, la respuesta tiene un rango de variación alto, por lo tanto, es necesario realizar algún tipo de transformación (normalmente logarítmica) que permita que los valores experimentales sean lo más cercanos posibles. Por otro lado, también es importante considerar la correcta representación de la estructura química de los compuestos, pues los descriptores moleculares (variables independientes) se obtienen directamente de esta representación.

Filtrado y pretratamiento de los datos

Previo al desarrollo del modelo se debe filtrar o pretratar la información disponible. Aquí se incluyen dos etapas fundamentales: 1) eliminar compuestos duplicados (moléculas que tienen la misma estructura) y 2) excluir descriptores moleculares con valores faltantes, constantes o casi constantes; así como los que se encuentran correlacionados más arriba de un cierto umbral.

Validación y predicción de los datos

Para desarrollar un modelo QSAR predictivo es necesaria la división de la base de datos en dos conjuntos: grupo de calibración (o entrenamiento) y grupo de predicción (o validación). Este proceso se puede realizar de forma casual o usando diversos métodos del aprendizaje automático, por ejemplo, análisis de agrupamiento, método de subconjuntos balanceados o el algoritmo Kennard–Stone. El grupo de calibración se usa para el desarrollo del modelo, para lo cual se utilizan las técnicas del aprendizaje supervisado más apropiadas, además de métodos de validación interna o cruzada. Seguidamente, se utiliza el grupo de predicción (moléculas no consideradas durante la calibración) para medir la capacidad predictiva del modelo. Otro aspecto importante es indicar las limitaciones de predicción del modelo mediante la definición del dominio de aplicabilidad (AD).

Interpretación del modelo

La interpretación del modelo se refiere al mecanismo de acción, es decir, explicar los descriptores moleculares del modelo y la forma en que se relacionan con la predicción de la actividad o propiedad en estudio. El mecanismo de acción es opcional, pues debido a la complejidad de las definiciones abstractas de los descriptores, no siempre se puede proveer una explicación de este tipo.

2.1.4. Principios de validación de un modelo QSAR

Para asegurar la confiabilidad y aplicabilidad de un modelo QSAR, es necesario asegurarse que el mismo haya sido rigurosamente validado; de esta manera 1) se prueba la validez del modelo, 2) se verifica si una molécula a ser predicha está dentro del dominio de aplicabilidad del modelo y 3) se debe documentar la confiabilidad en el enfoque del modelado, de tal forma que brinde un algoritmo subyacente claro. Para verificar la validez y, fundamentalmente, la aplicabilidad de un modelo para fines regulatorios, se han propuesto cinco principios básicos por parte de la Organización para la Cooperación Económica y el Desarrollo (OECD) [8].

Definición de la actividad/propiedad

Cualquier actividad o propiedad de interés puede ser medida siguiendo diversos protocolos experimentales y bajo diferentes condiciones

experimentales; por ejemplo, para la medición del gusto se usan estándares que brindan una respuesta particular. Por ejemplo, para medir el dulzor, particularmente el dulzor relativo (potencia de dulzor), se usan soluciones estándar de Sacarosa (estándar universal); sin embargo, también se pueden usar estándares de Glucosa o Sacarina o algún otro edulcorante. En consecuencia, este primer principio busca garantizar claridad en la definición de la actividad o propiedad objeto de modelado.

Algoritmo inequívoco

En la literatura se encuentran diversos enfoques para el desarrollo de los modelos, por lo que este principio procura asegurar transparencia en el algoritmo utilizado para el desarrollo del modelo QSAR. Por esta razón, en la actualidad es común encontrar disponible la base de datos junto con el algoritmo matemático o diagrama de flujo usado para calibrar el modelo. La desventaja frente a este principio se encuentra en modelos desarrollados con fines comerciales, pues la información es privada y no siempre está disponible.

Definición del dominio de aplicabilidad

Los modelos QSAR no son universales sino más bien reduccionistas, es decir, las predicciones están acotadas por la naturaleza química y la actividad/propiedad de los compuestos que se utilizaron para calibrar el modelo. Es decir, esta etapa se relaciona al principio de congenericidad: moléculas similares tienen actividades/propiedades similares.

Medida apropiada de la bondad de ajuste, robustez y predictividad

La evaluación de la calidad de un modelo se basa en saber si es robusto, si no está sobreajustado y fundamentalmente si tiene buena capacidad predictiva para la actividad/propiedad de nuevas moléculas. En este principio, las técnicas de validación interna aplicadas al grupo de calibración brindan la bondad de ajuste y robustez; mientras que el grupo de predicción provee una estimación de la predictividad del modelo.

Interpretación del mecanismo de acción de los descriptores

Este último principio busca dar una interpretación de los descriptores moleculares del modelo (siempre que sea posible) y la forma en que afectan a la predicción de la actividad/propiedad en estudio. La interpretación del

mecanismo de acción brinda mayor confianza al modelo desarrollado, principalmente para fines regulatorios. No obstante, no siempre es posible obtener tal interpretación desde un punto de vista científico, debido a la complejidad matemática involucrada en la definición de los descriptores moleculares.

2.2. Modelos moleculares

En la química computacional los compuestos químicos son representados mediante modelos moleculares. La creación de modelos matemáticos ha sido un tema de estudio que ha adquirido mayor importancia con el uso masivo de los computadores, acompañado de la evolución de las matemáticas y de la química y física teórica. De esta manera, cualquier molécula puede ser modelada a partir de un gráfico molecular (representación gráfica). Por esta razón, los programas que se emplean en la química computacional tienen una interfaz gráfica que facilita el diseño de los compuestos químicos [8].

Un modelo químico debe ser aplicable a cualquier tipo de molécula (sistema), independientemente del tamaño, por lo que la capacidad de cálculo debería ser el único limitante en esta clase de sistemas. Por lo tanto, el modelo teórico debe ser bien definido para una configuración dada de núcleos y electrones, de tal forma que permita una solución aproximada a la ecuación de Schrödinger. Una vez que se ha seleccionado el modelo teórico, se continúa con la implementación computacional en un programa específico. Un modelo químico tiene las siguientes características [32]:

- Consistencia de tamaño: debe reproducir el mismo resultado que se obtendría de la solución de la ecuación de Schrödinger.
- Ser variacional: las energías aproximadas no deben ser menores que la obtenida por la solución de la ecuación de Schrödinger.
- Ser eficiente: debe ser factible de implementarlo y calcular usando herramientas computacionales.
- Ser preciso: debe reproducir cuantitativamente los resultados experimentales.

2.2.1. Modelos moleculares bidimensionales

La forma más común de representar una estructura molecular de forma bidimensional (2D) es mediante los grafos moleculares [33,34].

Matemáticamente, el grafo $G = (V, E)$ es una representación de un grupo de vértices (V) y un grupo de aristas entre los vértices (E), donde los átomos corresponden a los vértices y las aristas a los enlaces químicos entre pares de átomos (Figura 2.3). De esta manera, la teoría de grafos permite aplicar algoritmos útiles y bien definidos para explorar las propiedades estructurales de los compuestos químicos [26]. Así, se pueden obtener matrices que dependen de las conexiones (moleculares) en el grafo, a partir de los cuales se obtienen diversos invariantes. Los invariantes del grafo molecular pueden ser valores simples, una secuencia de números o una característica polinomial. La teoría de invariantes de grafos moleculares es la base fundamental para la definición de diversos índices topológicos [26,35].

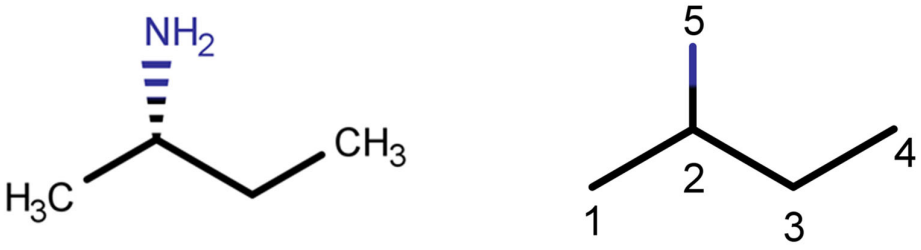

Figura 2.3 Representación de grafo molecular del 2–aminobutano

Por lo general, un grafo se representa como una colección de listas de adyacencia o como una matriz de adyacencia. Las listas de adyacencia son una colección de $|V|$ listas, una para cada *i*–ésimo átomo, donde cada lista $Adj[i]$ incluye los átomos conectados al *i*–ésimo átomo junto con el correspondiente orden de enlace. Por otra parte, la matriz de adyacencia **A** es una matriz cuadrada simétrica de dimensión $|V| \times |V|$, donde cada elemento $a_{ij} = 1$ si los átomos son adyacentes, es decir, si existe un enlace químico entre ellos.

A partir de la representación de grafo molecular se pueden calcular diversos tipos de matrices, las que reciben el nombre de matrices grafo–teóricas. Estas matrices permiten el cálculo de diversos tipos de descriptores moleculares para describir a las estructuras moleculares. La información que capturan estos descriptores va a depender del tipo de información que se incluye en dichas matrices. Por lo general, la mayoría de matrices grafo–teóricas se obtienen a partir de grafos moleculares libre de átomos de hidrógeno. Sobre esta representación grafo–teórica o matricial es posible

aplicar diversos operadores para calcular una amplia gama de descriptores moleculares. Las matrices grafo–teóricas más importantes son [26,33]:

Matrices de vértices

Son matrices cuadradas de dimensión $|V| \times |V|$, es decir, las filas y columnas representan a los átomos o vértices del grafo molecular, por lo que cada elemento codifica una propiedad asociada a un par de átomos presente en la estructura molecular. Aquí se encuentran las matrices de adyacencia, de distancias topológicas, Laplaciana y detour; a partir de las cuales se calculan diversos tipos de descriptores moleculares.

Matrices de aristas

Son matrices cuadradas de dimensión $|E| \times |E|$, cuyos elementos codifican la información de conexión entre pares de átomos, es decir, la información concerniente a los enlaces. En este grupo, la matriz de adyacencia $\mathbf{A}$ es la más importante.

Matrices de incidencia

Estas matrices brindan información concerniente a las relaciones entre dos diferentes conjuntos de objetos (átomos, enlaces, ciclos, subestructuras o trayectos moleculares), por lo tanto, no son matrices cuadradas y su dimensión dependerá de los objetos considerados.

2.2.2. Modelos moleculares tridimensionales

La representación tridimensional (geométrica) de la posición de los átomos de un compuesto brinda mayor información a la conectividad atómica. Este tipo de representación permite distinguir estructuras moleculares que presentan isomería. Los estereoisómeros están formados por los mismos átomos y enlaces, pero con disposición espacial distinta, por lo que tendrán actividades/propiedades completamente diferentes. En consecuencia, los compuestos químicos existen en diversas conformaciones de equilibrio, las cuales minimizan su energía, por lo que es importante definir la mejor conformación para el cálculo de los descriptores 3D [36].

El primer grupo de isómeros son los geométricos *cis*– (grupos funcionales a ambos lados de un doble enlace) y *trans*– (grupos funcionales en lados opuestos del doble enlace). Otro tipo de isomería se tiene cuando los cuatro

sustituyentes unidos a un átomo de carbono se pueden colocar en más de una forma. Este tipo de carbono se conoce como un centro quiral y en ocasiones genera dos moléculas que son la imagen especular la una de la otra (no se superponen). Por lo tanto, si dos compuestos son formas especulares que no se superponen, entonces se les conoce como enantiómeros; mientras que, si no son la imagen especular el uno del otro, se denominan diastereómeros. Para asignar la configuración a un carbono quiral, se jerarquizan los cuatro grupos sustituyentes: si el orden jerárquico va en sentido horario (hacia la derecha) se trata del isómero *R*–; mientras que si el orden jerárquico va en sentido antihorario (hacia la izquierda) se trata del isómero *S*– [37].

Los enantiómeros al ser expuestos a la luz polarizada interactúan con ella rotando su plano de luz en direcciones opuestas. Si la rotación es hacia la izquierda la molécula es levógira (*L*–), mientras que si la rotación es hacia la derecha es dextrógira (*D*–). Por otra parte, en el caso de los anillos bencénicos, existen tres tipos de isómeros posibles en cualquier benceno disustituido. Estos únicos tres isómeros se deben a la simetría planar del anillo aromático y se conocen como *orto*– (sustituyentes en las posiciones 1 y 2), *meta*– (sustituyentes en las posiciones 1 y 3) y *para*– (sustituyentes en las posiciones 1 y 4) [37].

Para obtener las coordenadas tridimensionales, la estructura química de los compuestos se modela en algún editor molecular (gratuito o comercial), por ejemplo, HyperChem [38] o MarvinSketch [39], entre otros. Los programas quimioinformáticos generan formatos computacionales específicos (SMILES, SYBYL o MDL), cuyos archivos almacenan la información de las posiciones tridimensionales de los átomos [8]. Complementariamente, se ha propuesto el lenguaje de marcado químico (CML) como formato estándar basado en el dialecto XML [40,41].

Las herramientas computacionales de modelización molecular permiten crear diversos modelos químicos. La mecánica molecular (MM) se fundamenta en las leyes de la mecánica clásica en la que las moléculas se tratan como conjuntos de átomos en el espacio (partículas puntuales dotadas de masa y carga), que se encuentran unidos entre sí por enlaces comparables a resortes y que se encuentran gobernados por un conjunto de funciones de potencial clásico [42,43]. De esta forma, los métodos de la mecánica molecular construyen una expresión para la energía potencial en función de

las posiciones atómicas V(x,y,z), en la que se analiza las contribuciones debidas a [32,43]:

- Alargamiento del enlace.
- Deformación del ángulo de enlace.
- Deformación fuera del plano.
- Rotación interna alrededor de un enlace (ángulo de torsión).
- Contribución cruzada debida a interacciones entre los 4 movimientos previamente citados.
- Atracciones y repulsiones de van der Waals entre los átomos no enlazados.
- Interacciones electrostáticas entre los átomos.

De esta manera, la energía total de una molécula se obtiene como la suma de estas contribuciones a la energía potencial. La MM es computacionalmente rápida en la ejecución de los cálculos y se puede aplicar a sistemas moleculares bastante grandes. El campo de fuerza MM+ es una extensión del método MM2, que primero intenta realizar un cálculo con los parámetros MM2 disponibles y luego usa un esquema predeterminado para trabajar cuando no existen parámetros para el campo de fuerza de la mecánica molecular versión 2 MM2(91) (falla en estas situaciones) [38].

Para los cálculos de la mecánica molecular se debe especificar las coordenadas iniciales de los átomos y la forma en que se encuentran conectados, esto se facilita cuando se usan programas con interfaz gráfica. De esta manera, la optimización se inicia con esta geometría de partida y continúa hasta encontrar un punto estacionario (mínima energía) en la superficie de energía potencial. La mecánica molecular se encuentra parametrizada por los valores que toman las constantes de fuerza y los valores geométricos en el equilibrio. Los cálculos de la mecánica molecular son computacionalmente menos demandantes con respecto a los métodos basados en la mecánica cuántica [32,42].

Por otra parte, los métodos mecano–cuánticos se basan en la solución de la ecuación de Schrödinger para describir el comportamiento de los núcleos y los electrones de un sistema molecular. En términos generales, la modelización tridimensional proporciona la geometría más estable, la energía, la distribución de cargas eléctricas o diferentes propiedades espectroscópicas del sistema en estudio, por ejemplo, espectros infrarrojo

(IR), Raman, resonancia magnética nuclear (NMR), ultravioleta–visible (UV–Vis), entre los más importantes. De igual forma, permite obtener datos termodinámicos y cinéticos, por ejemplo, calores de formación y de reacción, entropías, capacidades caloríficas y constantes de velocidad de una reacción química; así como diversas propiedades mecánicas (módulos elásticos y curvas de tensión–deformación) [42,43].

2.3. Descriptores moleculares

Los descriptores moleculares son las variables independientes con las cuales se predicen la actividad o propiedad de interés para un conjunto de moléculas. Por lo tanto, los científicos se han enfocado en la manera de capturar y convertir (de forma teórica) la información codificada dentro de la estructura química en números que se puedan usar para desarrollar modelos matemáticos predictivos para las actividades/propiedades [8,44].

2.3.1. Definición

Se define a un descriptor molecular como el resultado final de un procedimiento lógico y matemático que transforma la información química codificada en una representación simbólica de una molécula en un número útil o el resultado de algún experimento estandarizado [26]. El campo de los descriptores moleculares es amplio, por lo que se recurre a diferentes niveles de teoría para su desarrollo, por ejemplo, elementos de álgebra, teoría de grafos, teoría de la información, química computacional, teorías de reactividad química y fisicoquímica, entre otros [8,45].

2.3.2. Requisitos

Algunos descriptores moleculares se basan en teorías matemáticas complejas, sin embargo capturan únicamente una parte de la información química total contenida en una molécula. Por lo tanto, esta disciplina es un campo floreciente de investigación, y su evidencia está en el gran número de descriptores (miles) disponibles en la literatura y diversos programas específicos [46]. Debido a este hecho, se han propuesto ciertas reglas básicas que un descriptor molecular debe cumplir [8,47]:

1. Ser invariante al etiquetado y enumerado de los átomos.
2. Ser invariante a la roto–traslación de la molécula.

3. Ser definido por un algoritmo inequívoco.
4. Poseer una aplicabilidad bien definida en las estructuras moleculares.
5. Tener una interpretación estructural.
6. Tener buena correlación con al menos una propiedad experimental.
7. No tener relación trivial con otros descriptores.
8. No estar basado en propiedades experimentales.
9. Ser de preferencia continuo.
10. Poseer mínima degeneración.
11. Ser simple.
12. Ser aplicable a una amplia clase de moléculas.
13. Ser capaz de discriminar isómeros.
14. Poseer valores calculados en un rango numérico adecuado para el conjunto de moléculas sobre las cuales se utilizará.

Los primeros cuatro requisitos ayudan a entender si un descriptor molecular está bien definido; mientras que los siguientes requisitos se enfocan en el uso del descriptor, es decir, interpretabilidad, relación con al menos una propiedad experimental (pero no estrecha relación con los demás descriptores).

2.3.3. Tipos

Descriptores constitucionales

Son los descriptores más simples (0D) y engloban a aquellos que representan la estructura química sin considerar ni la topología ni la geometría [48]. Por ejemplo, el peso molecular, número de átomos (átomos terminales, heteroátomos y átomos ponderados), número de enlaces múltiples (dobles, triples y aromáticos), número de enlaces rotables, entre otros.

Descriptores de anillo

Estos descriptores codifican la información concerniente a la presencia de anillos en una molécula, es decir, estructuras cíclicas o anillos aromáticos [8]. Algunos descriptores que pertenecen a esta familia son el número de anillos, número de circuitos o ciclos, tamaño total del anillo, perímetro del anillo, conteo de puentes en el anillo, grado de ciclizado molecular, índice de complejidad del anillo o relación aromática.

Índices topológicos

Son descriptores bidimensionales que se obtienen de una representación topológica (grafo molecular) de los compuestos químicos, por lo tanto, no brindan ningún tipo de información sobre la distribución tridimensional de los átomos [26,35]. En este grupo se encuentra el número de trayectos moleculares, que cuenta el número de caminos, trayectos de ida y trayectos de ida y vuelta en un grafo molecular libre de hidrógenos usando diferentes longitudes topológicas [26]. En una segunda categoría se encuentran los índices de conectividad molecular [26,49], también calculados a partir de un grafo molecular libre de hidrógenos, en los que cada vértice se pondera por el grado de vértice (número total de átomos conectados). Estos índices tienen la siguiente forma general:

$$ {}^{m}\chi = \sum_{k=1}^{K} \left(\prod_{i=1}^{n} \delta_i \right)_k^{-1/2} \tag{2.2} $$

donde k recorre todos los subgrafos de orden m constituidos por n átomos ($n = m + 1$ para subgrafos acíclicos); K es el número total de trayectos de orden k presentes en el grafo molecular. Entre los principales descriptores de este tipo se tiene el índice de conectividad de Randić, índice modificado de Randić, índice de conectividad de Kupchik; así como los índices de conectividad molecular de Kier–Hall, índices de conectividad promedio, índices de conectividad de solvatación.

El tercer tipo constituyen los índices de información [26,30], los cuales indican el contenido de información de las moléculas aplicando diversos criterios para definir las clases de equivalencia (equivalencia de los átomos en una molécula), por ejemplo, la identidad química, formas de enlace en el espacio, topología molecular y simetría. Estos índices se derivan fundamentalmente a partir de un grafo molecular mediante la partición de sus elementos o los elementos de la matriz en clases de equivalencia.

Otro tipo importante constituyen los autovalores de Burden [50], los cuales se obtienen de la matriz de Burden $\mathbf{B}$ considerando un grafo molecular completo (incluido hidrógenos). Los autovalores de Burden se definen de la siguiente manera:

$$\left[\mathbf{B}\right]_{ij} = \begin{cases} \pi_{ij}^{*} \times 10^{-1} & si\left(i,j\right) \in E \\ Z_i & si\ i = j \\ 0.001 & si\left(i,j\right) \notin E \end{cases} \tag{2.3}$$

Los elementos diagonales B_{ii} son los números atómicos Z_i de los átomos. Los elementos no diagonales B_{ij} representan dos átomos enlazados que son iguales a $\pi^{*} \times 10^{-1}$, donde π^{*} es el orden de enlace convencional, igual a 1, 2, 3, y 1.5 para enlaces simples, dobles, triples y aromáticos, respectivamente.

Descriptores tipo P_VSA

Los descriptores tipo P_VSA [51] corresponden a una partición del área superficial de van der Waals condicionada por los valores atómicos de una determinada propiedad P dentro de un rango específico. Las propiedades que se utilizan son el coeficiente de reparto octanol–agua (logP), refractividad molar, masa, volumen de van der Waals, electronegatividad de Sanderson, polarizabilidad, energía de ionización y el estado intrínseco. Por otra parte, existen descriptores P_VSA que se calculan como la suma de las contribuciones del VSA de todos los átomos asignados a la presencia de potenciales sitios farmacóforos (PPPs); es decir, dado un determinado PPP, los descriptores P_VSA se calculan como la suma de las contribuciones VSA de todos los átomos asignados a aquel PPP [52].

Propiedades moleculares

Son descriptores heterogéneos que describen diversas características obtenidas por modelos teóricos [26]; por ejemplo, conteo de insaturaciones e índice de insaturación, factor hidrofílico, refractividad molar, área de superficie polar topológica, índice de densidad de empaquetamiento, coeficientes de reparto octanol–agua de Moriguchi y de Ghose–Crippen–Viswanadhan, volúmenes de van der Waals calculados a partir del volumen de McGowan y de la ecuación de Zhao–Abraham–Zissimos.

Fragmentos centrados en el átomo y número de grupos funcionales

Para el cálculo de estos descriptores se considera la composición química y las conectividades atómicas. Los fragmentos centrados en el átomo [53,54] cuentan los distintos tipos de átomos específicos (fragmentos) presentes en

un compuesto químico, en la que cada fragmento es un átomo en la molécula descrito por sus átomos vecinos. Por otro lado, el número de grupos funcionales [26] cuantifica la cantidad de estos tipos de átomos presentes en una estructura molecular.

Descriptores geométricos

Los descriptores geométricos brindan información adicional a la proporcionada por los descriptores constitucionales y topológicos. Estos descriptores se obtienen a partir de una representación tridimensional de la estructura molecular o a partir de un grafo 3D (índices topográficos), donde también se consideran las conexiones entre los mismos. En consecuencia, estos descriptores se calculan a partir de una geometría molecular optimizada por algún método computacional o a partir de las coordenadas cristalográficas [48,55].

En esta familia se encuentra a los descriptores de representación molecular 3D de estructuras basadas en la difracción de electrones (3D–MoRSE) [55,56], los que obtienen información mediante la transformación usada en los estudios de difracción de electrones para obtener curvas de dispersión teóricas. Otro tipo son los descriptores de función de distribución radial (RDF) [55,57], que brindan información sobre la distribución de probabilidad de encontrar un átomo dentro de un volumen esférico de radio R. Estos descriptores guardan similitud con la forma de cálculo de los descriptores 3D–MoRSE.

Por otro lado, los descriptores moleculares invariantes holísticos ponderados (WHIM) [55,58] brindan información importante relacionada con el tamaño, forma y simetría de la molécula, así como la distribución de los átomos con respecto a los ejes principales de la molécula (marco de referencia invariante). El cómputo de estos descriptores consiste en encontrar los autovalores y autovectores de la matriz de varianzas–covarianzas ponderada de las coordenadas cartesianas de la molécula. De esta forma, cada autovalor brinda información sobre el tamaño de una molécula a lo largo del eje principal.

Los descriptores de localización atómica holística ponderada y de forma de entidad (WHALES) [59] se han desarrollado como una forma de facilitar el andamio de salto «scaffold hopping» de productos naturales a compuestos sintéticos isofuncionales. El andamio de salto se refiere a la búsqueda de compuestos que presenten actividad similar pero que contengan diferentes

estructuras centrales (andamios). El enfoque holístico permite a estos descriptores capturar diversas propiedades moleculares contemporáneamente: distancias interatómicas geométricas, forma molecular y propiedades atómicas (por ejemplo la distribución de carga parcial).

Los descriptores de ensamblado de pesos de átomos, geometría y topología (GETAWAY) [55,60] se obtienen de la matriz de influencia ($\mathbf{H}=\mathbf{M}(\mathbf{M}^T\mathbf{M})\mathbf{M}^T$) construida con la matriz de información molecular ($\mathbf{M}$). Los elementos diagonales de la matriz $\mathbf{H}$ son los valores de influencia (entre 0 y 1), que codifican la información de la influencia de cada átomo de la molécula en la determinación de su forma con respecto al centro geométrico. En otras palabras, átomos cercanos al centro tendrán baja influencia, mientras que átomos en la periferia de la molécula mostrarán altos valores. De esta forma, estos descriptores son sensibles a los cambios conformacionales y las longitudes de los enlaces.

Descriptores topo–geométricos

En este grupo se encuentran los descriptores que se pueden calcular ya sea a partir de una representación topológica o una representación geométrica de la molécula.

Los pares de átomos [61] son descriptores que consideran a pares de átomos en una molécula con su respectiva separación interatómica. Así, los pares de átomos 2D utilizan la distancia topológica, mientras que los pares de átomos 3D usan la distancia euclidiana. Existen dos tipos: pares de átomos binarios (presencia/ausencia) y pares de átomos de frecuencia. De forma análoga, los descriptores de búsqueda de plantillas químicamente avanzadas (CATS) [62] están relacionados con la presencia de potenciales sitios farmacóforos (PPPs): donante de enlaces de hidrógeno (D), aceptor de enlaces de hidrógeno (A), positivo (P), negativo (N) y lipofílico (L). En consecuencia, los descriptores CATS2D consideran múltiples PPPs separados por una cierta distancia topológica; mientras que los descriptores CATS3D no permiten la asignación de múltiples potenciales sitios farmacóforos y utilizan la distancia euclidiana en el espacio tridimensional.

Los descriptores basados en la matriz 2D y matriz 3D [26] son índices topológicos e índices topográficos, respectivamente, que se calculan mediante la aplicación de un conjunto de operadores algebraicos. Para el cálculo de los descriptores 2D se usan matrices grafo–teóricas obtenidas a partir de un

grafo molecular libre de hidrógenos y la distancia topológica, mientras que para el cálculo de los descriptores 3D se usa la matriz de distancias geométricas obtenida a partir de un grafo molecular completo (incluido hidrógenos) y la distancia euclidiana.

Las autocorrelaciones 2D y las autocorrelaciones 3D [26,48] describen cómo se distribuye una determinada propiedad a lo largo de la estructura molecular. Las autocorrelaciones bidimensionales se obtienen a partir de un grafo molecular completo ponderado por las propiedades fisicoquímicas escaladas con respecto al valor del átomo de carbono y la distancia topológica; mientras que las autocorrelaciones tridimensionales se basan en las distancias euclidianas entre los átomos presentes en la superficie molecular.

2.3.4. Fragmentos moleculares

Claves moleculares del sistema de acceso molecular

Las claves moleculares del sistema de acceso molecular (MACCS) [48,63,64] son dos grupos de fragmentos moleculares que comprenden 960 y 166 características estructurales, respectivamente. Las claves MACCS han sido diseñadas y optimizadas para la búsqueda de diversas subestructuras. Las 166 claves moleculares MACCS son vectores boleanos de tamaño fijo que reflejan la presencia/ausencia de un grupo de características moleculares bien definidas.

Claves moleculares de PubChem

Las claves moleculares de PubChem [48,65] han sido concebidas como un grupo bien definido de fragmentos moleculares que se muestran como una lista ordenada de bits para indicar la presencia/ausencia de subestructuras moleculares específicas. Algunos ejemplos son el conteo de átomos o fragmentos, la presencia de anillos o pares de átomos, entre otros. Estas claves moleculares se usan en la quimioteca de acceso libre PubChem para la búsqueda de compuestos químicos con diversos niveles de similitud molecular.

2.3.5. Huellas dactilares moleculares

Las huellas dactilares moleculares (FPs) [48,66] describen a un compuesto mediante la captura de diversos aspectos locales de la estructura molecular.

Las huellas dactilares moleculares codifican la información de la estructura química de un compuesto mediante la obtención de todos los fragmentos (subestructuras posibles), los cuales generan vectores booleanos de dimensión fija que definen un conjunto de patrones (bits). Existen dos tipos principales de patrones que se pueden identificar: 1) átomos centrados (subestructuras circulares) que definen a las huellas dactilares moleculares de conectividad ampliada (ECFPs) y 2) trayectos (de longitud predefinida) que conciben a las huellas dactilares moleculares de trayecto (PFPs). Debido a que el número de los distintos fragmentos obtenidos para cualquier molécula puede ser muy largo, las FPs se procesan mediante una función de resumen (hash), de tal forma de reducir la longitud a un valor predefinido. En las huellas dactilares las características estructurales pueden colisionar, es decir, dos o más fragmentos de diversa naturaleza activan un mismo bit.

2.4. Modelos QSAR para el gusto de moléculas

La industria de alimentos (y farmacéutica) tienen gran interés en el descubrimiento de nuevos edulcorantes que puedan tener propiedades beneficiosas. Por consiguiente, los químicos tienen el reto de diseñar moléculas que presenten un sabor dulce puro, similar al de la Sacarosa. El uso de edulcorantes bajos en calorías (sin retrogusto) en productos alimenticios y medicinas es importante para personas que padecen diabetes [1,4]. La medición experimental del gusto se realiza mediante panelistas entrenados (o semientrenados), a los que se les entrega soluciones estándar para el gusto junto a soluciones de los compuestos sintetizados de interés. Los panelistas usan la metodología de análisis sensorial de “beber y escupir” «sip and spit». De esta manera, se asigna un gusto o mezcla de gustos a cada compuesto en estudio [6,67,68].

El descubrimiento de nuevos compuestos de gusto es complejo y costoso. Por un lado, existen múltiples factores que afectan a la percepción del mismo; por ejemplo, concentración, solubilidad, estabilidad en un amplio rango de pH y temperatura, gusto puro sin retrogusto desagradable, beneficio económico para usos industriales y, el más importante, seguridad e inocuidad para el consumo humano [69].

Ciertos compuestos químicos dulces aceptados para consumo humano han sido descubiertos por casualidad, por ejemplo, Sacarina (CAS 128–44–9 y PubChem CID 5143), Ciclamato (CAS 100–88–9 y PubChem CID 7533) y

Aspartamo (CAS 22839–47–0 y PubChem CID 134601) [8]. En otros casos, se ha observado que durante la síntesis de nuevos blancos moleculares, algunas variaciones de los radicales en el esqueleto químico de la molécula hacen que un compuesto deje de ser dulce (amargo, insípido, ácido, salado y umami) [1].

Por esta razón, existen ventajas en el desarrollo racional (descubrimiento) de compuestos con un gusto particular. Las relaciones cuantitativas estructura–actividad/propiedad son técnicas *in silico* que ayudan al desarrollo y síntesis de nuevos y más potentes edulcorantes [6,8,69]. En la Tabla 2.1 se presenta una revisión histórica hasta el año 2010 de los diversos modelos QSAR desarrollados para discriminar los gustos de diverso tipo de moléculas. Los modelos propuestos se han enfocado principalmente en discriminar compuestos dulces y amargos y moléculas dulces y no dulces.

Por otra parte, en la última década, ha existido aún mayor interés por sintetizar nuevos compuestos con un gusto particular, así como desarrollar bases de datos cada vez más extensas con las que se han desarrollado nuevos modelos basados en las relaciones cuantitativas estructura–actividad/propiedad, utilizando nuevos enfoques del aprendizaje automático. A continuación, se describirán con mayor detalle los modelos desarrollados en la última década (Tabla 2.2). En las Tablas 2.1 y 2.2, *ML* es el tipo de aprendizaje automático utilizado, *d* es el número de descriptores moleculares incluido en el modelo, n_{cal} y n_{pred} indica el número de moléculas en el grupo de calibración y predicción, respectivamente; mientras que NER_{cal} y NER_{pred} representa la tasa de aciertos en calibración y predicción, respectivamente. *NER* también se conoce como exactitud balanceada (Balanced Accuracy). Este parámetro se define en la sección 3.3.6 (Parámetros de evaluación en clasificación).

Tabla 2.1 Revisión cronológica de los modelos QSAR desarrollados para discriminar los gustos hasta el año 2010

Modelos	Gustos	Clases	ML	d	n_{cal}	n_{pred}	NER_{cal}	NER_{pred}
[13]	Dulce–Amargo	2	Regresión	3	49	—[a]	—[a]	—[a]
[14]	Dulce–Amargo	2	LDA	2	20	9	0.850	0.775
[70]	Dulce–No dulce	2	Gráfico	2	35	12	0.914[b]	0.917[b]
[71]	Dulce–Amargo	2	LLA	3	22	—[a]	1	—[a]
			*k*NN	6		—[a]	0.909	—[a]
[72]	Dulce–Amargo	2	LDA	3	33	—[a]	0.807	—[a]
[73]	Dulce–Amargo	2	LDA	3	22	9	1	0.775
				2			0.955	0.775
[74]	Dulce–Amargo	3	SIMCA	5	91	—[a]	0.840	—[a]
[75]	Dulce–No dulce	2	SIMCA	4	50	—[a]	0.798	—[a]
[76]	Dulce–No dulce	2	SIMCA	1[c]	25	—[a]	0.868	—[a]
					20	—[a]	0.808	—[a]
[77]	Dulce–No dulce	2	LDA	3	23	—[a]	0.642	—[a]
[78]	Dulce–No dulce	3	Gráfico	2	57	—[a]	0.860	—[a]
		2	LDA	3	33	—[a]	0.848[b]	—[a]
					23	—[a]	0.870[b]	—[a]
[67]	Dulce–No dulce	2	Gráfico	2	40	—[a]	0.833	—[a]
[79]	Dulce–Amargo	3	DA	11[c]	50	—[a]	1	—[a]
[80]	Dulce–No dulce	2	LDA	4	101	—[a]	0.665	—[a]
			QDA			—[a]	0.801	—[a]
			CART	3		—[a]	0.650	—[a]
[81]	Dulce–Amargo	2	Gráfico	2	23	—[a]	0.862	—[a]
			LDA	4		—[a]	0.850	—[a]
			QDA			—[a]	0.900	—[a]
[82]	Dulce–No dulce	2	LDA	4	132	—[a]	0.693	—[a]
			QDA			—[a]	0.683	—[a]
			CART	3		—[a]	0.815	—[a]
[83]	Dulce	3	LDA	8	75	8	0.547[b]	0.500[b]
			QDA				0.773[b]	0.250[b]
			CART				0.773[b]	—[a]
[84]	Amargo–No amargo	2	Naïve Bayes	10	14179	—[a]	0.805	—[a]
					287	—[a]	0.602	—[a]
[85]	Dulce	3	CART	6	82	—[a]	0.753	—[a]
				7		—[a]	0.580	—[a]
[86]	Dulce–No dulce	2	LDA	6	70	12	0.810	0.583[b]
				2	58	—[a]	0.655[b]	—[a]
			QDA	3		—[a]	0.759[b]	—[a]
		2	CART	6	48	10	0.950	0.700
		3	CART	6		10	0.908	0.611

[a] no disponible; [b] calculado como exactitud «*Acc*»; [c] número de componentes principales (PCs)

En el año 2016 se proponen dos modelos QSAR basados en el clasificador de los *k*–vecinos más cercanos (*k*NN) para discriminar los gustos dulce y amargo y los compuestos dulces e insípidos [87]. La base de datos dulce–amargo está constituida por 508 compuestos (427 dulces y 81 amargas) y la base de datos dulce–insípido está formada por 566 molécula (433 dulces y 133 insípidas). Para cada estructura molecular se han utilizado 2164 descriptores independientes de la conformación calculados en el programa Dragon, de los cuales se retuvieron 855 descriptores luego de aplicar la reducción no supervisada V–WSP a un umbral de correlación de 0.95. Seguidamente, ambas bases de datos se dividieron en conjuntos de calibración (70%) y predicción (30%) de forma casual y proporcional a la numerosidad de las clases. Con el conjunto de calibración se estableció el modelo QSAR usando el método de los *k*NN acoplado con la selección supervisada de los algoritmos genéticos (GAs), para lo cual se usó la validación cruzada de ventanas venecianas con 5 grupos, de tal forma de seleccionar el valor óptimo de k mediante maximización de la tasa de aciertos (NER_{cv}). El modelo dulce–amargo está constituido por 4 descriptores (NER_{cal} = 0.864, NER_{cv} = 0.861 y NER_{pred} = 0.789); mientras que el modelo dulce–insípido está definido por 9 descriptores moleculares (NER_{cal} = 0.838, NER_{cv} = 0.847 y NER_{pred} = 0.752).

Contemporáneamente, se propone la plataforma de acceso libre BitterX [88], en el que se proponen dos modelos basados en máquinas de soporte vectorial (SVM) para la predicción del amargor de diversas moléculas recopiladas de las quimiotecas PubMed y BitterDB. Para cada compuesto, se obtuvieron los archivos de estructura molecular (descriptores moleculares) de PubChem. El primer modelo (basado en el ligando) se desarrolló a partir de una base de datos de 539 compuestos amargos y 539 compuestos no amargos: 20 obtenidos de forma experimental y los restantes del directorio ACD (Available Chemicals Directory). De los 1078 compuestos se seleccionaron de forma aleatoria 862 moléculas (50% positivas y 50% negativas) para entrenar el modelo; mientras que las restantes 216 se usaron para medir la capacidad predictiva del mismo. Se acoplaron las SVM con los algoritmos genéticos (GAs) para seleccionar 46 descriptores fisicoquímicos. Con esta misma base de datos se obtuvieron dos nuevas particiones aleatorias de calibración y predicción (con el mismo número de moléculas) para evitar el sesgo debida a la división. La exactitud promedio de los tres modelos en calibración es $Acc = 0.879$ y en predicción es $Acc = 0.915$. Para

el segundo modelo (basado en el reconocimiento del receptor TAS2R) se usaron 260 moléculas positivas y 260 negativas, que se dividieron aleatoriamente en grupos de calibración y predicción en una proporción 4:1. Se usó el mismo diagrama de flujo que para el modelo basado en el ligando, obteniendo 35 descriptores (20 fisicoquímicos y 15 características del receptor). La exactitud promedio de los tres modelos en calibración es $Acc = 0.767$ y en predicción es $Acc = 0.798$.

Tabla 2.2 Modelos QSAR desarrollados en la última década para discriminar el gusto de moléculas

Modelos	Gustos	Clases	ML	d	n_{cal}	n_{pred}	NER_{cal}	NER_{pred}
[87]	Dulce–Amargo	2	kNN	4	356	152	0.864	0.789
	Dulce–Insípido			9	396	170	0.838	0.752
[88]	Amargo–No amargo	2	SVM	46	862	216	0.879[a]	0.915[a]
				35	416	104	0.767[a]	0.798[a]
[5]	Dulce–No dulce	2	Sistema experto	12	488	161	0.892	0.848
[23]	Amargo–No amargo	2	AdaBoost	16[b]	1827	781	0.921	0.812
[89]	Dulce–Amargo	2	RF	5[c]	796	200	0.997	0.914
[90]	Amargo–No amargo	2	Consenso	—[d]	1040	259	—[d]	0.929[a]
[9]	Dulce–Amargo	2	RF	—[d]	961	241	0.950[a]	0.967[a]
[91]	Dulce–No dulce	2	Consenso	—[d]	883	221	0.870	0.900
[24]	Amargo–No amargo	2	RF	—[d]	2257	154	0.754	0.819
	Dulce–No dulce		AdaBoost	—[d]	2205	161	0.856	0.834
[92]	Amargo–No amargo	2	SVM	36	512	128	0.930[a]	0.918[a]

[a] calculado como exactitud «*Acc*»; [b] descriptores con la contribución más significativa; [c] descriptores para la profundidad del árbol; [d] no disponible

En un estudio publicado en el año 2017 se fusionaron las bases de datos dulce–amargo y dulce–insípido previamente descritas para desarrollar un modelo QSAR para clasificar moléculas dulces y no dulces (fusión de amargo e insípido) basado en un sistema experto (aprendizaje semi–supervisado) [5]. La base de datos curada está formada por 649 moléculas (435 dulces y 214 no dulces). Cada estructura molecular fue representada por 875 descriptores moleculares independientes de la conformación y 2048 huellas dactilares moleculares de conectividad ampliada (ECFPs) calculados en el programa Dragon. Para la validación, la base de datos se dividió de forma casual y proporcional a la numerosidad de las clases en conjuntos de calibración (70%) y predicción (30%). En una primera etapa, se utilizó el análisis de similitud molecular basado en el escalado multidimensional (MDS) con las ECFPs, aquí se identificaron 2 grupos consistentes de moléculas dulces. En

una segunda etapa, utilizando el grupo donde las moléculas dulces y no dulces se superponen, se aplicaron modelos de clasificación basados en los *N*–vecinos más cercanos (N3) y el análisis discriminante de mínimos cuadrados parciales (PLSDA) acoplados con los GAs como estrategia de selección supervisada de los descriptores. Durante esta etapa se utilizó la validación cruzada basada en ventanas venecianas con 5 grupos de tal forma de maximizar (optimizar) la tasa de aciertos (NER_{cv}) y definir el parámetro alfa para N3 y las variables latentes (LVs) para PLSDA. De esta manera, el modelo basado en el sistema experto ensambla el análisis de similitud molecular con el consenso entre los dos métodos de clasificación ($NER_{cal} = 0.892$, $NER_{cv} = 0.887$ y $NER_{pred} = 0.848$).

Paralelamente, se propone un clasificador de aprendizaje automático denominado BitterPredict [23], el cual permite predecir si un compuesto es amargo o no en función de su estructura química. La base de datos está constituida por 2608 compuestos, divididos en 691 amargos (632 de BitterDB) y 1917 moléculas no amargas de distintas quimiotecas y referencias. Para cada estructura se calcularon 59 descriptores moleculares en diferentes programas y la base de datos se dividió de forma aleatoria en un conjunto de calibración (70%) y predicción (30%), manteniendo la proporción en la numerosidad de las dos clases. El modelo se calibró con el algoritmo de aprendizaje automático de "impulso adaptativo" AdaBoost «Adaptive Boosting», basado en árboles de decisión, con buenos resultados en calibración ($NER_{cal} = 0.921$ y $Acc_{cal} = 0.928$) y predicción ($NER_{pred} = 0.812$ y $Acc_{pred} = 0.832$). Posteriormente, el modelo se utilizó para la predicción prospectiva de 1553 moléculas de la quimioteca DrugBank (1375 moléculas dentro de dominio del gusto amargo), 20661 compuestos de FooDB (13588 dentro del dominio del amargor) y 28217 de la base de datos de productos naturales.

En el mismo año 2017, se aplicó el clasificador de bosques aleatorios (RF) para discriminar los compuestos dulces y amargos [89]. El modelo QSAR se desarrolló utilizando las bases de datos SweetenersDB (316 moléculas) y BitterDB (680 compuestos). Se emplearon 244 descriptores 2D calculados en el programa Dragon. La base de datos se dividió en un grupo de calibración con el 80% de las moléculas y de predicción con el 20% de los datos restantes. El modelo de RF se construyó con cien árboles, con una profundidad de árbol de 5 descriptores y el criterio de división de Gini. En el grupo de calibración, 253 moléculas dulces y 540 amargas fueron correctamente

clasificadas y únicamente 3 compuestos amargos fueron mal clasificados (NER_{cal} = 0.997 y MCC_{cal} = 0.990); mientras que en el conjunto de predicción 54 moléculas dulces y 133 amargas fueron correctamente asignadas y solo 4 compuestos amargos y 9 dulces mal clasificados (NER_{pred} = 0.914 y MCC_{pred} = 0.850). Adicionalmente, también se han desarrollado modelos del aprendizaje supervisado de RF y SVR para la predicción del dulzor relativo de los 316 compuestos de la base de datos SweetenersDB.

En el año 2018, se propone la plataforma e–Bitter [90] para la discriminación de compuestos amargos y no amargos. Para este propósito, se recopiló una base de datos completamente experimental y se definieron cuatro criterios para el curado de las moléculas. Posterior al curado, se obtuvieron 707 compuestos amargos y 592 no amargos (132 insípidos, 17 no amargos y 443 dulces), los cuales fueron divididos en grupos de calibración y predicción de la siguiente manera: se seleccionaron aleatoriamente 20% de los compuestos de cada clase (141 amargos y 118 no amargos) para constituir el grupo de predicción y los restantes (566 amargos y 474 no amargos) formaron el grupo de calibración. Las moléculas fueron representadas por distintas huellas dactilares moleculares de conectividad ampliada (ECFPs): 1024bit–ECFP4, 2048bit–ECFP4, 1024bit–ECFP6 y 2048bit–ECFP6. Los modelos se calibraron con los siguientes métodos: *k*NN, SVM, máquina de impulso de gradiente (GBM), RF, y dos redes neuronales profundas (DNN2 y DNN3). Para reducir el sesgo debida a la división de la base de datos, el procedimiento de partición se repitió 19 veces para los modelos con *k*NN, SVM, GBM y RF y solo tres veces para los modelos con DNN2 y DNN3 (computacionalmente demandantes). Se aplicó la validación de *k*–grupos de validación cruzada con 5 grupos y la aleatorización–Y. De esta forma, se generaron 1312 modelos individuales y 96 modelos de clasificación promediados. No obstante, se han propuesto nueve modelos de consenso (CM01–CM09) basados en el balance entre la precisión (*Pr*), velocidad y diversidad de métodos de aprendizaje automático. Entre estos modelos, el mejor es el CM01 con una exactitud en predicción $Acc = 0.929$.

Paralelamente, se publica un modelo de RF, denominado BitterSweetForest [9], para distinguir moléculas dulces y amargas en un algoritmo implementado en KNIME. Se utilizó una base de datos de 1202 compuestos tomados de SuperSweet (517 edulcorantes) y BitterDB (685 moléculas). Todas las estructuras fueron estandarizadas siguiendo diferentes criterios en el programa Instant JChem, para posteriormente representarlas

mediante cuatro tipos diferentes de huellas dactilares calculadas en el nodo RDKit: huellas de Morgan (2048 bits), pares de átomos (1024 bits), huellas de torsión (1024 bits) y huellas de Morgan Feat (2048 bits). La base de datos se dividió aleatoriamente en un grupo de calibración con 961 moléculas (80%) y de predicción con los restantes 241 compuestos (20%), asegurando la distribución original de las dos clases. Para el desarrollo de los modelos de RF se utilizaron los nodos Tree Ensemble Learner y Predictor, usando 100 árboles y el índice de Gini como criterio de división. Además, se utilizó una función de la raíz cuadrada para el atributo de muestreo y otros diversos atributos se fijaron para todos los árboles. Para evitar el sobreajuste del modelo se aplicó la validación cruzada de dejar–uno–fuera (LOO). El mejor modelo de RF se obtiene con las huellas dactilares de Morgan, para el cual la exactitud en validación cruzada LOO es $Acc = 0.950$ y la exactitud en predicción es $Acc = 0.967$.

En el año 2019, se propone la plataforma e–Sweet [91] para la discriminación entre compuestos dulces y no dulces, así como la predicción del dulzor relativo de diversos edulcorantes. La base de datos inicial está compuesta por 530 compuestos dulces y 850 no dulces (718 amargos de BitterDB y 132 insípidos), los cuales se sometieron a un proceso de curado. Cada molécula ha sido representada por 4 tipos de ECFPs (1024bit–ECFP4, 2048bit–ECFP4, 1024bit–ECFP6 y 2048bit–ECFP6). Para efectos de validación, la base de datos se dividió aleatoriamente en un conjunto de calibración y un conjunto de predicción de la siguiente manera: 80% de edulcorantes (339 moléculas) y 80% de no edulcorantes (544 compuestos) forman el grupo para la validación cruzada (calibración), mientras que el resto de los 221 compuestos se utilizan como grupo de predicción. Esta división se repitió varias veces para reducir el sesgo debido a la división aleatoria. Posteriormente, se usaron cinco métodos de aprendizaje automático: *k*NN, SVM, RF, GBM DNN, con los cuales se desarrollaron 1312 modelos individuales de clasificación (incluyendo 328 sin selección supervisada de descriptores y 984 con selección supervisada) y 96 modelos de clasificación promediados. Se aplicó la validación *k*–grupos de validación cruzada con 5 grupos y la aleatorización–Y. Sin embargo, es poco eficaz utilizar todos los modelos de clasificación, por lo que se sugieren cuatro modelos de consenso (CM01–CM04) en función del rendimiento, velocidad y diversidad de los modelos. Entre los 4 modelos, el mejor es el CM02 (NER_{cv} = 0.870, $F1_{cv}$ = 0.850, NER_{pred} = 0.900, $F1_{pred}$ = 0.880, MCC_{pred} = 0.810,

Sn_{pred} = 0.860, Sp_{pred} = 0.940 y Pr_{pred} = 0.900). Adicionalmente, se realizó la predicción del dulzor relativo (RS) mediante regresión para una base de datos compuesta por 352 edulcorantes.

Contemporáneamente, se publicaron diversos modelos del aprendizaje automático, denominados BitterSweet [24], para clasificar compuestos amargos y dulces. Para este propósito, se recopiló un conjunto extenso de 918 compuestos amargos, 1510 no amargos, 1205 dulces y 1171 no dulces. Las moléculas se han descrito mediante descriptores moleculares 2D/3D y ECFPs calculados en Dragon, propiedades fisicoquímicas y propiedades ADMET en el programa Canvas, así como propiedades estructurales y descriptores fisicoquímicos del programa ChemoPy. La base de datos amargo–no amargo se dividió en grupos de calibración con 2257 moléculas y de predicción con 154 compuestos; mientras que para el modelo dulce–no dulce se consideraron 2205 compuestos para la calibración y 161 moléculas para la predicción. Para el desarrollo de los modelos, se combinaron los grupos de descriptores previamente descritos con dos métodos de reducción no supervisada (algoritmo de Boruta y el PCA) y los métodos del aprendizaje automático de los RF, regresión logística RIDGE (RLR) y AdaBoost (árboles de decisión). Se ha aplicado la validación de *k*–grupos de validación cruzada con 5 grupos. El mejor modelo para la base de datos amargo–no amargo se tiene con los descriptores de ChemoPy, el PCA y los RF (NER_{cv} = 0.754, $F1_{cv}$ = 0.698, Sn_{cv} = 0.719, Sp_{cv} = 0.789, NER_{pred} = 0.819, $F1_{pred}$ = 0.838, Sn_{pred} = 0.790 y Sp_{pred} = 0.848). Por otra parte, para la base de datos dulce–no dulce el mejor modelo resulta de la combinación de los descriptores 2D/3D de Dragon, el algoritmo de reducción de Boruta y el aprendizaje supervisado AdaBoost (NER_{cv} = 0.856, $F1_{cv}$ = 0.858, Sn_{cv} = 0.853, Sp_{cv} = 0.859, NER_{pred} = 0.834, $F1_{pred}$ = 0.856, Sn_{pred} = 0.790 y Sp_{pred} = 0.878). Adicionalmente, los modelos BitterSweet se han aplicado para predecir el gusto de moléculas de las quimiotecas FlavorDB, FooDB, SuperSweet, Super Natural II, DSSTox y DrugBank.

Recientemente, se publicó el servidor web gratuito iBitter–Fuse [92], como una herramienta para identificar péptidos amargos. La base de datos, denominada BTP640, está constituida por 320 secuencias de péptidos amargos y 320 no amargos, los cuales se dividieron de forma aleatoria (en una relación 8:2) en un conjunto de calibración con 512 compuestos (256 de cada clase) y otro de predicción con 128 secuencias (64 de cada clase). Cada secuencia de péptidos ha sido representada por 544 índices de aminoácidos

(AAIs) obtenidos a partir del Amino acid index database (AAindex) versión 9.0. Luego de descartar los AAIs que tenían valores faltantes (NA), se utilizaron 531 AAIs como descriptores moleculares para el desarrollo del modelo basado en las SVM. Para ahorrar tiempo y recursos computacionales, se ha utilizado la selección supervisada basada en un algoritmo genético utilizando un informe de autoevaluación (GA–SAR), de tal forma de determinar el número óptimo de características informativas, a medida que se maximiza el rendimiento del modelo. Se ha utilizado la validación de k–grupos de validación cruzada con 10 grupos y se han obtenido 10 diferentes modelos, a partir de los cuales se seleccionó uno solo como el óptimo, con buen desempeño en validación cruzada ($Acc = 0.918$, $Sn = 0.918$, $Sp = 0.918$ y $MCC = 0.837$) y en predicción ($Acc = 0.930$, $Sn = 0.938$, $Sp = 0.922$ y $MCC = 0.859$).

Capítulo 3

APRENDIZAJE AUTOMÁTICO

3.1. Introducción

El aprendizaje automático, aprendizaje de máquina o máquinas de aprendizaje «machine learning» (ML) [93–96] son algoritmos que permiten que las computadoras aprendan de los datos; es decir, que tengan la capacidad de recibir datos de entrada y utilizar el análisis estadístico para predecir una salida, las que se actualizan a medida que existen nuevos datos disponibles. El aprendizaje automático se divide en cuatro clases principales:

1. Aprendizaje no supervisado: cada objeto de entrada se encuentra descrito por diversos tipos de variables (matriz de datos $\mathbf{X}$), sin tener asociado una respuesta. Por lo tanto, el objetivo es reducir la dimensionalidad de los datos y reconocer patrones (agrupamientos).
2. Aprendizaje supervisado: para este aprendizaje se requiere, además de la matriz de datos $\mathbf{X}$, un vector respuesta $\mathbf{Y}$, que es una variable aleatoria continua cuando se trata de regresión o de una variable aleatoria categórica en clasificación.
3. Aprendizaje semi–supervisado: es una combinación del aprendizaje no supervisado y supervisado. Los modelos basados en este tipo de aprendizaje se denominan «sistemas expertos».
4. Aprendizaje reforzado: es un algoritmo que aprende interactuando con su entorno, recibiendo recompensas por su correcto desempeño y castigo cuando su desempeño es incorrecto. El algoritmo aprende sin la intervención del humano, maximizando su recompensa y minimizando su penalización.

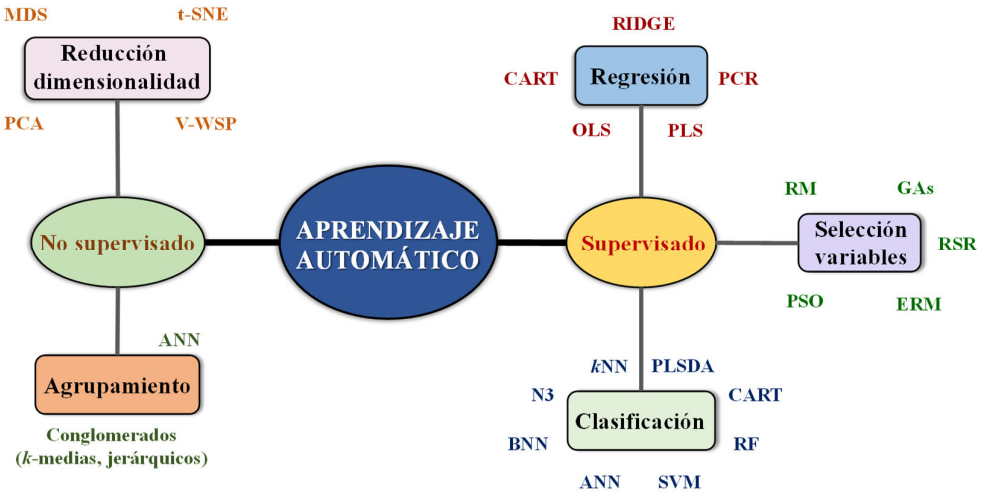

Figura 3.1 Diversas técnicas del aprendizaje automático para el desarrollo de modelos QSAR

El aprendizaje automático es la principal herramienta computacional para el desarrollo de las relaciones cuantitativas estructura–actividad/propiedad. En la Figura 3.1 se presentan las principales técnicas del ML que se utilizan para el desarrollo de los modelos QSAR. Si la variable respuesta es cuantitativa continua, es decir toma cualquier valor dentro de la escala de medida de dicha respuesta (por ejemplo el dulzor relativo o el índice de retención cromatográfico) los modelos QSAR/QSPR son de regresión. Por otro lado, si la variable respuesta es cualitativa nominal; es decir, se presenta en forma de categorías no ordenadas (por ejemplo compuestos dulces, amargos ácidos, umami o salados), los modelos QSAR/QSPR son de clasificación [8].

3.2. Aprendizaje no supervisado

3.2.1. Análisis de componentes principales

El análisis de componentes principales (PCA) es una de las técnicas de mayor uso para visualizar la estructura multivariada de los datos [97–99]. En el PCA las variables que describen los objetos se transforman, mediante combinaciones lineales, en nuevas variables denominadas componentes principales (PCs). La principal característica de las componentes es la de ser ortogonales entre sí, y se cumple que la PC1 tiene la máxima varianza, la

PC2 la segunda máxima varianza y así de forma decreciente hasta la última componente. El PCA permite reducir la dimensionalidad de los datos mediante la selección de un número de componentes reducido, por ejemplo, dos o tres componentes. Esta selección se realiza en función de la varianza explicada por cada una. El PCA también se utiliza para:

- Analizar la correlación entre las variables y su importancia.
- Identificar datos atípicos «outliers», grupos y/o clases.
- Mejorar la descripción de los datos mediante la eliminación de información inútil.
- Reduce la dimensionalidad y permite visualizar los datos en un sistema bidimensional o tridimensional.
- Define una representación de los datos dentro de un espacio ortogonal.

En esta técnica se desarrolla una rotación de los datos originales. Para una matriz $\mathbf{X}$ de dimensiones $n \times p$, la rotación se realiza de manera tal que el primer nuevo eje (PC1) se oriente en la dirección de máxima varianza, el segundo (PC2) es perpendicular al primero y en la dirección de la máxima varianza remanente; y así sucesivamente para todas las p–ésimas componentes. Matemáticamente, el PCA consiste en la diagonalización de la matriz de varianzas–covarianzas $\mathbf{S}$ obtenida a partir de la matriz de datos centrada $\mathbf{X}_c$ (matriz cuyas columnas tienen media cero):

$$diag\left(\mathbf{S}\right) = \left[\frac{\mathbf{X}_c^T \mathbf{X}_c}{n-1}\right] \tag{3.1}$$

n–1 se usa para obtener un estimador insesgado de las varianzas–covarianzas de la población.

La diagonalización permite obtener una matriz diagonal de autovalores $\mathbf{\Lambda}$, de dimensión $p \times p$, cuyos elementos diagonales son los autovalores λ_m, ordenados de manera decreciente. También se obtiene una matriz de cargas $\mathbf{L}$, de dimensión $p \times m$, cuyas columnas son los autovectores l_m de la matriz de varianzas–covarianzas, es decir, los coeficientes del autovector correspondiente. De esta manera, la matriz de varianzas–covarianzas $\mathbf{S}$ se puede descomponer en las dos matrices $\mathbf{L}$ y $\mathbf{\Lambda}$ mediante la descomposición en valores singulares:

$$\mathbf{S} = \mathbf{L}\Lambda\mathbf{L}^{T} \tag{3.2}$$

De esta manera se puede representar la matriz de datos $\mathbf{X}$ en un nuevo espacio (ortogonal) de acuerdo con:

$$\mathbf{T} = \mathbf{X}\,\mathbf{L} \tag{3.3}$$

donde $\mathbf{L}$ es la matriz de rotación y $\mathbf{T}$ es la matriz de puntuaciones.

Esta resolución matemática permite observar la información de forma simplificada en el gráfico de puntuaciones «scores» o proyección de los objetos y en el gráfico de cargas «loadings» o proyección de las variables.

3.2.2. Escalado multidimensional

El escalado multidimensional (MDS) [100,101] es una técnica que permite visualizar los datos y entender qué tan cerca se encuentran los objetos entre sí. Esta técnica reconstruye las similitudes (o disimilitudes) mediante distancias, de tal forma de proyectarlas en un número reducido de dimensiones; es decir, un arreglo de los objetos de manera tal que se reproduzcan las distancias en el espacio original. Para este propósito, se analiza la matriz de distancias $\mathbf{D}$; donde los elementos no diagonales D_{xy} representan la distancia entre el par de objetos x e y en el espacio p–dimensional y $\widehat{D}_{xy}$ es la distancia entre el par x e y en el subespacio M–dimensional ($M<p$). Normalmente, el escalado multidimensional busca proyecciones en 2 o 3 dimensiones (M = 2 o 3). De esta manera, los ejes (coordenadas principales) en el subespacio M–dimensional minimiza la siguiente expresión:

$$L^{*} = \min\left[\sum_{x}\sum_{y} w_{xy}\left(D_{xy} - \widehat{D}_{xy}\right)^{2}\right] \tag{3.4}$$

donde w_{xy} son los elementos de una matriz de pesos estadísticos. La versión mayormente usada del MDS considera el caso particular $w_{xy} = 1/D_{xy}$, por lo que la expresión anterior se reduce a:

$$L^* = \min\left[\sum_{x<y}\frac{\left(D_{xy} - \widehat{D}_{xy}\right)^2}{D_{xy}}\right] \tag{3.5}$$

De esta manera, se puede realizar un gráfico de dispersión de las M dimensiones, dentro del cual se observan reproducidas las distancias originales de los objetos en el espacio multidimensional.

3.2.3. Incrustación de vecinos estocásticos

La incrustación de vecinos estocásticos (SNE) [102] transforma las distancias entre pares de objetos en un espacio de alta dimensión ($\mathbb{R}^n$) en probabilidades condicionales que representan la similitud entre ellos. Así, la probabilidad condicional $p_{j|i}$ es la similitud del elemento x_j dado x_i, es decir, que x_i considere a x_j como su vecino, considerando que los vecinos se eligieran en proporción a su densidad de probabilidad bajo una distribución gaussiana centrada en el punto x_i. Si los puntos se encuentran muy cerca entre sí, $p_{j|i}$ será alta; mientras que para puntos lejanos $p_{j|i}$ puede llegar a ser infinitesimal (debida a σ_i). Como el interés es cuantificar la similitud entre pares de objetos, $p_{i|i} = 0$. La probabilidad condicional, para $i \neq j$, se calcula como:

$$p_{j|i} = \frac{\exp\left(-\left\|x_i - x_j\right\|^2 / 2\sigma_i^2\right)}{\sum_{k\neq i}\exp\left(-\left\|x_i - x_k\right\|^2 / 2\sigma_i^2\right)} \tag{3.6}$$

donde σ_i^2 es la varianza de la distribución gaussiana centrada en el punto x_i.

Para el mismo par de puntos en un espacio dimensionalmente pequeño ($\mathbb{R}^2$ o $\mathbb{R}^3$), y_i e y_j, la probabilidad condicional $q_{j|i}$ se calcula de la misma manera. En este caso, la varianza de la distribución gaussiana se considera como $1/\sqrt{2}$. De esta forma, la probabilidad condicional resulta:

$$q_{j|i} = \frac{\exp\left(-\left\|y_i - y_j\right\|^2\right)}{\sum_{k \neq i} \exp\left(-\left\|y_i - y_k\right\|^2\right)} \tag{3.7}$$

Aquí también se cumple que $q_{j|i} = 0$.

Si la similitud y_i e y_j modela de forma apropiada la similitud en el espacio multidimensional x_i y x_j, resulta que $p_{j|i} = q_{j|i}$. Una forma de medir esta propiedad es la divergencia de Kullback–Leibler, que en este caso es igual a la entropía cruzada hasta alcanzar una constante aditiva. Por lo tanto, SNE minimiza la suma de las divergencias de Kullback–Leibler sobre todos los pares de puntos empleando un método de descenso de gradiente. La función de costo resulta:

$$C = \sum_i KL\left(P_i \,||\, Q_i\right) = \sum_i \sum_j p_{j|i} \log \frac{p_{j|i}}{q_{j|i}} \tag{3.8}$$

donde P_i es la distribución de la probabilidad condicional con todos los objetos dado el punto x_i, mientras que Q_i es la distribución de la probabilidad condicional con todos los objetos dado el punto y_i.

Como la divergencia de Kullback–Leibler no es simétrica, los diferentes tipos de error de las distancias de los pares de puntos en el espacio de baja dimensión no se ponderan por igual. De forma particular, la función de costo es alta cuando se usa una probabilidad pequeña $q_{j|i}$ para modelar probabilidades grandes $p_{j|i}$; es decir, considerar pares de puntos muy separados en el espacio de baja dimensionalidad para representar pares de puntos cercanos en el espacio de alta dimensionalidad. Así, la función de costo busca preservar la estructura local de los datos en el espacio $\mathbb{R}^2$ o $\mathbb{R}^3$.

Otro parámetro que se debe definir es la varianza de la distribución gaussiana centrada en el punto x_i (σ_i^2). Cualquier valor de σ_i genera una distribución de probabilidad (P_i) sobre todos los demás datos, la cual tiene una entropía directamente proporcional a σ_i. En esta técnica, se desarrolla una búsqueda binaria del valor σ_i que genera el valor de P_i con un valor fijo de perplejidad (definido por el usuario):

$$Perp\left(P_i\right) = 2^{H(P_i)} \tag{3.9}$$

donde $H(P_i)$ es la entropía de Shannon de P_i medida en bits:

$$H\left(P_i\right) = -\sum_j p_{j|i} \log_2 p_{j|i} \tag{3.10}$$

La perplejidad se puede considerar un parámetro de suavizado para el número efectivo de vecinos. Los valores que normalmente asume están entre 5 y 50.

3.2.4. Incrustación de vecinos estocásticos distribuidos en t

La incrustación de vecinos estocásticos distribuidos en t (t–SNE) [103] es un enfoque que permite superar inconvenientes encontrados durante la optimización de la función de costo y el problema de hacinamiento en la incrustación de vecinos estocásticos. Al igual que la SNE, esta técnica permite visualizar datos multidimensionales mediante su proyección en un mapa de dos o tres dimensiones, es decir, $\mathbb{R}^n \to \mathbb{R}^2$ o $\mathbb{R}^n \to \mathbb{R}^3$, respectivamente. Este método usa una versión simétrica de la función de costo de SNE con gradientes simples y emplea la distribución t–Student (de colas pesadas), lo que ayuda a mejorar el problema de hacinamiento y optimización.

Como alternativa para minimizar la suma de las divergencias de Kullback–Leibler entre las probabilidades condicionales $p_{j|i}$ y $q_{j|i}$, también es posible minimizar una divergencia única de Kullback–Leibler entre la distribución de probabilidad conjunta en el espacio de alta dimensionalidad (P) y la distribución de probabilidad conjunta en el espacio de baja dimensionalidad (Q):

$$C = KL\left(P \,||\, Q\right) = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}} \tag{3.11}$$

Aquí también se cumple que $p_{ii} = q_{ii} = 0$. Esta versión simétrica de la función de costo tiene la propiedad de que $p_{ij} = p_{ji}$ y $q_{ij} = q_{ji} \;\; \forall i, j$.

La similitud entre pares de puntos en el espacio de baja dimensionalidad está dado por:

$$q_{ij} = \frac{\exp\left(-\left\|y_i - y_j\right\|^2\right)}{\sum_{k \neq l} \exp\left(-\left\|y_k - y_l\right\|^2\right)} \tag{3.12}$$

Mientras que la similitud entre pares de puntos en el espacio de alta dimensionalidad es:

$$p_{ij} = \frac{\exp\left(-\left\|x_i - x_j\right\|^2 / 2\sigma^2\right)}{\sum_{k \neq l} \exp\left(-\left\|x_k - x_l\right\|^2 / 2\sigma^2\right)} \tag{3.13}$$

Sin embargo, existen problemas cuando el punto x_i en el espacio de alta dimensionalidad es un dato atípico «outlier», es decir, todos los pares de distancias $\left\|x_i - x_j\right\|^2$ son grandes. Para este punto atípico, los valores de p_{ij} son extremandamente pequeños (para todo j), entonces su ubicación en el espacio de baja dimensionalidad y_i tiene un ligero efecto sobre la función de costo. En consecuencia, la posición de este punto en el mapa no se encuentra bien determinada por las posiciones de los otros puntos en el mismo mapa. Este problema se supera al definir que las probabilidades conjuntas p_{ij} en el espacio de alta dimensionalidad sean iguales a las probabilidades condicionales simétricas:

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2n} \tag{3.14}$$

De esta forma se garantiza que $\sum_j p_{ij} > \frac{1}{2n}$ para todos los puntos x_i, con lo que cada punto x_i tiene una contribución significativa en el función de costo. De esta forma, la principal ventaja de la versión simétrica de la SNE es la forma simple del gradiente, lo cual genera mayor velocidad de cálculo.

En la t–SNE también se tiene que definir el valor de exageración (E). Este parámetro minimiza la divergencia de Kullback–Leibler entre el modelo de

distribución gaussiana de las distancias entre los puntos en el espacio original (alta dimensión) y el modelo de distribución t–Student de las distancias entre los puntos correspondientes en el espacio incrustado (baja dimensión). De esta forma, el parámetro E define el tamaño de los conglomerados naturales en los datos. Valores altos de exageración generan que el espacio entre los grupos originales sea mayor en el espacio incrustado. El valor más típico de E durante la fase de exageración temprana es 12; sin embargo, valores más altos también funcionan adecuadamente en combinación con diferentes tasas de aprendizaje [104].

3.2.5. Reducción de variables V–WSP

La reducción de variables basada en el método V–WSP [8,105] se fundamenta en el algoritmo propuesto por Wootton, Sergent y Phan–Tan–Luu «WSP» [106] para seleccionar un subconjunto representativo de variables que tengan mínima correlación en un espacio multidimensional definido. El algoritmo involucrado es el siguiente:

1. Elegir una variable semilla j y un valor umbral de correlación (thr).
2. Calcular el coeficiente de correlación de Pearson (R) entre la j–ésima variable y todas las demás variables.
3. Eliminar las variables cuyo valor absoluto $R \geq thr$.
4. Se fija la variable j y se selecciona entre las restantes variables aquella que tenga la correlación absoluta más alta con la misma.
5. Repetir los pasos 2, 3 y 4 hasta que $R < thr$, es decir, hasta que no existan variables correlacionadas por encima del umbral definido.

3.3. Aprendizaje supervisado

3.3.1. Árboles de clasificación

Los árboles de clasificación y regresión (CART) [94,107] realizan una secuencia de particiones binarias recurrentes del espacio multidimensional de los datos en diversos subespacios de clase. De esta manera, se forma un árbol de decisión que define ciertas reglas con las cuales se realiza la clasificación de un objeto. El árbol de decisión está formado por tres partes: la raíz o nodo superior, las ramas o nodos intermedios y las hojas o nodos terminales. En el nodo superior todos los objetos están agrupados previo al

inicio de la clasificación. Una vez que el algoritmo ha iniciado, los objetos se colocan temporalmente en los nodos intermedios para ser clasificados de acuerdo a la regla de decisión. Finalmente, los objetos se asignan en los nodos terminales al final de la secuencia de decisiones. A cada hoja se le asocia una clase.

En cada uno de los nodos intermedios se selecciona la variable que brinda la mayor separación de los datos. Para ello se recurren a las medidas de impureza del nodo, dentro de las cuales se incluye el índice de Gini:

$$I_G = 1 - \sum_{j=1}^{g} p_j^2 \tag{3.15}$$

y la entropía cruzada:

$$E = -\sum_{j=1}^{g} p_j \log p_j \tag{3.16}$$

donde p_j es la probabilidad de la j–ésima clase.

El índice de Gini mide la frecuencia con la que un elemento elegido al azar del conjunto de datos se etiquetaría incorrectamente. Al igual que el índice de Gini, la entropía cruzada mide el grado de desorden de clasificación generada por una variable en el nodo. La división óptima se elige por la clasificación con menos entropía. Estas dos funciones son diferenciables y, por lo tanto, son más susceptibles a la optimización numérica; es decir, son más sensibles a los cambios en las probabilidades en los nodos que la tasa de error en clasificación [95,96].

La partición se desarrolla de forma recursiva, es decir, los objetos que cumplen la regla de decisión se colocan en una hoja, caso contrario irán a otro grupo o a un nuevo nodo para realizar una nueva partición de acuerdo a la regla definida por la variable correspondiente. De esta forma, el modelo de clasificación (Figura 3.2) estará constituido por una serie de nodos que definen las reglas de clasificación.

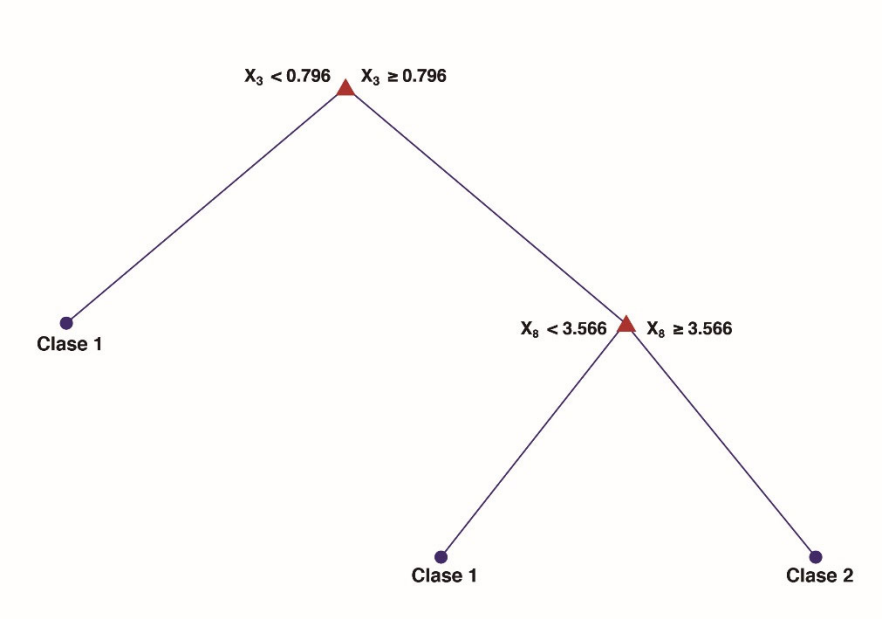

Figura 3.2 Esquema de una partición para generar un árbol de clasificación para dos clases

Debido a que este método usa únicamente las variables que permiten la mejor separación de las clases, se lo considera también como una técnica de selección de variables. Adicionalmente, tiene la ventaja de ser invariante al escalado de los datos y robusto a la presencia de valores atípicos [8].

Para evitar el problema de sobreajuste debido a un crecimiento demasiado profundo, se recurre al podado de los árboles [94,95]. La poda «pruning» consiste en reemplazar uno o más subárboles con hojas, cada uno etiquetado con la clase más frecuente de los objetos de calibración que alcanzan en el árbol CART original. La poda se desarrolla de forma secuencial: primero se reemplaza un subárbol con una hoja, luego otro, y así sucesivamente; siempre que los reemplazos sean favorables de acuerdo con algún criterio razonable, por ejemplo el criterio de la complejidad de costo [95]:

$$R_{\alpha}\left(T\right) = R\left(T\right) + \alpha\left|T\right| \tag{3.17}$$

donde R(T) es el error de calibración del árbol completo, |T| es el número de nodos terminales (hojas) y α es un parámetro que reduce la complejidad del árbol al controlar el número de hojas (lo que eventualmente reduce el sobreajuste). El parámetro alfa ($\alpha \geq 0$) se define mediante validación cruzada.

A medida que el parámetro α se incrementa, existe mayor poda del árbol, lo cual incrementa la impureza total de sus hojas. La poda basada en el criterio de la complejidad de costo genera una serie de árboles, donde la medida de la complejidad de costo para cada subárbol T_t es:

$$R_\alpha\left(T_t\right) = R\left(T_t\right) + \alpha\left|T_t\right| \tag{3.18}$$

Dentro de las desventajas que tienen los CART se puede citar la inestabilidad de los árboles asociada a la alta varianza; es decir, pequeños cambios en la base de datos (conjunto de calibración) puede resultar en una serie de divisiones muy diferente, lo que hace que la interpretación sea algo difícil. La razón principal de esta inestabilidad es la naturaleza jerárquica del proceso; es decir, el efecto de error en la división en un nodo superior se propaga a todos los nodos inferiores. Por otra parte, se tiene también la falta de suavidad de la superficie de predicción y la dificultad para capturar estructuras aditivas (especialmente para modelos de regresión). Estos inconvenientes se superan con el proceso de ensacado de los árboles (para usar el voto mayoritario) en los bosques aleatorios [95].

3.3.2. Bosques aleatorios

Los bosques aleatorios (RF) [95,108] se fundamentan en la idea del ensacado «bagging», es decir, en promediar varios modelos ruidosos, pero aproximadamente insesgados de tal manera de reducir la varianza. Particularmente, los árboles CART son notoriamente ruidosos, por lo que son los candidatos ideales para el ensacado, debido a que son capaces de capturar estructuras de interacción complejas en los datos y, si crecen abundantemente (suficientemente profundo), se benefician enormemente del promedio (sesgo relativamente bajo). Además, dado que cada árbol generado en el ensacado es independiente e idénticamente distribuido (i.d.d.), la esperanza de un promedio de B árboles es la misma que la esperanza de cualquiera de ellos. Esto significa que el sesgo de los árboles ensacados es el mismo que el de los árboles individuales (bootstrap), y la única alternativa para mejorar es a través de la reducción de la varianza. Esto está en contraste con el ensacado, donde los árboles crecen de manera adaptativa para eliminar el sesgo y, por lo tanto, no son i.d.d.

El promedio de las B variables aleatorias i.d.d. (cada una con varianza σ^2) tiene varianza $\frac{1}{B}\sigma^2$. Si las variables son simplemente idénticamente distribuidas (i.d.) pero no necesariamente independientes, con correlación positiva ρ entre ellas, la varianza del promedio es:

$$\rho\sigma^2 + \frac{1-\rho}{B}\sigma^2 \tag{3.19}$$

Se observa que a medida que B aumenta el segundo término desaparece, mientras que el primero permanece constante y, por lo tanto, el tamaño de la correlación entre pares de árboles ensacados limita los beneficios de realizar el promedio. La idea de los bosques aleatorios es mejorar la reducción de la varianza del ensacado al reducir la correlación entre los árboles, sin aumentar demasiado la varianza. Esto se logra en el proceso de crecimiento de los árboles mediante la selección aleatoria de las variables de entrada. El algoritmo involucrado en los bosques aleatorios es el siguiente:

1. Para $b = 1$ hasta B:
    a. Tomar una muestra bootstrap $\mathbf{Z}^*$ de tamaño N a partir de los datos de calibración.
    b. Hacer crecer un árbol del bosque aleatorio T_b para los datos bootstrap, repitiendo recursivamente los siguientes pasos para cada nodo terminal del árbol, hasta que se alcance el tamaño mínimo de nodo n_{min}.
        i. Seleccionar m variables aleatoriamente a partir de las p variables ($m \leq p$). Los valores de m normalmente usados son $\sqrt{p}$ (o incluso tan bajos como 1).
        ii. Elegir la mejor variable a partir de las m mejores.
        iii. Dividir el nodo en dos nodos secundarios.
2. Salida del conjunto de árboles $\left\{T_b\right\}_1^B$. Para realizar la predicción de un nuevo objeto x, se considera a $\widehat{C}_b\left(x\right)$ como la clase predicha

del *b*–ésimo árbol del bosque aleatorio. Entonces $\widehat{C}_{rf}^{B}(x) = voto\ mayoritario\left\{\widehat{C}_b(x)\right\}_1^B$.

Para el voto mayoritario se supone que se tienen diferentes reglas de clasificación, $h_1(x), h_2(x), \ldots, h_B(x)$, definidas por cada árbol aleatorio. De esta manera se pueden combinar estas reglas de tal forma que se produzca un modelo de clasificación (bosque aleatorio) que es superior con respecto a los modelos individuales. La forma en la que se realiza el voto mayoritario es la siguiente [109]:

$$\widehat{C}_{rf}^{B}(x) = moda\left\{h_1(x), h_2(x), \ldots, h_B(x)\right\} \tag{3.20}$$

De esta manera, cada valor de x se asigna a la clase que posee el mayor número de frecuencia (o votos). En los bosques aleatorios se puede entender el efecto del ensacado en términos de un análisis de consenso entre diversos árboles aleatorios e independientes [95]. Los modelos individuales contienen diferentes fuentes de ruido (especialmente cuando se trata de bases de datos grandes y heterogéneas), que se pueden reducir promediando las predicciones de los modelos. El supuesto principal del análisis de consenso es que las fortalezas de un modelo deben compensar las debilidades de los demás modelos y viceversa [5,8,110].

Muestras fuera de la bolsa

Para el desarrollo del modelo de los RF se crean dos conjuntos independientes. El primer conjunto, denominado de inicio o arranque, es elegido a partir de la base de datos inicial mediante muestreo con reemplazo y se denomina «dentro de la bolsa». Complementariamente, el conjunto «fuera de la bolsa» (OOB) son todos los datos que no se eligen durante el muestreo. Para cada observación $z_i = (x_i, y_i)$ del conjunto OOB, se construye su predictor del bosque aleatorio promediando solo aquellos árboles correspondientes a las muestras bootstrap del conjunto dentro de la bolsa en las que el objeto z_i no fue considerado.

La estimación del error de clasificación del conjunto OOB es similar a la obtenida mediante *k*–grupos de validación cruzada de dejar–varios–fuera (LMO). Por lo tanto, a diferencia de otros estimadores no lineales, los bosques aleatorios se validan de forma secuencial mediante procesos de

LMO. Una vez que el error del conjunto OOB se estabiliza, termina el proceso de calibración del modelo.

Importancia de las variables

En los bosques aleatorios se pueden construir gráficos para ver la importancia de las variables. En cada división de cada árbol, la mejora en el criterio de división es la medida de importancia atribuida a la variable de división, y se acumula sobre todos los árboles del bosque por separado por cada variable. La selección de una variable de división candidata aumenta la posibilidad de que cualquier variable se incluya en el bosque aleatorio.

Los RF también usan las muestras OOB para construir una medida diferente de la importancia de las variables, aparentemente para medir la fuerza de predicción de cada variable. Cuando se ha desarrollado el b–ésimo árbol, las muestras OOB pasan a través del mismo y se registra la exactitud (Acc) de la predicción. Entonces, los valores para la j–ésima variable se permutan aleatoriamente en las muestras OOB, y se calcula nuevamente la Acc. Como resultado de esta permutación, existe una disminución de la exactitud, cuyo promedio sobre todos los árboles se usa como una medida de la importancia de la variable j en el bosque aleatorio. La aleatorización anula el efecto de una variable, es decir, corresponde a una variable poco importante en el modelo. Sin embargo, esto no mide el efecto sobre la capacidad predictiva al excluir esta variable, pues si el modelo se recalibró sin ella, otras variables podrían usarse en su lugar.

3.3.3. k–vecinos más cercanos

El método de los k–vecinos más cercanos (kNN) [94,111] es un clasificador no lineal que no considera *a priori* las distribuciones estadísticas que siguen las variables (no paramétrico). Este método clasifica en función de analogía, es decir, según la clase a la que pertenecen la mayoría de los k objetos más cercanos en el espacio multidimensional. Para este propósito se calcula y analiza la matriz de distancias $\mathbf{D}$ (normalmente la distancia euclidiana) y se selecciona un número entero de entornos k (normalmente entre 1 y 10) al objeto a clasificar. Posteriormente, los objetos se ordenan de acuerdo a sus distancias y se clasifican en función de la clase a la cual pertenecen la mayoría de los k vecinos. El clasificador kNN es robusto y presenta buen desempeño cuando las superficies de separación entre las clases son no lineales o cuando una clase está contenida en otra.

El algoritmo computacional de los *k*NN inicia con el escalado de los datos, luego se selecciona la distancia a usar y se optimiza el número de vecinos *k*, posteriormente se calcula la matriz de distancias **D** para finalmente clasificar cualquier objeto de acuerdo a la clase más representativa de los entornos *k* más cercanos (voto mayoritario). Para definir el valor óptimo de *k*, se evalúan diversos valores y se elige aquel valor que genere la menor tasa de aciertos en validación cruzada (NER_{cv}). De esta manera, el modelo *k*NN está constituido por los objetos del conjunto de calibración, el valor óptimo de *k* y la matriz **D** (no existe una función matemática). En consecuencia, un objeto nuevo a ser predicho se introduce en la matriz de datos y se corre el algoritmo para evaluar la clase mayoritaria de los *k* vecinos del conjunto de calibración.

3.3.4. *N*–vecinos más cercanos

El clasificador de los *N*–vecinos más cercanos (N3) [8,112] utiliza todos los n – 1 vecinos para clasificar un objeto, ordenados desde el más similar hasta el menos similar para obtener el vector de similitud r, el cual mide la contribución de los vecinos a la asignación de las clases y se encuentra modulado por un parámetro α. Así, para cada *i*–ésimo objeto a ser clasificado, la contribución de la *g*–ésima clase se calcula de acuerdo con:

$$w_{ig} = \frac{1}{\hat{n}_g} \times \sum_{\substack{j=1 \\ j \neq i}}^{n-1} \frac{s_{ij}}{r_{ij}^{\alpha}} \times \delta_i \tag{3.21}$$

donde s_{ij} es la similitud entre la *i*–ésima y la *j*–ésima observación; r_{ij} es el valor de ranking de la similitud del *j*–ésimo objeto con respecto al *i*–ésimo objeto; α es un parámetro de valor real a ser optimizado dentro del intervalo [0.1, 2.5]; δ_j es la delta de Dirac que es igual a 1 cuando el *j*–ésimo objeto pertenece a la *g*–ésima clase y su contribución a la ponderación de la clase es mayor que ε:

$$\delta_j = \begin{cases} 1 & si \quad c_j = g \wedge \dfrac{s_{ij}}{r_{ij}^{\alpha}} > \varepsilon \\ 0 & caso\ contrario \end{cases} \tag{3.22}$$

donde c_j es la clase del j–ésimo objeto. Finalmente, $\hat{n}_g$ es el número de vecinos que contribuyen al peso de la clase de la siguiente manera:

$$\hat{n}_g = \sum_j \delta_j \tag{3.23}$$

El exponente α de la Ec. 3.21 se optimiza de tal forma que genere la mayor tasa de aciertos en validación cruzada (NER_{cv}).

3.3.5. Vecinos más cercanos agrupados

El clasificador de los vecinos más cercanos agrupados (BNN) [8,112] predice la clase de un objeto mediante el uso de valores distintos de los *k–vecinos* de acuerdo al criterio del voto mayoritario; es decir, para la clasificación de un objeto se consideran todos aquellos vecinos que tienen la similitud más grande y comparable con el mismo. Para seleccionar los vecinos más similares se definen intervalos de similitud denominados «bins», dentro de los cuales se distribuyen los vecinos de acuerdo a la similitud que presentan con respecto al objeto a clasificar. De esta forma, para la predicción de la clase de un objeto se consideran todos los vecinos que caen dentro del intervalo con la similitud más grande. Los intervalos de similitud se definen mediante la optimización del parámetro α, el cual define el ancho del intervalo:

$$bin_m^{\alpha} = \left[S_m^{\alpha}, S_{m+1}^{\alpha} \right] \qquad m = 1,2,3,... \tag{3.24}$$

donde S son los distintos valores de similitud (vectores) definidos de forma decreciente con variación fija de 0.1 dentro del rango [0.1, 1] (es decir 1, 0.9, 0.8, 0.7, 0.6, 0.5, 0.4, 0.3, 0.2 y 0.1); mientras que para el rango [0, 0.1] se inicia con 10^{-2} y se disminuye en una unidad en la potencia para cada entorno (es decir 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}, 10^{-7} y 10^{-8}). Se incluyen estos últimos siete intervalos para evitar que el intervalo $S = 0.1$ sea demasiado grande, cuando se trata de valores pequeños de α, y así evitar el riesgo de que todos los n – 1 objetos se agrupen en el mismo.

El valor óptimo de α usado para definir el umbral del intervalo se optimiza dentro del rango [0.1, 1.5] con variación de 0.05, de tal forma que provea la mayor tasa de aciertos en validación cruzada (NER_{cv}). Una vez que se han

definidos los intervalos, el algoritmo de clasificación de los BNN es el siguiente:

1. Se calcula para cada elemento la similitud con respecto al objeto a clasificar.
2. Los elementos se distribuyen dentro del intervalo de similitud de acuerdo a su similitud con respecto al objeto a clasificar.
3. Se seleccionan únicamente los elementos que se encuentran dentro del primer intervalo no vacío como los vecinos más cercanos para la clasificación.
4. La predicción se la considera mediante el voto mayoritario. Cuando varias clases comparten el mismo número de vecinos, el objeto a clasificar se asigna a la clase que tiene la suma máxima de las contribuciones de similitud.

3.3.6. Selección de variables

Debido a que las moléculas que forman un conjunto de datos para un estudio QSAR están representadas por miles de descriptores moleculares, es importante la búsqueda de un subconjunto óptimo de descriptores de tal forma de desarrollar modelos de clasificación que posean buena calidad de ajuste (calidad descriptiva) de los datos y buena capacidad predictiva. Desarrollar modelos con demasiadas variables aumenta su capacidad de ajuste, pero la capacidad predictiva tiende a disminuir después de cierto punto (óptimo). En consecuencia, se debe apuntar a un modelo óptimo y parsimonioso, que balancee la calidad de ajuste y la capacidad predictiva. La búsqueda de un subconjunto óptimo de variables permite expresar de forma simple la relación entre las mismas y la respuesta, separar las variables relevantes de aquellas que no lo son, mejorar la precisión de los estimadores estadísticos y la predictividad del modelo [8,113]. Para los estudios QSAR se han propuestos diferentes métodos de selección basados en diferentes teorías, por ejemplo, la búsqueda exacta, métodos paso a paso, algoritmos genéticos, método de reemplazo, método de reemplazo modificado y ampliado, reemplazo secuencial reconfigurado, optimización por enjambre de partículas y optimización de colonia de hormigas, entre otros [8].

Otro enfoque son los algoritmos genéticos (GAs) [94,114], que es un método de optimización particularmente idóneo para tratar problemas de

relacionados con el desarrollo de los mejores modelos de aprendizaje automático (regresión o clasificación); es decir, se utilizan para la selección de aquellas variables que generen un máximo en una función objetivo (por ejemplo la *NER*). Los GAs se basan en la teoría de la evolución de Darwin, para lo cual se inicia con una población de «cromosomas» generada de manera aleatoria. El cromosoma es un vector binario de p bits (variables), es decir, un número igual a 0 o 1 al que se denomina «gen». De esta manera, cada cromosoma es una representación de un punto en el espacio p–dimensional de las variables para seleccionar (genes), que se asocia a un modelo particular, donde bits con valor 1 indican a las variables presentes en el modelo y bits igual a 1 indican que esas variables no se encuentran en el modelo. A cada cromosoma se asocia una cierta respuesta que se debe optimizar, por ejemplo, la *NER* en los modelos de clasificación. Los GAs se desarrollan en tres etapas (Figura 3.3).

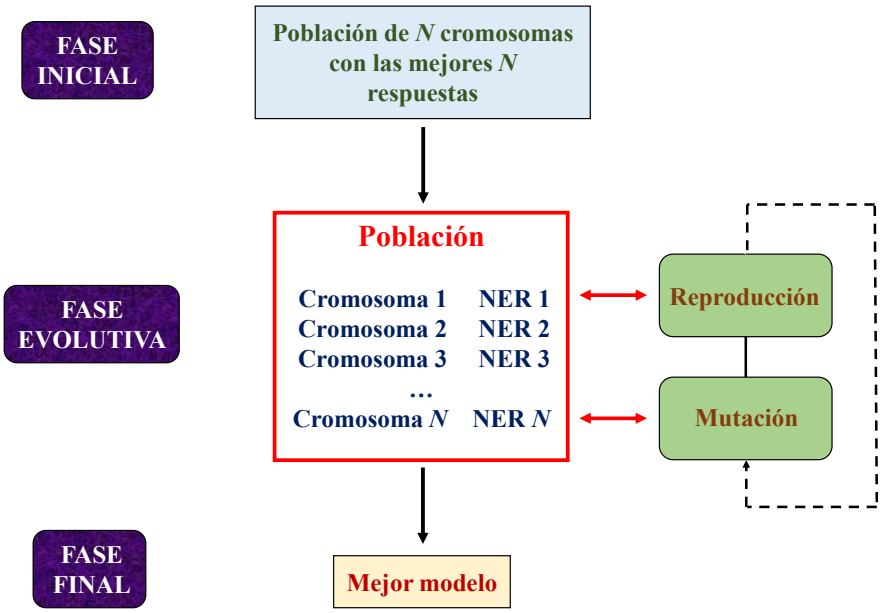

Figura 3.3 Diagrama de funcionamiento de los GAs para desarrollar modelos de clasificación

Fase inicial

Se define la dimensión N de la población de partida, es decir, el número de cromosomas a partir de los cuales se desarrollará la evolución (por ejemplo 100). Seguidamente, se construye de forma casual un cierto número de cromosomas, normalmente mayor a la dimensión de la población inicial (por

ejemplo 300) y se evalúa la función objetivo (*NER*) de cada uno. Finalmente, se ordenan los cromosomas de forma decreciente en la respuesta y se incluyen los mejores en la población inicial (los peores se eliminan). Normalmente el tamaño de la población se mantiene constante.

Fase evolutiva

La fase evolutiva de los GAs para generar nuevos cromosomas se desarrolla de dos maneras:

1. Reproducción «crossover»: en este proceso se seleccionan dos cromosomas padres de la población *N*, a partir de los cuales se generan cromosomas hijos. La selección de los padres puede ser casual o proporcional a la calidad de los cromosomas (sesgada hacia los mejores). Los cromosomas hijos comparten el patrimonio genético de los padres; es decir, los bits igual a 0 y 1 en ambos padres se mantienen, mientras que los valores de los bits dispares se fijan de acuerdo a una regla de probabilidad. Para cada cromosoma hijo se evalúa la *NER* y si es mejor a alguno de los cromosomas de la población de partida, este cromosoma ingresa en la población en el lugar correspondiente de orden, caso contrario se descarta.
2. Mutación: una vez que se ha desarrollado un cierto número de reproducciones, se analiza cada cromosoma en base a una probabilidad de mutación, de tal forma que en los bits se cambia el código binario; es decir, 0 a 1 y viceversa. La mutación limita la posibilidad de que la población se quede atrapada en un mínimo local. Para evitar que la población se aleje demasiado de la probable región óptima, la probabilidad de mutación es menor a la probabilidad de reproducción.

Fase final

El algoritmo computacional finaliza en función de criterios preestablecidos; por ejemplo, cuando los GAs han superado un número

máximo de iteraciones, cuando la población no mejora o no se renueva luego de un cierto número de iteraciones.

Para evitar el riesgo de sobreajuste en los modelos, se ha implementado una variante en la que, en lugar de realizar una simple corrida y muchas iteraciones, se realiza un número pequeño de corridas independientes a partir de varias poblaciones iniciales. A continuación, se registra la frecuencia de selección de las variables en cada una de las corridas, para desarrollar el modelo mediante la inclusión de las variables más frecuentes [114].

3.3.7. Medidas de evaluación en clasificación

Para la evaluación de la calidad de un modelo de clasificación, se parte de la matriz de confusión [115,116], la cual es una tabla de contingencia cuadrada que se obtiene al comparar las clases verdaderas y las clases predichas por el modelo. Cuando se trata de modelos que involucran G clases (multiclase), resulta una matriz de dimensión G × G (Tabla 3.1).

Tabla 3.1 Matriz de confusión para el cálculo de las medidas de evaluación de los modelos de clasificación con G clases

		Clases asignadas				
		1	2	...	G	
Clases verdaderas	1	c_{11}	c_{12}	...	c_{1G}	n_1
	2	c_{21}	c_{22}	...	c_{2G}	n_2
	...	...	...	...	...	...
	G	c_{G1}	c_{G2}	...	c_{GG}	n_G
		n_1'	n_2'	...	n_G'	n

donde cada elemento de la matriz c_{gk} representa el número de elementos que pertenecen a la clase g y que son clasificados en la clase k, los elementos diagonales c_{gg} representan el número de objetos clasificados correctamente, los elementos fuera de la diagonal son los objetos incorrectamente clasificados (número de errores de clasificación), n es el número total de elementos, G es el número total de clases, n_g es el número de objetos que pertenecen a la g–ésima clase y n_g' es el número de elementos asignados a la g–ésima clase. La matriz de confusión es normalmente asimétrica, debido a que el número de objetos que pertenecen a la g–ésima clase y son clasificadas en la k–ésima clase (c_{gk}) no es igual al número de objetos que pertenecen a la k–ésima clase y son asignados a la g–ésima clase (c_{kg}).

Índices primarios

A partir de la matriz de confusión es posible calcular diversos índices primarios, también conocidos como medidas de clase [115,116].

Precisión: se define como la pureza de la clase y mide la capacidad del modelo de no incluir objetos de otras clases en la g–ésima clase:

$$\mathrm{Pr}_g = \frac{c_{gg}}{n'_g} \tag{3.25}$$

Sensibilidad: mide la capacidad del modelo de reconocer correctamente elementos que pertenecen a la g–ésima clase:

$$Sn_g = \frac{c_{gg}}{n_g} \tag{3.26}$$

Especificidad: mide la capacidad de la g–ésima clase del modelo para rechazar objetos de todas las demás clases:

$$Sp_g = \frac{\sum_{\substack{k=1 \\ k \neq g}}^{G} \left(n_k - c_{kg} \right)}{n - n_g} \tag{3.27}$$

Medida F1: se calcula como la media armónica de la Sn y la Pr de una clase:

$$F1_g = \frac{2}{\frac{1}{Sn_g} + \frac{1}{Pr_g}} = 2\frac{Sn_g \cdot Pr_g}{Sn_g + Pr_g} \tag{3.28}$$

Índices globales

Las medidas primarias brindan únicamente una idea de cómo trabaja el método de clasificación sobre una clase específica, pero no proporcionan una

evaluación general de la calidad global de clasificación. Por lo tanto, las medidas globales de clasificación son [115]:

Exactitud: se calcula como la suma de los elementos diagonales de la matriz de confusión (proporción de elementos correctamente clasificados) sobre el número total de elementos. Este indicador no considera ninguna información concerniente a la calidad de clasificación de las clases individuales:

$$Acc = \frac{\sum_{g=1}^{G} c_{gg}}{n} \tag{3.29}$$

Tasa de aciertos o exactitud balanceada: se calcula como el promedio de las sensibilidades de las clases. La tasa de aciertos (NER) estima mejor la calidad de los modelos de clasificación con respecto a la exactitud, particularmente cuando las clases son desbalanceadas. Por esta razón, también se la conoce como exactitud balanceada (balanced accuracy).

$$NER \equiv Sn = \frac{\sum_{g=1}^{G} Sn_g}{G} \tag{3.30}$$

Precisión promedio: se calcula como el promedio de las precisiones de las clases.

$$Pr = \frac{\sum_{g=1}^{G} Pr_g}{G} \tag{3.31}$$

Coeficiente de correlación de Matthew: es una medida bien conocida para medir el rendimiento de modelos de clasificación binaria:

$$MCC = \frac{c_{11} \cdot c_{22} - c_{21} \cdot c_{12}}{\sqrt{n_1 \cdot n_1^{'} \cdot n_2 \cdot n_2^{'}}} \tag{3.32}$$

El coeficiente de correlación de Matthew varía entre –1 y +1. Para tratar sistemas multiclase, se ha propuesto el coeficiente de correlación de Matthew extendido (EMCC).

Medida F1: se calcula como la media armónica de la sensibilidad y precisión:

$$F1 = 2\frac{Sn \cdot Pr}{Sn + Pr} = 2\frac{NER \cdot Pr}{NER + Pr} \tag{3.33}$$

Finalmente, no ha sido propuesta en la literatura una medida global de clasificación calculada como el promedio de las especificidades de las clases. Posiblemente, esto se debe al sesgo que esta medida tiene con relación al número total de clases G.

Índices para clasificación binaria

Cuando el número de clases G es igual a 2 (clasificación binaria), la matriz de confusión tiene la siguiente forma (Tabla 3.2).

Tabla 3.2 Matriz de confusión para el cálculo de las medidas de evaluación de los modelos de clasificación con dos clases

		Clases asignadas	
		1	2
Clases verdaderas	1	TP	FN
	2	FP	TN

donde TP es el verdadero positivo (predice la clase positiva como positiva), FP es el falso positivo (predice la clase negativa como positiva), FN es el falso negativo (predice la clase positiva como negativa) y TN es el verdadero negativo (predice la clase negativa como negativa).

De esta forma, las medidas de clasificación toman la siguiente forma:

$$\text{Pr} = \frac{TP}{TP + FP} \qquad Sn = \frac{TP}{TP + FN}$$

$$Sp = \frac{TN}{TN + FP} \qquad Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.34}$$

$$NER = \frac{Sn + Sp}{2} \qquad F1 = \frac{2TP}{2TP + FP + FN}$$

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{\left(TP + FN\right) \cdot \left(TP + FP\right) \cdot \left(TN + FP\right) \cdot (TN + FN)}}$$

En modelos de clasificación que involucren únicamente dos clases, la sensibilidad de la clase 1 corresponde a la especificidad de la clase 2 y viceversa. Por esta razón la tasa de aciertos se calcula como el promedio entre la sensibilidad y especificidad.

3.4. Técnicas de validación

La validación [116,117] es un aspecto importante para garantizar que el modelo no esté sobreajustado y para maximizar la capacidad predictiva al realizar pequeñas perturbaciones. De hecho, al aumentar la complejidad del modelo, aumenta la capacidad de ajuste (calidad descriptiva) del mismo; sin embargo, un incremento descontrolado de la complejidad provoca una caída en la capacidad predictiva. Los modelos con alta complejidad intentan capturar todas y cada una de las variaciones en los datos, por lo que son modelos de alta varianza. Estos modelos tienden a tener un desempeño pobre en predicción, debido a que, al capturar cada variación en los puntos de calibración, también lo hace con los valores atípicos, datos aleatorios (correlación casual) y puntos con alto valor de influencia (apalancamiento). Esto produce que el modelo se sobreajuste a los datos de calibración. Por lo tanto, para mejorar el rendimiento de un modelo en predicción, se reduce la complejidad y varianza en el conjunto de calibración. Al reducir la varianza del modelo, se introduce un error en el mismo (denominado sesgo).

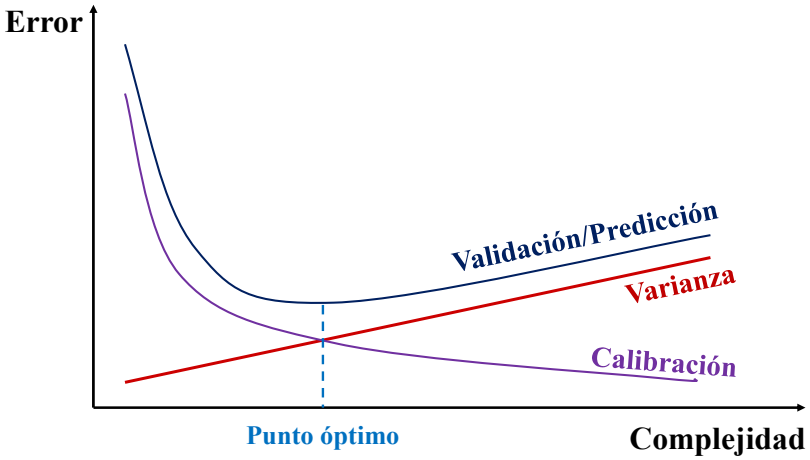

Figura 3.4 Representación esquemática de la complejidad del modelo versus el error para el conjunto de calibración, validación y/o predicción

Por lo tanto, al reducir la varianza del modelo aumenta el sesgo y viceversa (compensación sesgo–varianza). La compensación sesgo–varianza se esquematiza en la Figura 3.4. Se observa que a medida que aumenta la complejidad del modelo, la varianza del mismo aumenta y el error de calibración disminuye. El error de validación también comienza a disminuir a medida que aumenta la complejidad, pero hasta cierto punto, después del cual comienza a aumentar. Este punto indica la complejidad óptima del modelo, donde existe contemporáneamente buena capacidad descriptiva de los datos y estabilidad independientemente de los mismos (predictividad).

3.4.1. Validación interna

La validación interna es un procedimiento de remuestreo que se utiliza para evaluar modelos de aprendizaje automático, en el que se considera un solo parámetro llamado k que se refiere al número de grupos en los que se dividirá una conjunto de datos determinado. Por lo tanto, el método se denomina k–grupos de validación cruzada «k–Fold Cross Validation» [95,116] y permite realizar la partición del conjunto de datos de calibración siguiendo una lógica. Esta metodología consiste en dividir el conjunto de calibración en k grupos de validación, los que se excluyen una sola vez del modelo, se recalibra y se realiza la predicción de la respuesta del grupo excluido. Esta metodología también se conoce como dejar–varios–fuera (LMO). La Tabla 3.3 muestra los valores de k generalmente usados y el porcentaje de objetos que se incluyen en el subconjunto de validación.

Tabla 3.3 Valores de k y porcentaje de objetos que son colocados en el subgrupo de validación en la técnica de k–grupos de validación cruzada

k	% de objetos
2	50
3	33.3
4	25
5	20
10	10

Si $k = n$, se trata de la modalidad de dejar–uno–fuera (LOO). En LOO se calculan n modelos, en cada uno de los cuales se excluye un objeto a la vez. De esta manera, los modelos construidos con los $n - 1$ objetos se usan para predecir la respuesta del objeto excluido [118]. Por otra parte, si $k < n$, los grupos de validación se generan de acuerdo a dos procedimientos: 1) bloques continuos «contiguous blocks» y 2) ventanas venecianas «venetian blinds» [119]. Cada grupo contiene n/k objetos. Cada grupo k de validación se excluye una sola vez, se recalibra el modelo y luego se realiza la predicción de los objetos excluidos.

En bloques continuos, cada subconjunto k de validación se selecciona a partir de los primeros n/k elementos del conjunto de calibración ordenados de forma secuencial (bloques continuos), es decir, se excluye el primer bloque, luego el segundo bloque y así sucesivamente hasta el k–ésimo bloque. Por otra parte, en ventanas venecianas cada objeto del conjunto de validación es seleccionado a partir del primer objeto del conjunto de calibración y los subsiguientes cada k–ésimo objeto, es decir, el primer grupo de validación contiene el primer elemento de cada bloque continuo, el segundo grupo contiene el segundo elemento de cada bloque continuo y así sucesivamente hasta el último elemento de cada bloque continuo. En modelos supervisados de clasificación, la elección entre estos dos tipos de validación dependerá de la forma en la que se distribuyen las clases en el vector respuesta; es decir, bloques continuos se usa cuando las clases se encuentran distribuidas aleatoriamente; mientras que ventanas venecianas es útil cuando las clases siguen un orden lógico.

Una técnica similar a k–grupos de validación cruzada es el método Monte Carlo, en el que también se definen en varias oportunidades (iteraciones) distintos subconjuntos de forma aleatoria (validación LMO). En cada iteración los objetos del grupo de calibración se dividen en un subconjunto de entrenamiento (por ejemplo el 80%) y un subconjunto de validación

(20%); se calibra el modelo con las moléculas del subconjunto de calibración y luego se utiliza para predecir las respuestas de las moléculas de validación. La calidad de la validación Monte Carlo se obtiene por comparación de las predicciones acumuladas versus las clases del subconjunto de evaluación [120]. Debido a que la partición se realiza de forma independiente para cada iteración, los objetos aparecerán varias veces en los subconjuntos de validación. La validación Monte Carlo se repite tantas veces como sea posible, por ejemplo, 10000 iteraciones.

Otra técnica de validación interna es el método Bootstrap [121], en el que se utiliza el muestreo aleatorio con reemplazo. Es decir, este método conserva constante el tamaño original del conjunto de calibración n, por lo que el subconjunto de entrenamiento estará conformado por objetos repetidos (incluso más de una vez) a fin de que se equiparen con el número de elementos que fueron excluidos para formar el conjunto de validación. Al igual que las demás técnicas de validación, el modelo se calcula con los elementos del conjunto de calibración, para posteriormente realizar las predicciones de las respuestas de los elementos del conjunto de validación. El procedimiento de extracción de los subconjuntos n–dimensionales se repite tantas veces como sea posible.

3.4.2. Validación externa

La validación externa del modelo es un aspecto importante en los modelos QSAR. Para esto, la base de datos se divide en dos grupos: 1) conjunto de calibración (training set) y 2) conjunto de predicción (test set). El primero se utiliza para construir el modelo y el segundo se usa para predecir la respuesta de sus elementos con el modelo desarrollado. El balance entre las clases tiene influencia en la calidad del modelo QSAR, debido a que los modelos se sesgan hacia la clase más numerosa. Por este motivo, es común que el conjunto de calibración sea formado tomando en consideración la numerosidad de las clases [122]. Es decir, la partición se realiza de forma casual y proporcional a la numerosidad de las mismas, de tal forma de obtener similar representatividad en los dos conjuntos [8]. La característica fundamental del grupo de predicción es que no se utiliza durante la calibración y validación interna del modelo. Por lo general, en el conjunto de predicción se coloca entre el 10% y 50% de los objetos. Este procedimiento se puede realizar una sola vez «single evaluation set» o varias veces «repeated evaluation set». La validación externa se utiliza principalmente

durante el aprendizaje supervisado para estimar la capacidad de un modelo de predecir nuevos datos.

Capítulo 4

APLICACIONES

Este capítulo inicia con la descripción de la forma en que se ha compilado la base de datos de moléculas de gusto, así como la manera en que se ha filtrado y curado la información. Posteriormente, se utiliza el aprendizaje no supervisado para definir el espacio químico del gusto (mapa bidimensional), seguido del uso de diversas estrategias del aprendizaje supervisado (clasificación) para calibrar modelos que permitan predecir el gusto de los compuestos químicos estudiados en función de diversas formas de representación molecular.

4.1. Aprendizaje no supervisado para definir el espacio químico del gusto

4.1.1. Materiales y métodos

Para la generación de la base de datos se ha realizado una búsqueda en múltiples fuentes bibliográficas de la información de diversos compuestos químicos para los cuales se haya medido de forma experimental el gusto. De esta manera, se han obtenido 4580 moléculas a partir de 37 artículos científicos, 3 libros y 53 capítulos de libros (a partir de 8 libros). Cada molécula está asociada a un gusto básico (dulce, amargo, umami, ácido o salado), mezcla de estos gustos básicos (multigusto) y otras sensaciones gustativas tales como insípido (neutro), no dulce, no amargo, astringente, refrescante, picante, quemante o pungente.

Durante el desarrollo de la base de datos, se han adoptado los siguientes criterios:

1. No se han incluido proteínas (macromoléculas), por ejemplo, miraculina, brazzeína, curculina, pentadina, monelina (I y II), taumatina (I, II, III, a, b y c), mabinlina (I y II).

2. Se han eliminado las moléculas de agua de los compuestos hidratados. Esto debido a que la evaluación sensorial del gusto se realiza mediante la metodología de "beber y escupir" «sip and spit» [6,67,68].
3. En la clase de compuestos umami se han incluido moléculas que presentan este gusto, así como moléculas moduladoras y mejoradoras del sabor umami [2,123].
4. Se ha utilizado la proyección de Haworth para la representación estructural de los monosacáridos, por ejemplo, fructosa, glucosa, psicosa o tagatosa.

Seguidamente, cada compuesto químico fue diseñado y optimizado en el programa HyperChem [38]. Se utilizaron los campos de la mecánica molecular (MM+) y el algoritmo de gradiente conjugado para la optimización de las geometrías. El criterio de convergencia para la optimización es que el elemento máximo del vector gradiente de la energía total con respecto a las coordenadas atómicas sea menor a 0.01 $kcal \times (Å \times mol)^{-1}$. Cuando se disponía de la información de los estereocentros, se utilizó para diferenciar estereoisómeros. Si esta información no estaba disponible en la fuente, se la obtuvo de otras referencias bibliográficas o quimiotecas de acceso libre (por ejemplo PubChem); caso contrario se utilizó la estructura molecular generada por defecto por el modelador de HyperChem (no se ha realizado un análisis conformacional).

Para garantizar la confiabilidad de la base de datos generada, se ha procedido a realizar un curado de los compuestos para identificar la presencia de potenciales errores en las estructuras químicas, por ejemplo, falta de átomos o grupos funcionales, incorrecta orientación de los sustituyentes en los estereocentros, desplazamiento de átomos (ubicación incorrecta) o intercambio de grupos químicos. Estos errores influyen en el cálculo de los descriptores moleculares y en las subsiguientes aplicaciones de las máquinas de aprendizaje [124]. Para este propósito, se utilizó el programa alvaMolecule [125] para identificar moléculas con estructuras múltiples, valencia inusual, con carga total, estructuras con átomos cargados, átomos no estándar (H, C, N, O, P, S, F, Cl, Br e I) y moléculas sin estandarización de los anillos aromáticos. Además, las moléculas se verificaron en la quimioteca PubChem [16] a través de una opción implementada en alvaMolecule y se obtuvo el número de registro CAS y el código PubChem CID (cuando esta información

estaba disponible). También se generó en el mismo programa la notación lineal de cadena SMILES canónico, con la cual se realizó el filtrado de la base de datos en un diagrama de flujo programa en KNIME [126] (Figura 4.1).

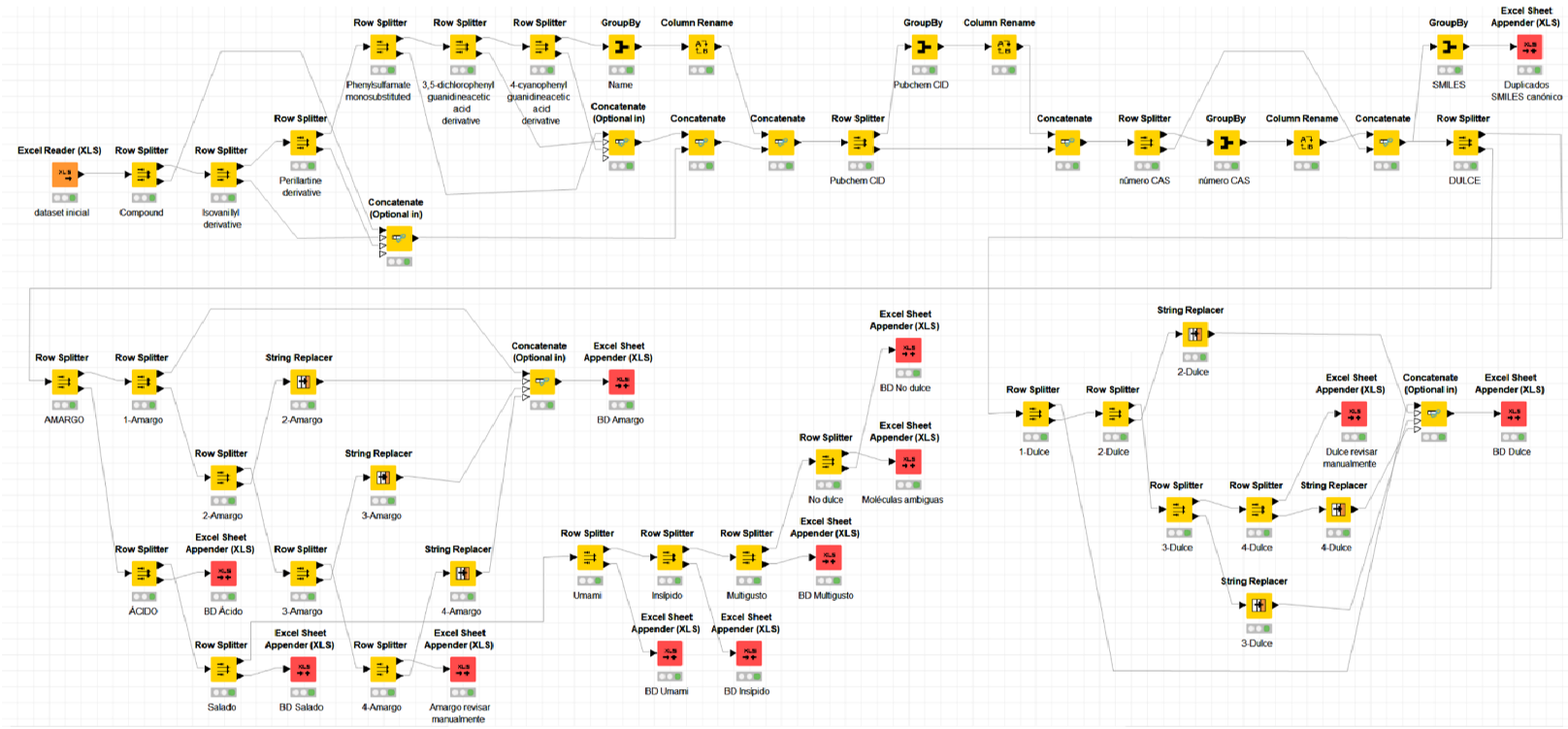

Figura 4.1 Diagrama de flujo KNIME para el filtrado y curado de la base de datos de gustos

Seguidamente, con las moléculas filtradas, se calcularon 166 claves moleculares del sistema de acceso molecular (MACCS) [48,64] en el programa alvaDesc [52]. Con estas claves moleculares binarias se definió el espacio químico mediante la incrustación de vecinos estocásticos distribuidos en t (t–SNE) [103] para proyectar las similitudes/disimilitudes en el espacio de las 166 claves MACCS en un mapa bidimensional (es decir $\mathbb{R}^{166} \rightarrow \mathbb{R}^2$), utilizando el coeficiente de similitud de Jaccard–Tanimoto [127]. Este coeficiente enfatiza la presencia de características comunes (etiquetadas como *a*) omitiendo la ausencia de características comunes (etiquetadas como *d*). De esta manera, el gráfico de dispersión bidimensional define el espacio químico o mapa químico de los compuestos de gusto. El algoritmo t–SNE se ha implementado en el lenguaje de programación MATLAB [128].

4.1.2. Resultados y discusión

La información de las 4580 moléculas recopiladas de las distintas fuentes bibliográficas, así como el número de registro CAS, el PubChem CID y la notación lineal de cadena SMILES canónico, se importaron en el programa KNIME para realizar el filtrado de la base de acuerdo al diagrama de flujo de la Figura 4.1. Durante el filtrado de la base de datos se consideraron los siguientes criterios:

1. Fusionar los compuestos duplicados por el nombre químico.
2. Se ha usado el número de registro CAS y el PubChem CID para identificar los compuestos duplicados que se encuentran etiquetados como “compuesto”, “fenilsulfamato monosustituido”, “derivado de la isovanillina”, “derivado de la perillartina”, “derivado del ácido guanidinoacético 3,5–diclorofenil” y “derivado del ácido guanidinoacético 4–cianofenil”.
3. La notación lineal de cadena SMILES se ha aplicado como último filtro para identificar los compuestos duplicados que no fueron identificados en los pasos 1 y 2. Si los mismos correspondían a estereoisómeros, no se fusionaron (por ejemplo D–Glucosa y L–Glucosa).

4. Se han aplicado filtros para separar los compuestos que pertenecen a las clases dulce, amargo, umami, ácido, salado, insípido y no dulce.
5. Aquellos compuestos que presenten dos o más gustos distintos se incluyeron en la clase "multigusto", siempre que todas las clases sean distintas o cuando exista un empate en la numerosidad de las mismas.
6. Para las moléculas que presenten tres o más gustos no empatados, se usó el voto mayoritario para asignarlas a la clase del gusto predominante.
7. Finalmente, en la clase "misceláneos" se han colocado a los compuestos etiquetados como astringente, refrescante, picante, quemante o pungente; así como los compuestos ambiguos: amargo/quemante, amargo/insípido, no amargo/quemante, no dulce/dulce, dulce/amargo, dulce/insípido.

De esta manera, se obtuvo una base de datos curada con 2944 estructuras moleculares (ChemTastesDB), donde se han definido 9 clases a las que puede pertenecer cualquier compuesto:

- 977 dulcificantes.
- 1183 compuestos amargos.
- 98 moléculas umami.
- 38 estructuras moleculares ácidas.
- 12 compuestos salados.
- 233 moléculas no dulces.
- 203 estructuras insípidas.
- 113 compuestos que presentan multigusto.
- 87 moléculas en la clase misceláneos.

Seguidamente, se utilizaron las 166 claves moleculares del sistema de acceso molecular (MACCS) para definir el espacio químico mediante la incrustación de vecinos estocásticos distribuidos en t (t–SNE), disponible en el "Statistics and Machine Learning Toolbox" de MATLAB. Se ha utilizado el coeficiente de similitud de Jaccard–Tanimoto y el algoritmo de control exacto (optimiza la divergencia de las distribuciones de Kullback–Leibler

entre el espacio original y el espacio incrustado). Para automatizar la búsqueda de los diversos parámetros que se deben considerar en la t–SNE, se ha programado una función específica. De esta manera, se ha explorado valores de exageración E = [2, 4, 500, 100], perplejidad P = [20, 30, 40, 50] y tasa de aprendizaje L = [100, 500, 900, 1300]. Así, al realizar la combinación de estos tres parámetros, se ha obtenido un total de 64 mapas químicos. La selección del mejor espacio químico se ha realizado en términos de la mayor formación de grupos de moléculas y la separación de las clases analizadas. En la Figura 4.2 se presenta el espacio químico del gusto, diseñado considerando los parámetros E = 100, P = 30 y L = 100.

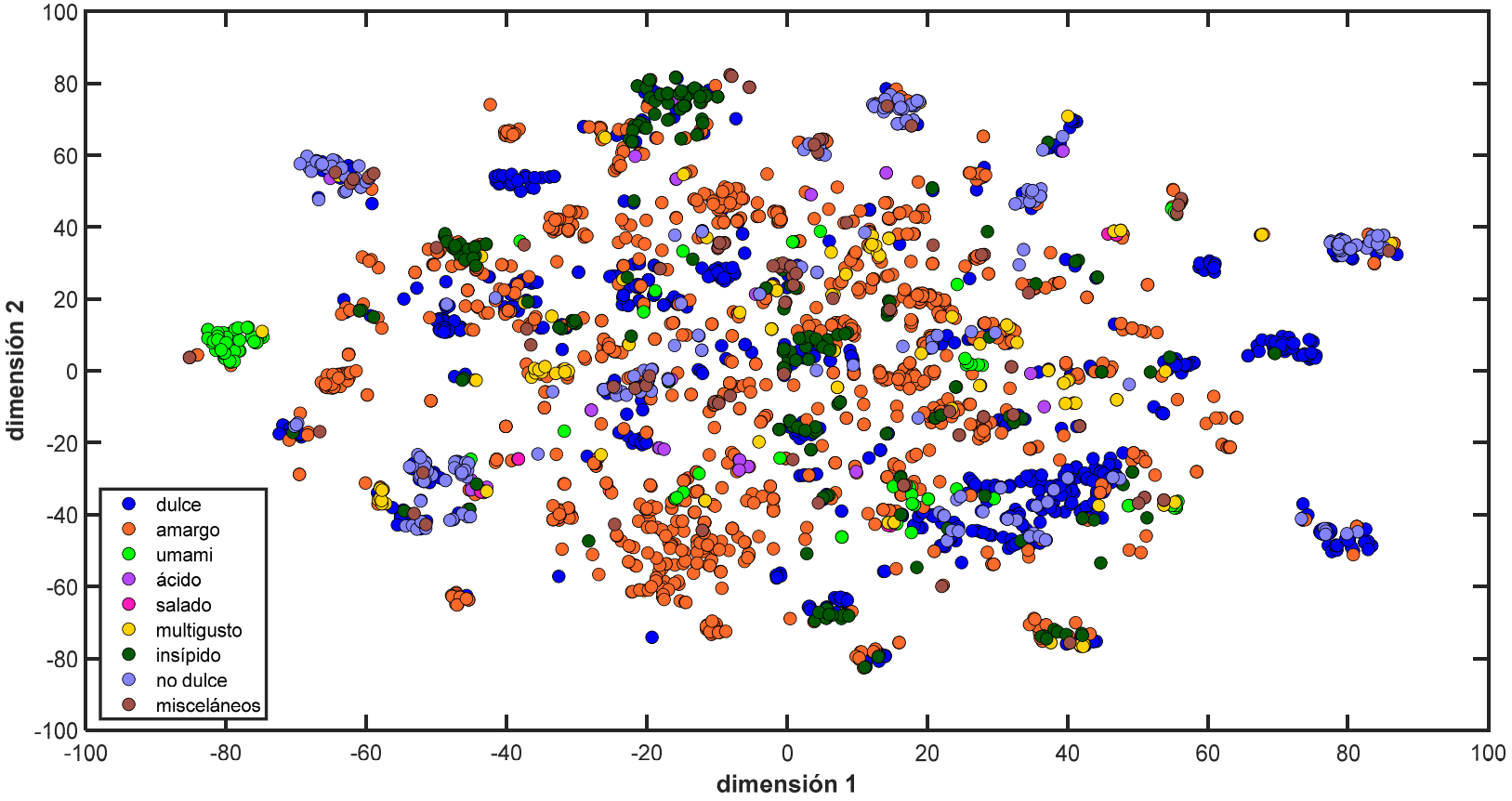

Figura 4.2 Espacio químico del gusto mediante la t–SNE. Exageración 100, perplejidad 30 y tasa de aprendizaje 100

Para un mejor entendimiento y análisis del espacio químico para cada uno de los cinco gustos básicos, se proponen figuras clase/no clase.

En la Figura 4.3 se presenta el espacio químico para el gusto dulce. Se identifican dos grupos (D1 y D2) formados por derivados del ácido guanidinoacético. El grupo D1 posee 30 compuestos, entre ellos el Ácido sucrononico. Complementariamente, el grupo D2 está constituido por 44 derivados dulces y 1 compuesto insípido (Ácido fenil guanidineacético), entre los que se destacan los edulcorantes Bernardame, Carrelame y Lugduname. Por otra parte, en el grupo D3 se encuentran 19 derivados del acesulfamo, entre ellos el acesulfamo, acesulfamo de potasio, 6–Etil–acesulfamo y Aspartamo–acesulfamo.

En el extremo positivo de las dos componentes se localiza el grupo D4, conteniendo 18 compuestos dulces que tienen la característica de poseer dentro de su estructura el grupo funcional fenilsulfonil. En este grupo se encuentra el Sulfone, ASA 1 (2,2–dimethil–1–fenilsulfonilalcanoic ácido), ASA 3 (2,4–dimetil–1–fenilsulfonilciclohexanecarboxilico ácido) y ASA 5 (1–fenilsulfonil–6–metil–3–ciclohexenecarboxilico ácido. En este grupo se encuentran también 7 compuestos amargos, 3 multigusto, 27 no dulces y 1 etiquetado como misceláneos. A continuación, en el grupo D5 se localizan 51 compuestos dulces que pertenecen mayoritariamente a derivados halogenados (mono–, di–, tri– y tetra– sustituidos) de la sacarosa y galactosacarosa; así como el edulcorante sucralosa y tres de sus análogos. Este grupo comparte el espacio químico con tres 3 compuestos amargos y 6 no dulces.

Por otra parte, en el extremo negativo de la primera componente y positivo de la segunda, se encuentra el grupo D6, que alberga a 22 compuestos dulces. Estos dulcificantes se solapan con 3 compuestos amargos, 4 ácidos, 2 multigusto, 25 no dulces y 5 misceláneos. La característica fundamental de estos compuestos dulces es la de pertenecer a la familia de sales sódicas del sulfamato. El edulcorante internacional Sacarosa se localiza en el grupo D7, que es relativamente pequeño pero compacto (16 moléculas). Aquí también se encuentra la Glucosil sacarosa, Lactosacarosa, Lactulosacarosa, Galactosacarosa, D–Lactulosa, Palatinosa, Raffinosa, Sedoheptulosa, Stachiosa. Estos edulcorantes comparten el espacio químico con diversos derivados de la Sacarosa que pertenecen a la clase amargo, insípido, no dulce y misceláneo. Muy cerca, en la región de puntuaciones negativas para las dos componentes, se encuentra un grupo de 36 moléculas

dulces (D8), entre las cuales se encuentran el Ciclamato de sodio, Ciclamato de calcio y diversos derivados de sales sódicas del sulfamato. Este grupo se encuentra solapado con algunos compuestos no dulces, insípidos, amargos y ácidos, principalmente.

Por otra parte, en el grupo D9 se encuentran los compuestos dulces Nitrobenzeno, P 4000, 3–Aminobenzonitrilo, 3–Amino–4–bromobenzonitrilo, ácido guanidinoacético 4–nitrofenil, Suosan (y 3 de sus derivados), Nitroanilina (junto a 12 de sus derivados). Asimismo, se encuentran 11 compuestos híbridos del Suosan–aspartamo, 9 derivados del ácido guanidinoacético (fenílicos y cianofenílicos) y 3 compuestos derivados de la aril úrea, entre los más importantes. Estos compuestos comparten el espacio químico con 16 compuestos amargos (Am6) y 14 moléculas insípidas. A continuación, en el grupo D10 se encuentra el edulcorante Sacarina junto a sus sales de sodio, potasio y calcio, así como 7 derivados sustituidos de la misma. Aquí también se localizan otros derivados de la Sacarina etiquetados como amargos (Am5), insípidos y multigusto.

El grupo D11 contiene 193 dulcificantes que se traslapan con compuestos amargos, umami, no dulces y misceláneos, principalmente. Aquí se encuentra principalmente derivados del Ácido aspártico, por ejemplo, el Aspartamo, Advantamo y Neotamo. En este dominio químico también están dos α–aminoácidos (D–Asparagina y D–Glutamina) y el Ácido guanidino–acético. Finalmente, en el grupo D12 se encuentran 39 compuestos dulces solapados con compuestos insípidos y amargos (Am8), y en menor cantidad con compuestos ácidos y misceláneos. Aquí se identifican a la Hesperetina DHC, Floroglucinol, Resorcinol, trans–Anetol, trans–Cinnamaldehido, así como dos derivados de la Dihidrochalcona y ciertos derivados de la Isocoumarina, entre los más importantes. Los compuestos de este grupo de caracterizan por tener dentro de su estructura química fragmentos moleculares del tipo Ácido benzoico, Resorcinol, Floroglucinol o Anisole.

Los demás compuestos dulces se encuentran dispersos a lo largo del espacio químico y compartiendo espacio con las moléculas de los demás gustos básicos, así como compuestos multigusto, insípidos, no dulce y misceláneos

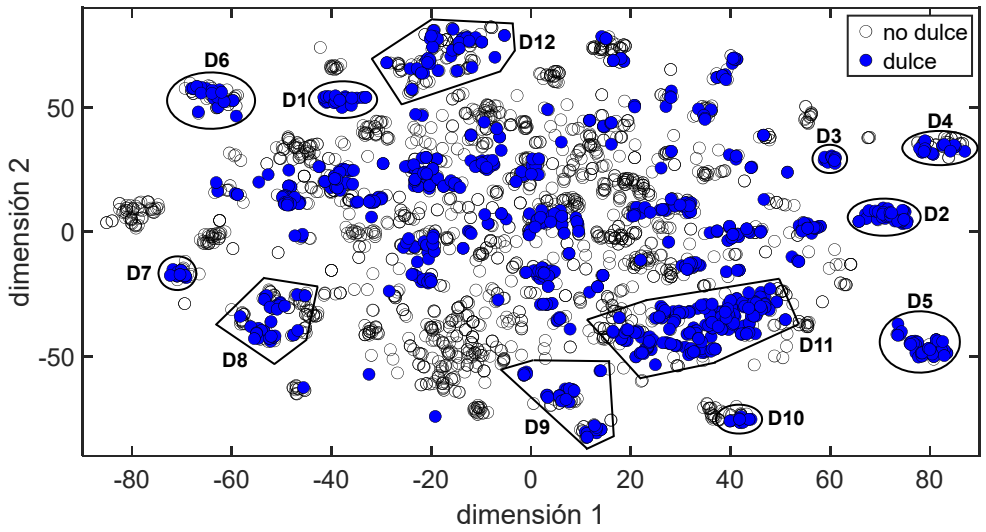

Figura 4.3 Espacio químico para el gusto dulce mediante la t–SNE.

El espacio químico para el gusto amargo se presenta en la Figura 4.4. Aquí se observa que los compuestos que pertenecen a esta clase se dispersan a lo largo de la proyección de las dos componentes, por lo que tienen alta superposición con moléculas de las otras clases. No obstante este hecho, se logran identificar algunos grupos consistentes de moléculas amargas. El grupo Am1 está constituido por 35 compuestos amargos, dentro de los cuales se encuentra el Barbital y diversos derivados del mismo, por ejemplo, Ácido fenilmetilbarbitúrico, Allobarbital, Amobarbital, Aprobarbital, Butallilonal, Ciclobarbital, Ciclopentobarbital, Heptabarbital, Narcobarbital, Nealbarbital y Talbutal. Asimismo, existen las sales sódicas del Barbital, Amobarbital, Aprobarbital, Butabarbital, Butallilonal, Hexobarbital, Pentobarbital, Fenobarbital, Secobarbital y Vinbarbital; así como el Ciclobarbital de calcio.

En el grupo Am2, conformado por 29 moléculas amargas, se ubican principalmente el Xanthohumol (y los derivados B, C, D, G, H, I, L, M, N y P), Isoxantohumol (H, M y P), 1',2'–Dihidroxanthohumol (C, F y K), 2'–Hidroxi–xanthohumol M, 5'–Prenilxanthohumol y 1',2'–Dihidroisoxanthohumol C. A continuación, se identifican 14 moléculas en el grupo Am3, que corresponden en su totalidad a derivados del Lupone: Dehidrotricicloadlupone, Dehidrotriciclocolupone, Dehidrotriciclolupone, Hidroperoxitricicloadlupone, Hidroxitricicloadlupone, Hidroxitriciclocolupone, Hidroxitriciclolupone, Nortricicloadlupone,

Nortriciclocolupone, Nortriciclolupone, Tricicloadlupone, Triciclocolupone y Triciclolupone.

Los demás grupos que se describirán a continuación tienen sobreposición con compuestos de las otras clases. En el espacio químico del grupo Am4 se ubican principalmente compuestos nitrogenados cíclicos, tales como la Pirazina (y los derivados Acetilpirazina y Etilpirazina), gamma,gamma'–dipiridil, Harman, Imidazol, Picolina, Purina, Pirazol, Piridazina, Piridina, Pirimidina, Pirrol y Quinazolina. Además, aquí se encuentra el compuesto dulce 5–Bromopirimidina–2–carbonitrilo, que presenta dentro de su esqueleto el fragmento molecular Pirimidina.

Seguidamente, en el grupo Am5 se identifican 19 derivados amargos de la Sacarina (5–Metoxisacarina, 5–Nitrosacarina, 6–Nitrosacarina, 7–Nitrosacarina y Denatonio sacárido), así como los compuestos amargos Camfotamida, Glimepirida, Sulfisoxazola y Trimetafán camsilato. Estos compuestos comparten el espacio químico con derivados de la sacarina que presentan gusto dulce (grupo D10), insípido (10 compuestos) y multigusto (5 moléculas). Muy cerca se localiza el grupo Am6, constituido por 15 moléculas amargas que comparten el espacio químico con el grupo dulce D9 junto con 14 compuestos insípidos. Aquí se destacan los compuestos amargos Cloramfenicol, Nitrofurazona, Ranitidina hidrocloruro, 1–Nitronaftaleno, 2–Nitroanilina, Picrato de amonio, Azatioprina, Ácido crisamminico, m–Nitrobenzeno y Ácido pícrico. En el extremo opuesto de la componente 2, se ubica el grupo Am7 con 16 sales de sodio derivadas del sulfamato, las que tienen alta similitud molecular con 63 sales sódicas derivadas del sulfamato (10 dulces, 3 ácidos, 6 multigusto, 38 no dulce y 6 misceláneos).

Por otra parte, se identifican 42 compuestos amargos (Am8) que se sobreponen con los compuestos dulces D12, así como con moléculas insípidas ácidas y misceláneas. Entre los compuestos que se ubican en este grupo se tiene al Benzaldehído y compuestos que contienen dentro de su estructura a este compuesto como fragmento molecular, por ejemplo, Acetanisol, Acetofenon, Atranorin, Ácido bencílico, Benzoína, Buteina, Ácido cafeico, Chalcona, Cinchofeno, Dietil ftalato, Eriodictiolchalcona, Isoliquiritigenina, Fenilacetaldehido, Floretina, Pirocatequina, Resveratrol y Salsalato.

Finalmente, el estándar Quinina (con sus sales sulfato, hidrocloruro y dihidrocloruro) se encuentra en el grupo Am9, el más numeroso con 238 compuestos amargos. En este grupo, con alta diversidad molecular, también se encuentran 15 derivados del Denatonio, diversas cadenas de aminoácidos

(6 secuencias lineales y 28 secuencias cíclicas), así como el Ácido pantoténico, Aconina, Acortatarina A, Alerlisina, Berberina, Brucina, Cefaelina, Clorhexidina, Creatinina, Diisobutilamina, Enalapril, Lafutidina, Matrina, Piperazina, Piperidina, Solanidina, Estrichnina, Trimetadiona, α–Chaconina y α–Solanina; entre otros. En términos generales, este grupo se solapa con algunas moléculas dulces, umami, ácidas, multigusto, insípidas y misceláneas. Las demás moléculas amargas se encuentran dispersas a lo largo del resto del espacio químico, con la característica de superponerse con compuestos de las otras clases.

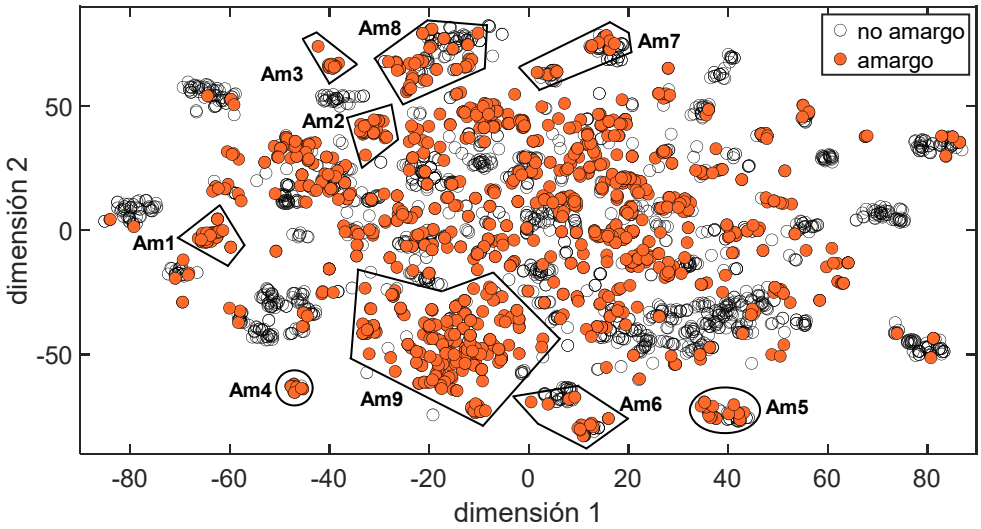

Figura 4.4 Espacio químico para el gusto amargo mediante la t–SNE.

El espacio químico para el gusto umami se presenta en la Figura 4.5, donde se identifican 5 grupos. El grupo U1 se encuentra constituido por la mayoría de compuestos umami (49 moléculas), entre las que se encuentran principalmente diversas sales de sodio del guanilato, inosinato, adelinato, adenósido, ribósido y xantosinato. Aquí también se tiene tres compuestos amargos (Adenosina, Inosina y Pironaridina tetrafosfato), uno multigusto (Cefazolina de sodio) y dos etiquetados como misceláneos (1–(1–b–glucopiranosil)–1H–indole–3–acético ácido y Metil 1–(1–β–glucopiranosil)–1H–indole–3–acetato). Por otra parte, en el grupo U2 se identifican 5 compuestos que tienen la característica principal de poseer dentro de su estructuras grupos amida. Estas moléculas se solapan con cuatro compuestos amargos (Acecarbromal, Dietilbromoacetamida, N–Isobutil acetamida y

Valpromida) y seis moléculas de la clase misceláneos (cis–Pellitorina, Evercool 180, Spilantole, trans–Pellitorina, WS3 y WS23).

El grupo U3 está formado por el estándar Glutamato monosódico (MSG) y otros tres glutamatos (monopotásico, monoamónico y monosódico D,L–treo–β–hidroxi). En este grupo también se incluyen dos Diglutamatos (cálcico y magnésico), el L–aspartato monosódico, el L–α–amino adipato monosódico y los aminoácidos Thr–Glu y Glu–Asp–Glu. Muy cerca se encuentra el grupo U4, en el que se ubica el Ácido L–Iboténico, Ácido L–Tricolómico (forma eritro), L–Teanina, Asp–Glu–Ser, γ–L–Glutamil–L–(S–metil) metionina, γ–L–Glutamil–L–cisteinil–glicina, N–fenetil–4–hidroxipentanamida, N–2,4–Dimetoxibenzil–N–(2–piridil)etil oxalamida, Etil 4–((2–isopropil–5–metilciclohexiloxi)carbonil)butanoato, N–(3–metoxi–4–hidroxi–benzil)–5–hidroxipentanamida. En este grupo también se ubican los compuestos N–(4–hidroxifenetil) de la eritronamida, gluconamida y succinamida.

Finalmente, en el grupo U5 se encuentran siete derivados del inosinato, entre ellos el compuesto 2–Mercaptoinosinato 5''–monofosfato, dos sales de calcio (Inosinato de calcio y 2–alliloxi–5''–inosinato de calcio) y cuatro sales disódicas (Disodio 2–metoxi–5''–inosinato, Disodio 2–metil–5''–inosinato, Disodio N1–metil–5''–inosinato y Disodio N1–metil–2–metiltio–5''–inosinato). Los demás compuestos umami se encuentran dispersos a lo largo del espacio químico del gusto y sobrepuestos con otras clases de compuestos.

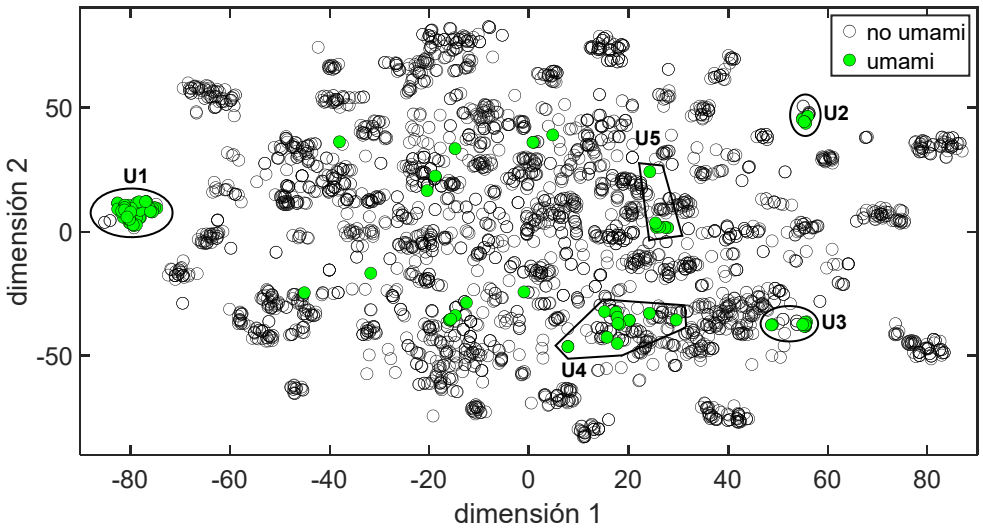

Figura 4.5 Espacio químico para el gusto umami mediante la t–SNE.

El espacio químico para el gusto ácido se presenta en la Figura 4.6. El grupo Ac1 está conformado por 4 sales sódicas del sulfamato: Sodio N–(1,3–benzotiazol–2–il)sulfamato, Sodio N–(4–metil–5–propil–1,3–tiazol–2–il)sulfamato, Sodio N–(5–benzil–1,3,4–tiadiazol–2–il)sulfamato y Sodio N–(5–tert–butil–1,2–oxazol–3–il)sulfamato. A continuación, en el grupo Ac2 se encuentran ocho compuestos ácidos que corresponden a sales disódicas del ácido imidodisulfúrico, las cuales se superponen contemporáneamente con unas pocas moléculas saladas, multigusto, insípidas y no dulces. Por otra parte, en el grupo Ac3 se encuentran los ácidos orgánicos presentes en algunos alimentos (acético, propiónico, cítrico, láctico, málico y tartárico), así como el ácido fórmico, carbónico, fosfórico y dos sales sódicas (Sodio 3–(sulfonatoamino)benceno–1–sulfonato y Sodio N–[4–(butan–2–il)fenil]sulfamato). Los restantes compuestos salados se encuentran dispersos y sobrepuestos en el espacio químico de los demás gustos básicos, así como multigusto y no dulce.

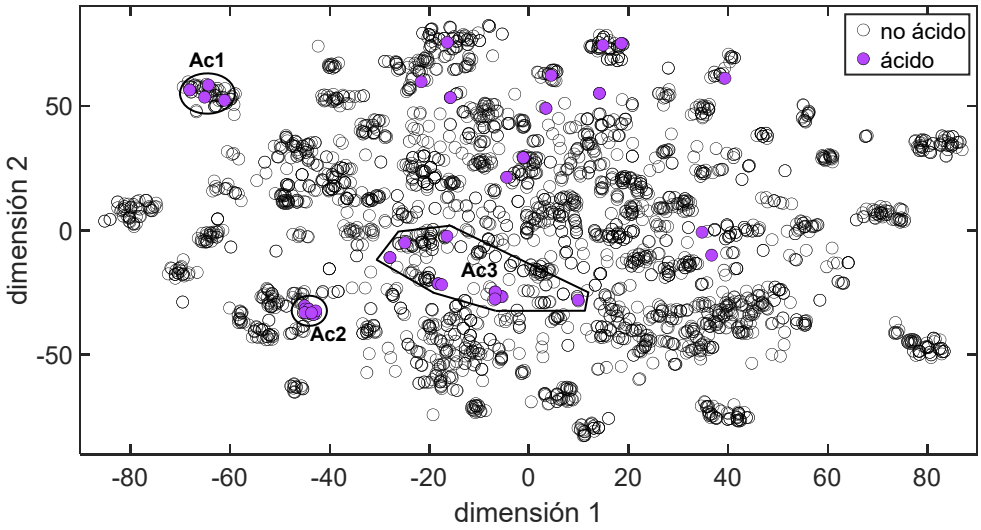

Figura 4.6 Espacio químico para el gusto ácido mediante la t–SNE.

La Figura 4.7 define el espacio químico para el gusto salado (menos numeroso), en el que se identifica la presencia de 3 grupos consistentes de moléculas que se encuentran superpuestas con estructuras de los otros cuatro gustos básicos y algunos compuestos con multigusto. Así, el grupo S1 está formado por las sales cloruro de sodio, cloruro de potasio, cloruro de litio y cloruro de amonio. Por otra parte, en el grupo S2 se encuentran los compuestos L–Ornitil–γ–ácido aminobutírico, Orn–βAla.HCl y Orn–

γAbu.HCl. Finalmente, en el grupo S3, que es el más disperso, se localizan los compuestos L–Ornitiltaurina, Lys–Tau.HCl, Orn–Tau.HCl, Disodio N,N–disulfonatopentan–1–amina y N–(5–metilheptil)ácido imidodisulfúrico disodio sal.

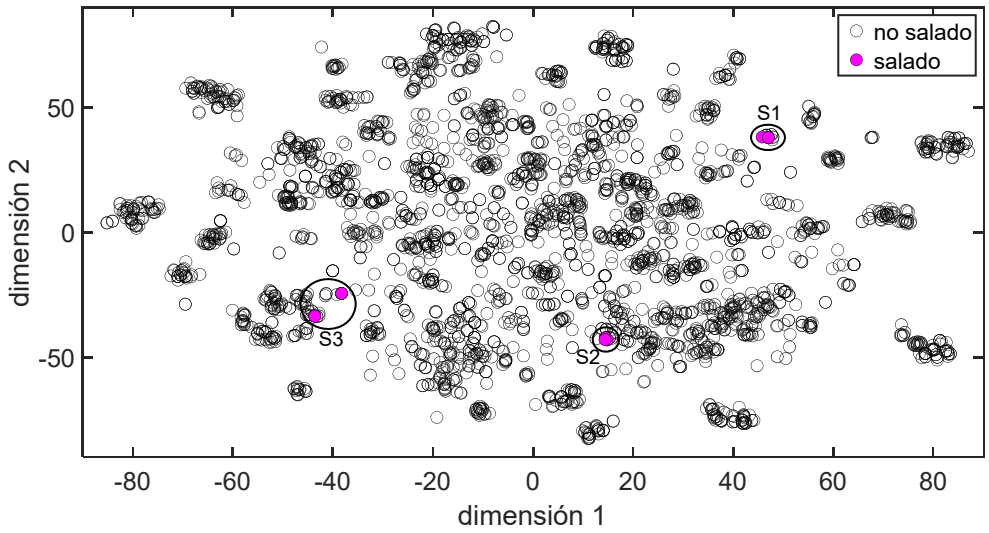

Figura 4.7 Espacio químico para el gusto salado mediante la t–SNE.

4.1.3. Conclusiones

El espacio químico desarrollado en esta aplicación del aprendizaje no supervisado constituye el primero para analiza una base de datos extensa de los cinco gustos básicos, así como otras clases de percepciones que se logran al analizar sensorialmente diversos compuestos naturales o sintetizados en el laboratorio. La incrustación de vecinos estocásticos distribuidos en t (t–SNE) permitió establecer un gráfico de dispersión bidimensional a partir de un espacio de alta dimensionalidad definido por las claves moleculares del sistema de acceso molecular (MACCS). La automatización del algoritmo t–SNE permitió definir el mapa bidimensional con una exageración E = 100, perplejidad P = 30, tasa de aprendizaje L = 100 y el coeficiente de similitud de Jaccard–Tanimoto. El espacio químico permite establecer las regiones de alta similitud molecular para cada gusto, así como definir grupos consistentes de moléculas. Sin embargo, se observa también alta superposición entre los compuestos de los cinco gustos básicos, así como con las estructuras moleculares de las demás clases.

Es importante indicar que la asignación de los gustos es altamente ruidosa, debido al error humano involucrado en la medición de esta actividad molecular mediante paneles de catadores. Otro factor que contribuye al ruido experimental es que la mayoría de compuestos no presentan un gusto puro y limpio, es decir, sin la percepción de otros gustos o retrogustos (deseables o indeseables). También se debe considerar que la percepción del gusto depende de la naturaleza del quimiorreceptor (células gustativas) y la forma en que un ligando (compuesto) interacciona con el mismo; por ejemplo, algunos compuestos interaccionan con el quimiorreceptor pero no generan ningún estímulo (falsos positivos); mientras que otros ligandos que no interactúan con dicho quimiorreceptor generan una señal de sensación del gusto (verdaderos negativos).

Finalmente, la facilidad de implementación del algoritmo de la t–SNE permite que nuevas moléculas sean proyectadas en el espacio químico para tener una idea cualitativa del gusto predominante a partir de los compuestos con alta similitud molecular. De esta forma, constituye una herramienta útil para los investigadores que trabajan en la búsqueda de nuevos compuestos de sabor como potenciales aditivos para ser usados en la industria de alimentos y farmacéutica.

4.2. Aprendizaje supervisado para predecir los gustos básicos

4.2.1. Materiales y métodos

Para esta aplicación se ha utilizado la base de datos ChemTastesDB, presentada en la sección anterior, considerando los compuestos de los cinco gustos básicos, es decir, 977 dulces, 1183 amargos, 98 umami, 38 ácidos y 12 salados. Por lo tanto, se han excluido las 636 moléculas que pertenecen a las clases no dulce, insípido, multigusto y misceláneos. Para cada compuesto se han calculado 3 grupos de variables descriptoras de la estructura química en el programa alvaDesc [52]:

- 166 claves moleculares del sistema de acceso molecular (MACCS).
- 1024 huellas dactilares moleculares de conectividad ampliada (ECFPs).
- 4176 descriptores moleculares independientes de la conformación.

Para optimizar el tiempo de cálculo durante la selección supervisada, se han eliminado descriptores no informativos, es decir, 1147 constantes, 127 casi constantes y 849 con al menos un valor faltante. De esta manera se han retenido 2053 descriptores moleculares para el desarrollo de los modelos de clasificación. Posteriormente, se afinó aún más la reducción de los descriptores al aplicar la reducción no supervisada basada en el algoritmo de Wootton, Sergent y Phan–Tan–Luu (V–WSP), para de esta manera seleccionar un subconjunto representativo de los mismos de tal forma que se encuentre una mínima correlación entre ellos y cubran todo el espacio multidimensional.

Debido a que cada molécula está etiquetada por una respuesta (variable aleatoria) discreta nominal, definida por la pertinencia a una clase de los cinco gustos básicos, se han aplicado diversos métodos del aprendizaje supervisado de clasificación. Considerando las MACCS y las ECFPs, se han usado los clasificadores basados en similitudes locales de los *k*–vecinos más cercanos (*k*NN), *N*–vecinos más cercanos (N3) y los vecinos más cercanos agrupados (BNN). Debido al hecho de que estos dos grupos de variables (MACCS y ECFPs) son binarias, se ha utilizado el coeficiente de similitud de Jaccard–Tanimoto. Por otra parte, con el grupo de descriptores moleculares se han aplicado los métodos de clasificación de los bosques aleatorios (RF) y el análisis discriminante de mínimos cuadrados parciales (PLSDA).

Para medir la capacidad predictiva de los modelos, la base de datos se dividió de forma aleatoria en conjuntos de calibración (70% de compuestos) y predicción (30% de las moléculas), de tal forma que se conserve la proporción en la numerosidad de las clases. De esta forma se obtiene similar representatividad de los cinco grupos y se evita sesgos hacia las clases más numerosas (dulce y amargo). En el caso de los modelos basados en los clasificadores *k*NN, N3, BNN y RF, el grupo de calibración se usó para ajustar los modelos, mientras que el grupo de predicción se utilizó para medir la capacidad de realizar predicciones de moléculas no consideradas durante la calibración de los mismos. Por otro lado, para desarrollar el modelo PLSDA, las moléculas del grupo de calibración se usaron para la selección supervisada de descriptores mediante los algoritmos genéticos (GAs).

Para todos los modelos se utilizó la validación cruzada de ventanas venecianas con 5 grupos para obtener los parámetros óptimos de cada clasificador, es decir, *k* para *k*NN, α para N3 y BNN, árboles y hojas para

RF y variables latentes (LVs) para PLSDA. De esta manera, se maximiza la tasa de aciertos o exactitud balanceada en esta etapa (NER_{cv}).

Todos los modelos fueron calibrados y validados en el lenguaje de programación MATLAB. Para los modelos *k*NN y PLSDA se recurrió al "classification toolbox versión 5.3" [119], mientras que para N3 y BNN se utilizó el "N3–BNN toolbox versión 1.0" [112]. Finalmente, los RF se implementaron mediante algoritmos propios que invocan el "Statistics and Machine Learning Toolbox"; mientras que para los GAs se programaron diversas funciones específicas.

4.2.2. Resultados y discusión

A partir de los 2053 descriptores retenidos luego de la exclusión de los valores constantes, casi constantes y al menos un valor faltante, se aplicó el método de reducción no supervisado V–WSP a un umbral de correlación de 0.95. De esta manera se excluyeron 836 variables redundantes y multicolineales, de tal forma que se consideren únicamente 1217 descriptores para el desarrollo de los modelos con los RF y PLSDA. Este aspecto reduce significativamente el cálculo computacional involucrado durante la selección supervisada. A continuación, los 2308 compuestos de la base de datos se dividieron en conjuntos de calibración con 1616 moléculas y predicción con 692 compuestos. El detalle del número de moléculas presente en cada clase se encuentra en la Tabla 4.1.

Tabla 4.1 Detalle del número de las moléculas de los conjuntos de calibración y predicción para cada uno de los cinco gustos básicos

	Dulce	Amargo	Umami	Ácido	Salado
Calibración	684	828	69	27	8
Predicción	293	355	29	11	4

De esta manera, las 1616 estructuras moleculares del conjunto de calibración se usaron para el ajuste de los modelos. Para evitar la presencia de sobreajuste, se ha optimizado la tasa de aciertos o exactitud balanceada en validación cruzada de ventanas venecianas (NER_{cv}). Únicamente para el clasificador PLSDA se ha usado el conjunto de calibración para la selección supervisada con los GAs, también optimizando la tasa de aciertos en validación cruzada. De esta forma, se obtienen ocho modelos de clasificación que permiten predecir la pertinencia de una molécula a una de los cinco gustos básicos. Los resultados se presentan en la Tabla 4.2.

Tabla 4.2 Parámetros globales de calidad de los modelos QSAR para la predicción de los cinco gustos básicos

Modelo	Parámetro óptimo	Tasa de aciertos (*NER*)			Exactitud (*Acc*)		
		cal	val	pred	cal	val	pred
MACCS–*k*NN	k = 2	0.746	0.725	0.697	0.861	0.857	0.871
MACCS–N3	α = 1.25	0.825	0.774	0.853	0.780	0.778	0.795
MACCS–BNN	α = 0.3	0.726	0.726	0.739	0.860	0.860	0.867
ECFPs–*k*NN	k = 2	0.739	0.732	0.810	0.860	0.855	0.864
ECFPs–N3	α = 1.25	0.852	0.816	0.876	0.807	0.807	0.806
ECFPs–BNN	α = 0.45	0.753	0.748	0.791	0.864	0.855	0.860
Descriptores–RF	árboles = 50 hojas = 1	0.999	0.634	0.607	0.998	0.882	0.879
Descriptores–PLSDA	d = 14 LVs = 5	0.640	0.640	0.650	0.610	0.620	0.630

Se observa que el clasificador de los *N*–vecinos más cercanos tiene buena calidad al aplicarlo con las huellas dactilares moleculares de conectividad ampliada ($NER_{cal} = 0.852$, $NER_{val} = 0.816$ y $NER_{pred} = 0.876$) y las claves moleculares del sistema de acceso molecular ($NER_{cal} = 0.825$, $NER_{val} = 0.774$ y $NER_{pred} = 0.853$). Por otra parte, los dos peores modelos para realizar nuevas predicciones se obtienen con los descriptores moleculares y los métodos de clasificación de los RF ($NER_{pred} = 0.607$) y los PLSDA ($NER_{pred} = 0.650$). De forma particular, se observa que los bosques aleatorios sobre ajustan el modelo ($NER_{cal} = 0.999$); sin embargo, en validación y particularmente en predicción decrece rápidamente la calidad del mismo. Esto indica que el modelo aprende demasiado de la información del grupo de calibración, por lo que no presenta buena predictividad para los compuestos que no fueron considerados durante la calibración del mismo.

Tabla 4.3 Matriz de confusión del grupo de predicción para el modelo de los bosques aleatorios con los descriptores moleculares

		Clases predichas				
		Dulce	Amargo	Umami	Ácido	Salado
Clases verdaderas	Dulce	259	34	0	0	0
	Amargo	26	326	0	1	2
	Umami	6	5	18	0	0
	Ácido	3	4	0	4	0
	Salado	0	3	0	0	1

Para entender qué sucede con este modelo, se ha analizado la matriz de confusión del conjunto de predicción (Tabla 4.3). A partir de la matriz de confusión se ha obtenido los valores de los índices primarios de clasificación para cada una de las clases: sensibilidad, especificidad y precisión (Tabla 4.4). Particularmente, la sensibilidad brinda la información relacionada a la capacidad del modelo de reconocer correctamente las moléculas que pertenecen a cada clase. Las clases dulce y amargo presentan buena capacidad de reconocimiento ($Sn = 0.884$ y $Sn = 0.918$, respectivamente), decreciendo para la clase umami con un 62.1 % ($Sn = 0.621$) y empeorando drásticamente para las clases ácido (36.4 %) y salado (25 %). La habilidad relativamente baja del modelo de reconocer moléculas umami y muy mínima para las clases ácido y salado se debe posiblemente a la poca numerosidad de los compuestos de estas tres clases con relación a la base de datos completa. De hecho, entre estas tres clases se alcanza únicamente el 6.4 % de la información total. Es lógico intuir que las clases dulce y amargo van a tener mayor sensibilidad debido a su amplio dominio químico (sección anterior) y por consiguiente tender a sobreestimar las predicciones. Adicionalmente, en la sección 4.1 se pudo observar la alta sobreposición de las moléculas dulces y amargas con los compuestos de las otras tres clases.

Tabla 4.4 Parámetros primarios de clasificación para el grupo de predicción para el modelo de los bosques aleatorios con los descriptores moleculares

Clase	*Sn*	*Sp*	*Pr*
Dulce	0.884	0.912	0.881
Amargo	0.918	0.864	0.876
Umami	0.621	1.000	1.000
Ácido	0.364	0.999	0.800
Salado	0.250	0.997	0.333

No obstante esta limitación del desbalance en la representatividad de las clases umami, ácido y salado, el modelo desarrollado con las huellas dactilares moleculares de conectividad ampliada y el método de clasificación de *N*–vecinos más cercanos resulta ser el mejor. La tasa de aciertos es comparable en calibración, validación y predicción, lo que claramente indica que el modelo no sufre sobreajuste y puede ser aplicado con una bondad del 87.6 % de confianza para la predicción de nuevas moléculas. Una posible explicación para el buen desempeño del clasificador N3 está en el uso de todos los n–1 objetos para clasificar un compuesto. Es particularmente interesante que el clasificador N3 considera la similitud (s) y el vector el

vector de ranking (r) para obtener una ponderación en la asignación de las clases (Ec. 3.21). Es decir, para cualquier molécula existe una mayor contribución de los compuestos más cercanos en el espacio multidimensional (más similares), la cual va disminuyendo hasta la menor contribución brindada por los objetos más lejanos. De esta forma, es posible superar el efecto de la superposición de las clases y el desbalance en la numerosidad de las mismas.

4.2.3. Conclusiones

En esta aplicación se ha desarrollado modelos QSAR basados en diversas máquinas del aprendizaje automático para la predicción de los gustos básicos. El mejor modelo se obtiene al aplicar el clasificador de los N–vecinos más cercanos utilizando las huellas dactilares moleculares de conectividad ampliada. De forma similar, N3 también funciona bien cuando se aplican las claves moleculares MACCS. Por el contrario, los modelos de los bosques aleatorios y el análisis discriminante de mínimos cuadrados parciales son los métodos menos apropiados para enfrentar problemas de multiclase, particularmente donde existe un desbalance marcado en la numerosidad de los compuestos de las mismas. De esta manera, este estudio refleja el primer enfoque para estudiar contemporáneamente los cinco gustos básicos y permite realizar predicciones de nuevos compuestos con una confianza del 87.6 %.

4.3. Aprendizaje supervisado para discriminar compuestos dulces y amargos

4.3.1. Materiales y métodos

Debido a que las clases más numerosas corresponden a la dulce y amargo, se han considerado las mismas para desarrollar modelos de clasificación que permitan discriminar entre las moléculas de estos dos gustos básicos. Por lo tanto, de la base de datos ChemTastesDB (sección 4.1.2) se han considerado únicamente las 977 moléculas dulces y 1183 amargas (se han excluido los 784 compuestos umami, ácido, salado, no dulce, insípido, multigusto y misceláneo). Para el desarrollo de los modelos, se ha seguido la metodología detallada en la sección 4.2.1 para las máquinas de aprendizaje con los cinco gustos básicos.

4.3.2. Resultados y discusión

Se usaron las 2160 estructuras moleculares de las clases dulce y amargo para realizar una partición aleatoria de la base de datos en conjuntos de calibración y predicción en una proporción 70:30 (manteniendo la proporción en la numerosidad). De esta manera, 1512 moléculas constituyen el conjunto de calibración (684 dulces y 828 amargas), mientras que los restantes 648 compuestos definen el conjunto de predicción (293 dulces y 355 dulces). Los modelos QSAR basados en métodos de clasificación se presentan en la Tabla 4.5.

Tabla 4.5 Parámetros globales de calidad de los modelos QSAR para la discriminación de moléculas dulces y amargas

Modelo	Parámetro óptimo	Tasa de aciertos (*NER*)			Exactitud (*Acc*)		
		cal	val	pred	cal	val	pred
MACCS–*k*NN	k = 3	0.888	0.880	0.869	0.884	0.876	0.864
MACCS–N3	$\alpha = 1.25$	0.885	0.881	0.871	0.880	0.876	0.864
MACCS–BNN	$\alpha = 0.9$	0.885	0.879	0.870	0.881	0.875	0.866
ECFPs–*k*NN	k = 2	0.873	0.870	0.878	0.870	0.867	0.873
ECFPs–N3	$\alpha = 1$	0.871	0.872	0.870	0.864	0.866	0.861
ECFPs–BNN	$\alpha = 0.3$	0.879	0.879	0.886	0.876	0.876	0.881
Descriptores–RF	árboles = 100 hojas = 1	0.998	0.908	0.914	0.998	0.909	0.912
Descriptores–PLSDA	d = 7 LVs = 2	0.760	0.760	0.740	0.770	0.770	0.750

En el caso de este problema de clasificación binaria, todos los modelos presentan resultados comparables en calibración y validación (interna y externa), con lo que se verifica la ausencia de sobreajuste en los mismos. Prácticamente todos los modelos (excepto el PLSDA) presentan una predictividad del 86 % o superior. Entre estos modelos destaca el basado en los bosques aleatorios usando los descriptores, el cual tiene la calidad de predicción más alta de todos (91.4 % de confianza). Por el contrario, el modelo basado en el análisis discriminante de mínimos cuadrados parciales es el que genera la predicción más baja (74 %). No obstante este valor, en términos generales se puede asumir como un modelo con una calidad aceptable.

Se había presentado previamente en la sección 2.4 (Tablas 2.1 y 2.2) once modelos QSAR que se han desarrollado a lo largo de la historia para discriminar los gustos dulce y amargo. De todos estos modelos, solo dos

utilizan los bosques aleatorios. El primero, publicado en el año 2017 [89], utiliza una base de datos de 992 compuestos y tiene una tasa de aciertos del 91.4 %; mientras que el segundo, publicado un año más tarde [9], usa 1202 moléculas y alcanza una exactitud del 96.7 %. En términos generales, la calidad del modelo desarrollado en este estudio es comparable con los dos modelos basados en el mismo clasificador, aunque es importante destacar que aquí se ha utilizado una cantidad significativamente mayor de moléculas, lo que permite cubrir mayor dominio de diversidad química para la predicción de nuevas moléculas.

4.3.3. Conclusiones

Las relaciones cuantitativas estructura–actividad desarrolladas para las clases de compuestos dulce y amargo presentan buena calidad en calibración, validación interna de dejar–varios–fuera y particularmente en predicción. De entre todos los clasificadores utilizados, recurriendo a tres formas de representación de la estructura molecular, destaca el de los bosques aleatorios con una predictividad del 91.4 %. El modelo obtenido es comparable con sus análogos reportados en la literatura.

4.4. Aprendizaje supervisado para predecir el dulzor

4.4.1. Materiales y métodos

En esta aplicación se desarrollarán modelos QSAR para el gusto dulce, considerando una clasificación binaria clase/no clase, es decir, dulce/no dulce. Para este propósito, se han usado las 977 moléculas dulces, mientras que para definir la clase no dulce se han fusionado los compuestos de las clases amargo, umami, ácido, salado, no dulce e insípido (1767 compuestos) de la base de datos ChemTastesDB. Se han excluido las moléculas de las clases multigusto y misceláneo. Aquí también se han aplicado los métodos descritos para los modelos con los cinco gustos básicos (sección 4.2.1) y las clases dulce–amargo (sección 4.3.1).

4.4.2. Resultados y discusión

Los 2744 compuestos dulces y no dulces presentes en esta base de datos, se han utilizado para definir los conjuntos de calibración (70 %) y predicción (30 %) mediante una división aleatoria y proporcional a la numerosidad de las clases. Luego de aplicar el algoritmo de partición, se han asignado 1921 moléculas al conjunto de calibración (684 dulces y 1237 no dulces) y 823 moléculas al grupo de predicción (293 dulces y 530 no dulces). Utilizando las tres formas de representación de la estructura molecular, se han desarrollado los diversos modelos de clasificación binaria que se muestran en la Tabla 4.6.

Tabla 4.6 Parámetros globales de calidad de los modelos QSAR para la predicción del dulzor

Modelo	Parámetro óptimo	Tasa de aciertos (*NER*)			Exactitud (*Acc*)		
		cal	val	pred	cal	val	pred
MACCS–*k*NN	k = 5	0.817	0.822	0.809	0.818	0.821	0.804
MACCS–N3	$\alpha = 1.5$	0.821	0.829	0.829	0.812	0.818	0.817
MACCS–BNN	$\alpha = 1.25$	0.810	0.818	0.828	0.815	0.823	0.825
ECFPs–*k*NN	k = 6	0.817	0.813	0.820	0.817	0.814	0.815
ECFPs–N3	$\alpha = 1.5$	0.811	0.812	0.831	0.803	0.805	0.819
ECFPs–BNN	$\alpha = 1.05$	0.814	0.819	0.842	0.816	0.821	0.840
Descriptores–RF	árboles = 25 hojas = 1	0.997	0.811	0.853	0.997	0.837	0.870
Descriptores–PLSDA	d = 9 LVs = 5	0.730	0.730	0.740	0.760	0.760	0.770

Se observa que no existe sobreajuste de los modelos y la calidad es comparable entre ellos, excepto el modelo PLSDA que es el peor de todos. Aquí también presenta muy buen desempeño el clasificador de los bosques aleatorios, con una capacidad de predicción del 85.3 %. Se nota que la calidad de descripción de los datos es bastante alta (99.7 %), con una variación porcentual menor al 15 % con respecto a la predictividad; por lo que podría considerarse un modelo aceptable para la predicción del dulzor de nuevas moléculas. Es importante también destacar que el modelo de los vecinos más cercanos agrupados (BNN) aplicado a las huellas dactilares moleculares de conectividad ampliada también es apropiado para la predicción del dulzor. De hecho, este modelo presenta variaciones mínimas en calibración, validación cruzada y predicción.

En la literatura se encuentran citados doce modelos para la predicción del gusto dulce (sección 2.4). Destacan los modelos presentados en la Tabla 2.2

que se basan en un sistema experto ($NER_{pred} = 0.848$) [5], en el análisis de consenso ($NER_{pred} = 0.900$) [91] y en el clasificador AdaBoost ($NER_{pred} = 0.834$) [24]. Al comparar con sus análogos, se observa que el clasificador RF logra un desempeño similar. Un aspecto que destaca del modelo desarrollado en esta aplicación, es el uso de una base de datos más extensa que las previamente citadas, con lo que se cubre mayor dominio químico en términos de diversidad molecular.

4.4.3. Conclusiones

En este estudio se han desarrollado diversos modelos del aprendizaje automático con la finalidad de predecir el dulzor de nuevas moléculas. El mejor modelo corresponde al de los bosques aleatorios junto con los descriptores moleculares, alcanzando una confianza del 85.3 % para discriminar entre moléculas dulces y no dulces. Este modelo complementa a los previamente reportados en la literatura y es aplicable para el descubrimiento de nuevos edulcorantes como sustitutos adecuados de la sacarosa y ciertos edulcorantes bajos en calorías (sacarina y ciclamato) y que puedan ser potencialmente usados en la industria como aditivos para la elaboración de alimentos bajos en calorías.

4.5. Aprendizaje supervisado para predecir el amargor

4.5.1. Materiales y métodos

Debido a que la discriminación del amargor ha adquirido prácticamente la misma importancia que el dulzor, en esta última aplicación se desarrollarán modelos de clasificación QSAR binarios para discriminar moléculas amargas y no amargas (clase/no clase). Con esta finalidad, partiendo de la base de datos ChemTastesDB, se ha definido la clase no amargo mediante la fusión de las clases dulce, umami, ácido, salado e insípido (1328 compuestos). Para evitar ambigüedades de clasificación, se ha excluido la clase etiquetada como no dulce, pues el gusto amargo corresponde a una de las opciones al etiquetar un compuesto como no dulce. También se han excluido las moléculas con multigusto y misceláneos. Para esta última aplicación también se han seguido la misma metodología indicada para los modelos anteriores (secciones 4.2.1, 4.3.1 y 4.4.1).

4.5.2. Resultados y discusión

La base de datos se encuentra conformada por 2511 estructuras moleculares, divididas en dos clases: 1183 amargas y 1328 no amargas. Para efectos de validación de los modelos, la base de datos se ha dividido en grupos de calibración y predicción en una relación 70:30. De esta forma, el algoritmo asigna 1758 compuestos al grupo de calibración (828 amargos y 930 no amargos) y las restantes 753 moléculas al grupo de validación (355 amargas y 398 no amargas). A continuación se han desarrollado los ocho modelos QSAR basados en el aprendizaje supervisado, cuyos resultados se presentan en la Tabla 4.7. Al igual que en el modelo para predecir el dulzor (sección 4.4), el clasificador de los bosques aleatorios brinda la mejor capacidad predictiva (89.8 % de bondad), seguido muy de cerca por el modelo de los *N*–vecinos más cercanos (con las ECFPs) con el 87.2 % de predictividad. También en este caso el modelo PLSDA tiene la más baja calidad para realizar nuevas predicciones (78 %).

Tabla 4.7 Parámetros globales de calidad de los modelos QSAR para la predicción del amargor

Modelo	Parámetro óptimo	Tasa de aciertos (*NER*)			Exactitud (*Acc*)		
		cal	val	pred	cal	val	pred
MACCS–*k*NN	k = 1	0.868	0.871	0.845	0.870	0.873	0.846
MACCS–N3	$\alpha = 1.5$	0.873	0.876	0.841	0.876	0.879	0.843
MACCS–BNN	$\alpha = 0.5$	0.869	0.864	0.844	0.871	0.867	0.845
ECFPs–*k*NN	k = 1	0.855	0.862	0.855	0.858	0.865	0.862
ECFPs–N3	$\alpha = 1.25$	0.864	0.867	0.872	0.868	0.871	0.874
ECFPs–BNN	$\alpha = 1$	0.861	0.862	0.865	0.864	0.866	0.867
Descriptores–RF	árboles = 150 hojas = 1	0.999	0.899	0.898	0.999	0.901	0.899
Descriptores–PLSDA	d = 13 LVs = 6	0.820	0.820	0.780	0.830	0.830	0.790

De los seis modelos citados en la literatura, cinco han sido publicados en la última década (Tabla 2.2); de los cuales dos modelos utilizan las máquinas de soporte vectorial (SVM) [88,92], otro el clasificador AdaBoost [23], mientras que el cuarto se basa en el análisis de consenso [90]. Finalmente, el único modelo basado en los RF [24] considera 2411 estructuras moleculares y alcanza una capacidad predictiva del 81.9 %. Claramente se evidencia que el modelo obtenido en esta aplicación supera en un 8 % a la predictividad del mismo. Por otra parte, los demás modelos basados en otros enfoques de

clasificación usan bases de datos con menor numerosidad de compuestos (ver Tabla 2.2), con lo que el dominio químico de los mismos presente mayores limitaciones de generalización en la predicción de nuevos compuestos amargos.

4.5.3. Conclusiones

La mejor predicción del amargor se logra con el clasificador de los bosques aleatorios junto con los descriptores moleculares. El modelo tiene una capacidad discriminante del 89.8 %, que supera al único modelo basado en el mismo enfoque de modelado previamente reportado en la literatura. La superioridad de los RF en los modelos dulce–amargo, dulce–no dulce y amargo–no amargo, se debe a su característica de ensacado «bagging», lo que le permite realizar el consenso (voto mayoritario) entre los árboles (diversos modelos individuales).

DISCUSIÓN FINAL Y PERSPECTIVAS FUTURAS

En la presente tesis de maestría se ha recopilado una base de datos de moléculas de gusto extensa a partir de la información que se encuentra dispersa en la literatura. Un aspecto crucial ha constituido el curado de las estructuras moleculares, la verificación y filtrado de la información, de tal forma que la base de datos a ser usada con el aprendizaje automático sea completamente válida. Este proceso minucioso ha permitido compilar la base de datos ChemTastesDB que cubre un amplio dominio químico, es decir, los cinco gustos básicos más otras categorías de compuestos. En consecuencia, se constituye en la más completa para los cinco gustos básicos que se dispone hasta el momento.

La primera forma de estudiar la base de datos ha sido mediante la aplicación del aprendizaje no supervisado de la incrustación de vecinos estocásticos distribuidos en t (t–SNE), para generar un diagrama de dispersión bidimensional donde se encuentren reproducidas las similitudes/disimilitudes del espacio de alta dimensionalidad. Esta metodología, relativamente nueva ha sido poco explorada para definir el espacio químico del gusto y en las dos únicas aplicaciones reportadas en la literatura no se realiza un estudio pormenorizado del significado de los diferentes grupos de compuestos para cada clase. El análisis desarrollado en esta aplicación contribuye a un mejor entendimiento del carácter estructural de las moléculas que tienen alta similitud química. Adicionalmente, en esta aplicación se han analizado diversos valores de los parámetros involucrados en la t–SNE, de tal forma de tener una mejor exploración de los distintos mapas químicos generados.

En las aplicaciones subsiguiente se han recurrido a diversas técnicas del aprendizaje supervisado de clasificación, tales como los k–vecinos más cercanos (kNN), N–vecinos más cercanos (N3), vecinos más cercanos

agrupados (BNN), bosques aleatorios (RF) y el análisis discriminante de mínimos cuadrados parciales (PLSDA). Estos enfoques supervisados han permitido desarrollar diversos modelos de clasificación a partir de tres formas modernas de representación de la estructura molecular (MACCS, ECFPs y descriptores moleculares), de tal forma que se logre relacionar el gusto con los atributos químicos de los compuestos. Se han propuesto modelos para los cinco gustos básicos (multiclase), constituyéndose así, hasta donde se conoce, en los primeros modelos de clasificación que consideran todos los gustos básicos. Esto se debe a que históricamente ha existido mayor interés por parte de los químicos en la síntesis y estudio de compuestos dulces y amargos, lo que ha provocado que estas clases de compuestos sean predominantes sobre las demás. Este hecho se ve reflejado en la baja representatividad de compuestos umami, ácidos y salados. No obstante esta limitación el modelo ECFPs–N3 brinda una capacidad de predicción del 81.6%. Por otra parte, también se han calibrado modelos binarios al considerar estos dos gustos mayormente representados, es decir, dulce–amargo, dulce–no dulce y amargo–no amargo; donde el clasificador de los RF resulta ser el más apropiado. De hecho, los resultados son comparables con los reportados previamente en la comunidad científica y en algunos casos superan en la capacidad de realizar nuevas predicciones.

Existe aún aspectos por explorar dentro de la línea de investigación propuesta en la presente tesis de maestría, que tentativamente se pueden enfocar en los siguientes aspectos:

1. Profundizar, desde el punto de vista químico, el análisis de los diferentes espacios químicos, particularmente los grupos consistentes de moléculas, de tal forma que se posibilite la identificación de los fragmentos que brindan la similitud entre los compuestos que pertenecen a clases distintas. De esta manera se brindaría una herramienta adicional a los químicos experimentales para optimizar el diseño y síntesis de nuevas moléculas.
2. Utilizar otros métodos de optimización de las geometrías moleculares, basados en métodos semiempíricos y de la mecánica cuántica, de tal forma que se puedan usar los descriptores tridimensionales, los cuales probablemente contribuyen a una mejor descripción de la estructura molecular.

3. Analizar la contribución de los descriptores 3D en la redefinición de los espacios químicos basados en la incrustación de vecinos estocásticos distribuidos en t. Asimismo, se pueden implementar otros métodos del aprendizaje no supervisado para analizar los dominios químicos, por ejemplo, el análisis de componentes principales, el escalado multidimensional y las variantes derivadas de los mismos.
4. Se pueden explorar nuevos métodos del aprendizaje automático para intentar mejorar las predicciones de los modelos de clasificación generados. Por ejemplo, máquinas de soporte vectorial (SVM), redes neuronales de retropropagación (BPNNs), redes neuronales de contrapropagación (CP–ANNs), redes neuronales profundas (DNNs), aprendizaje automático de "impulso adaptativo" (AdaBoost), modelado independiente suave por analogía de clase (SIMCA), Funciones Potenciales, Naive Bayes, entre los más utilizados en las relaciones cuantitativas estructura–actividad.
5. Realizar un análisis de consenso entre las predicciones de diversos clasificadores del aprendizaje automático, de tal forma que las debilidades de un modelo sean compensadas con las fortalezas de los otros y viceversa. Es posible que un enfoque basado en el consenso permita incrementar la capacidad descriptiva y fundamentalmente predictiva de los modelos.
6. Finalmente, utilizar el aprendizaje automático supervisado enfocado en los métodos de regresión para estudiar otras propiedades que se derivan de la base de datos desarrollada, por ejemplo, el dulzor relativo, amargor relativo y umami relativo para los compuestos dulces, amargos y umami.

PUBLICACIONES Y TRABAJOS PRESENTADOS EN CONGRESOS

El presente trabajo de tesis de maestría dio lugar a la publicación de la siguiente base de datos:

- Rojas, C., Ballabio, D., Pacheco Sarmiento, K., Pacheco Jaramillo, E., Mendoza, M., & García, F. (2021). ChemTastesDB: A Curated Database of Molecular Tastants (1.0) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.5747393.

los siguientes artículos científicos:

- Rojas, C., Ballabio, D., Pacheco Sarmiento, K., Pacheco Jaramillo, E., Mendoza, M., & García, F. ChemTastesDB: A Curated Database of Molecular Tastants. (*En revisión*).
- Machine Learning Approaches to Predict the Tastes of Molecular Tastants. (*Manuscrito en preparación*).

así como al siguiente trabajo presentado en un evento científico:

- Pacheco Sarmiento, K., Rojas, C., Tripaldi, P. & Ballabio. D. Predicción *in silico* del gusto de compuestos amargos, dulces e insípidos. II Congreso Internacional de Alimentos, Ciencia y Tecnología, Capítulo Ciencia y Emprendimiento. Quito, Ecuador. Diciembre 2019. (*Premio al mejor póster en la temática de Análisis de Alimentos, Categoría Senior*).

REFERENCIAS

1. Damodaran, S.; Parkin, K.L.; Fennema, O.R. *Fennema's Food Chemistry*, Fourth ed.; CRC Press: USA, **2008**.
2. Wong, D.W. *Mechanism and Theory in Food Chemistry*; Springer: Cham (Switzerland), **2018**.
3. Shallenberger, R.S. *Taste Chemistry*, First ed.; Springer Science+Business Media, B.V: Dordrecht (Netherlands), **1993**.
4. Morini, G.; Bassoli, A.; Borgonovo, G. Molecular Modelling and Models in the Study of Sweet and Umami Taste Receptors. A Review. *Flavour and Fragrance Journal* **2011**, *26*, 254–259.
5. Rojas, C.; Todeschini, R.; Ballabio, D.; Mauri, A.; Consonni, V.; Tripaldi, P.; Grisoni, F. A QSTR–Based Expert System to Predict Sweetness of Molecules. *Frontiers in Chemistry* **2017**, *5*, 1–12.
6. Rojas, C.; Duchowicz, P.R.; Pis Diez, R.; Tripaldi, P. Applications of Quantitative Structure–Relative Sweetness Relationships in Food Chemistry. In *Chemometrics Applications and Research: QSAR in Medicinal Chemistry*, Mercader, A.G., Duchowicz, P.R., Sivakumar, P.M., Eds.; Apple Academic Press: Boca Raton (USA), **2016**; pp. 317–339.
7. Bassoli, A.; Drew, M.G.B.; Hattotuwagama, C.K.; Merlini, L.; Morini, G.; Wilden, G.R.H. Quantitative Structure–Activity Relationships of Sweet Isovanillyl Derivatives. *Quantitative Structure–Activity Relationships* **2001**, *20*, 3–16.
8. Rojas, C.; Duchowicz, P.R. *Química Computacional de los Alimentos: Relaciones Cuantitativas Estructura–Actividad/Propiedad (QSAR/QSPR)*; Editorial Acribia, S.A.: Zaragoza (España), **2021**.
9. Banerjee, P.; Preissner, R. BitterSweetForest: A Random Forest Based Binary Classifier to Predict Bitterness and Sweetness of Chemical Compounds. *Frontiers in Chemistry* **2018**, *6*, 93.

10. Yamaguchi, S. The Umami Taste. In *Food Taste Chemistry*, Boudreau, J.C., Ed.; American Chemical Society: Washington (USA), **1979**; pp. 33–51.
11. Suess, B.; Festring, D.; Hofmann, T. Umami Compounds and Taste Enhancers. In *Flavour Development, Analysis and Perception in Food and Beverages*, Parker, J.K., Elmore, J.S., Methven, L., Eds.; Woodhead Publishing: Cambridge (UK), **2015**; pp. 331–351.
12. Ley, J.; Reichelt, K.; Obst, K.; Krammer, G.; Engel, K.H. Important Tastants and New Developments. In *Food Flavors. Chemical, Sensory and Technological Properties*, Jeleń, H., Ed.; CRC Press: Boca Raton (USA), **2012**; pp. 19–33.
13. Iwamura, H. Structure–Taste Relationship of Perillartine and Nitro– and Cyanoaniline Derivatives. *Journal of Medicinal Chemistry* **1980**, *23*, 308–312.
14. Kier, L.B. Molecular Structure Influencing either a Sweet or Bitter Taste Among Aldoximes. *Journal of Pharmaceutical Sciences* **1980**, *69*, 416–419.
15. Ahmed, J.; Preissner, S.; Dunkel, M.; Worth, C.L.; Eckert, A.; Preissner, R. SuperSweet–A Resource on Natural and Artificial Sweetening Agents. *Nucleic Acids Research* **2011**, *39*, D377–D382.
16. Kim, S.; Chen, J.; Cheng, T.; Gindulyte, A.; He, J.; He, S.; Li, Q.; Shoemaker, B.A.; Thiessen, P.A.; Yu, B. PubChem 2019 Update: Improved Access to Chemical Data. *Nucleic Acids Research* **2019**, *47*, D1102–D1109.
17. Dagan–Wiener, A.; Di Pizio, A.; Nissim, I.; Bahia, M.S.; Dubovski, N.; Margulis, E.; Niv, M.Y. BitterDB: Taste Ligands and Receptors Database in 2019. *Nucleic Acids Research* **2019**, *47*, D1179–D1185.
18. Sterling, T.; Irwin, J.J. ZINC 15–Ligand Discovery for Everyone. *Journal of Chemical Information and Modeling* **2015**, *55*, 2324–2337.
19. Ruddigkeit, L.; Reymond, J.L. The Chemical Space of Flavours. In *Foodinformatics: Applications of Chemical Information to Food Chemistry*, Martinez–Mayorga, K., Medina–Franco, J.L., Eds.; Springer: Cham (Switzerland), **2014**; pp. 83–96.
20. Bouysset, C.; Belloir, C.; Antonczak, S.; Briand, L.; Fiorucci, S. Novel Scaffold of Natural Compound Eliciting Sweet Taste Revealed by Machine Learning. *Food Chemistry* **2020**, *324*, 126864.

21. Medina–Franco, J.L.; Martínez–Mayorga, K.; Giulianotti, M.A.; Houghten, R.A.; Pinilla, C. Visualization of the Chemical Space in Drug Discovery. *Current Computer–Aided Drug Design* **2008**, *4*, 322–333.
22. Zabolotna, Y.; Volochnyuk, D.; Ryabukhin, S.; Horvath, D.; Gavrylenko, K.; Marcou, G.; Moroz, Y.; Oksiuta, O.; Varnek, A. A Close–Up Look at the Chemical Space of Commercially Available Building Blocks for Medicinal Chemistry. *ChemRxiv* **2021**.
23. Dagan–Wiener, A.; Nissim, I.; Ben Abu, N.; Borgonovo, G.; Bassoli, A.; Niv, M.Y. Bitter or not? BitterPredict, A Tool for Predicting Taste from Chemical Structure. *Scientific Reports* **2017**, *7*, 12074.
24. Tuwani, R.; Wadhwa, S.; Bagler, G. BitterSweet: Building Machine Learning Models for Predicting the Bitter and Sweet Taste of Small Molecules. *Scientific Reports* **2019**, *9*, 7155.
25. Hansch, C.; Leo, A. *Exploring QSAR. Fundamentals and Applications in Chemistry and Biology*; American Chemical Society: USA, **1995**.
26. Todeschini, R.; Consonni, V. *Molecular Descriptors for Chemoinformatics*, Second ed.; Wiley–VCH Verlag GmbH & Co: Weinheim (Germany), **2009**.
27. Kubinyi, H. *QSAR: Hansch Analysis and Related Approaches*; VCH: Weinheim (Germany), **2008**; Volume 1.
28. Todeschini, R.; Consonni, V.; Ballabio, D.; Grisoni, F. Chemometrics for QSAR Modeling. In *Comprehensive Chemometrics: Chemical and Biochemical Data Analysis*, Second ed.; Brown, S., Tauler, R., Walczak, B., Eds.; Elsevier: Amsterdam (Netherlands), **2020**; Volume 4, pp. 599–634.
29. Cronin, M.T. Quantitative Structure–Activity Relationships (QSARs)–Applications and Methodology. In *Recent Advances in QSAR Studies: Methods and Applications*, Puzyn, T., Leszczynski, J., Cronin, M.T., Eds.; Springer Science+Business Media B.V.: Dordrecht (Netherlands), **2010**; pp. 3–11.
30. Roy, K.; Kar, S.; Das, R.N. *A Primer on QSAR/QSPR Modeling: Fundamental Concepts*; Springer: Cham (Switzerland), **2015**.
31. Golbraikh, A.; Wang, X.S.; Zhu, H.; Tropsha, A. Predictive QSAR Modeling: Methods and Applications in Drug Discovery and Chemical Risk Assessment. In *Handbook of Computational*

Chemistry, Second ed.; Leszczynski, J., Kaczmarek–Kedziera, A., Puzyn, T., Papadopoulos, M.G., Reis, H., Shukla, M.K., Eds.; Springer International Publishing: Cham (Switzerland), **2017**; Volume 3, pp. 2303–2340.

32. Cuevas, G.; Cortés, F. *Introducción a la Química Computacional*; Fondo de Cultura Económica: México, **2003**.
33. Janežič, D.; Miličević, A.; Nikolić, S.; Trinajstić, N. *Graph–Theoretical Matrices in Chemistry*; CRC Press: Boca Raton (USA), **2015**.
34. Polansky, O.E. Elements of Graph Theory for Chemists. In *Chemical Graph Theory: Introduction and Fundamentals*, Bonchev, D., Rouvray, D.H., Eds.; Abacus Press: Amsterdam (Netherlands), **1991**; Volume 1, pp. 41–96.
35. Bonchev, D. On the Concept for Overall Topological Representation of Molecular Structure. In *Advances in Mathematical Chemistry and Applications*, Basak, S.C., Restrepo, G., Villaveces, J.L., Eds.; Elsevier: Amsterdam (Netherlands), **2015**; pp. 42–75.
36. Garcia, J.; Duchowicz, P.R.; Castro, E.A. Considering the Molecular Conformational Flexibility in QSAR Studies. In *Chemometrics Applications and Research: QSAR in Medicinal Chemistry*, Mercader, A.G., Duchowicz, P.R., Sivakumar, P.M., Eds.; Apple Academic Press: Boca Raton (USA), **2016**; pp. 129–158.
37. Morrison, R.T.; Boyd, R.N.; Bhattacharjee, S.K. *Organic Chemistry*, Seventh ed.; Pearson: India, **2010**.
38. Hypercube Inc. *HyperChem™ Professional version 8,* http://www.hyper.com.
39. ChemAxon Ltd. *MarvinSketch version 20.17,* http://www.chemaxon.com.
40. Engel, T. Computer Processing of Chemical Structure Information. In *Chemoinformatics: Basic Concepts and Methods*, Engel, T., Gasteiger, J., Eds.; WILEY–VCH: Weinheim (Germany), **2018**; pp. 43–119.
41. Polanski, J.; Gasteiger, J. Computer Representation of Chemical Compounds. In *Handbook of Computational Chemistry*, Second ed.; Leszczynski, J., Kaczmarek–Kedziera, A., Puzyn, T., Papadopoulos, M.G., Reis, H., Shukla, M.K., Eds.; Springer International Publishing: Cham (Switzerland), **2017**; Volume 3, pp. 1997–2039.

42. Frau, J.; Sánchez–Marcos, E.; Andrés, J. Modelización Molecular. In *Química Teórica y Computacional*, Andrés, J., Beltrán, J., Eds.; Publicacions de la Universitat Jaume I: Castelló (España), **2000**; pp. 417–532.
43. Jensen, F. *Introduction to Computational Chemistry*, Third ed.; Wiley: Chichester, West Sussex (UK), **2017**.
44. Sippl, W.; Robaa, D. QSAR/QSPR. In *Applied Chemoinformatics: Achievements and Future Opportunities*, Engel, T., Gasteiger, J., Eds.; Wiley–VCH Verlag GmbH & Co. KGaA: Weinheim (Germany), **2018**; pp. 9–52.
45. Gasteiger, J. Introduction. In *Chemoinformatics: A Textbook*, Gasteiger, J., Engel, T., Eds.; WILEY–VCH Verlag GmbH & Co. KgaA: Weinheim (Germany), **2003**; pp. 1–13.
46. Grisoni, F.; Ballabio, D.; Todeschini, R.; Consonni, V. Molecular Descriptors for Structure–Activity Applications: A Hands–On Approach. In *Computational Toxicology: Methods and Protocols*, Nicolotti, O., Ed.; Humana Press: New York (USA), **2018**; pp. 3–53.
47. Guha, R.; Willighagen, E. A Survey of Quantitative Descriptions of Molecular Structure. *Current Topics in Medicinal Chemistry* **2012**, *12*, 1946–1956.
48. Mauri, A.; Consonni, V.; Todeschini, R. Molecular Descriptors. In *Handbook of Computational Chemistry*, Second ed.; Leszczynski, J., Kaczmarek–Kedziera, A., Puzyn, T., Papadopoulos, M.G., Reis, H., Shukla, M.K., Eds.; Springer International Publishing: Cham (Switzerland), **2017**; Volume 3, pp. 2065–2093.
49. Kier, L.B.; Hall, L.H. *Molecular Connectivity in Structure–Activity Analysis*; John Wiley & Sons: USA, **1986**.
50. Burden, F.R. Molecular Identification Number for Substructure Searches. *Journal of Chemical Information and Computer Sciences* **1989**, *29*, 225–227.
51. Labute, P. A Widely Applicable Set of Descriptors. *Journal of Molecular Graphics and Modelling* **2000**, *18*, 464–477.
52. Alvascience *alvaDesc (software for Molecular Descriptors Calculation) version 2.0.10,* https://www.alvascience.com, **2021**.
53. Ghose, A.K.; Viswanadhan, V.N.; Wendoloski, J.J. 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* **1998**, *102*, 3762–3772.

54. Viswanadhan, V.N.; Ghose, A.K.; Revankar, G.R.; Robins, R.K. 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* **1989**, *29*, 163–172.
55. Todeschini, R.; Consonni, V. Descriptors from Molecular Geometry. In *Handbook of Chemoinformatics: From Data to Knowledge in 4 Volumes*, Gasteiger, J., Ed.; WILEY–VCH Verlag GmbH & Co. KGaA: Weinheim (Germany), **2003**; Volume 3, pp. 1004–1033.
56. Schuur, J.H.; Selzer, P.; Gasteiger, J. The Coding of the Three–Dimensional Structure of Molecules by Molecular Transforms and its Application to Structure–Spectra Correlations and Studies of Biological Activity. *Journal of Chemical Information and Computer Sciences* **1996**, *36*, 334–344.
57. Hemmer, M.C.; Steinhauer, V.; Gasteiger, J. Deriving the 3D Structure of Organic Molecules from Their Infrared Spectra. *Vibrational Spectroscopy* **1999**, *19*, 151–164.
58. Todeschini, R.; Gramatica, P. New 3D Molecular Descriptors: the WHIM Theory and QSAR Applications. In *3D QSAR in Drug Design: Ligand–Protein Interactions and Molecular Similarity*, Kubinyi, H., Folkers, G., Martin, Y.C., Eds.; Kluwer Academic Publishers: New York (USA), **2002**; Volume 2, pp. 355–380.
59. Grisoni, F.; Merk, D.; Consonni, V.; Hiss, J.A.; Tagliabue, S.G.; Todeschini, R.; Schneider, G. Scaffold Hopping from Natural Products to Synthetic Mimetics by Holistic Molecular Similarity. *Communications Chemistry* **2018**, *1*, 44.
60. Consonni, V.; Todeschini, R.; Pavan, M. Structure/Response Correlations and Similarity/Diversity Analysis by GETAWAY Descriptors. 1. Theory of the Novel 3D Molecular Descriptors. *Journal of Chemical Information and Computer Sciences* **2002**, *42*, 682–692.
61. Carhart, R.E.; Smith, D.H.; Venkataraghavan, R. Atom Pairs as Molecular Features in Structure–Activity Studies: Definition and

Applications. *Journal of Chemical Information and Computer Sciences* **1985**, *25*, 64–73.

62. Renner, S.; Fechner, U.; Schneider, G. Alignment–Free Pharmacophore Patterns–A Correlation–Vector Approach. In *Pharmacophores and Pharmacophore Searches*, Langer, T., Hoffmann, R.D., Eds.; WILEY–VCH Verlag GmbH & Co. KGaA: Weinheim (Germany), **2006**; pp. 49–79.
63. O'Donnell, T.J. *Design and Use of Relational Databases in Chemistry*; CRC Press: Boca Raton (USA), **2009**.
64. Durant, J.L.; Leland, B.A.; Henry, D.R.; Nourse, J.G. Reoptimization of MDL Keys for Use in Drug Discovery. *Journal of Chemical Information and Computer Sciences* **2002**, *42*, 1273–1280.
65. Bolton, E.E.; Wang, Y.; Thiessen, P.A.; Bryant, S.H. PubChem: Integrated Platform of Small Molecules and Biological Activities. *Annual Reports in Computational Chemistry* **2008**, *4*, 217–241.
66. Mauri, A. alvaDesc: A Tool to Calculate and Analyze Molecular Descriptors and Fingerprints. In *Ecotoxicological QSARs*, Roy, K., Ed.; Humana Press: New York (USA), **2020**; pp. 801–820.
67. Spillane, W.J.; Sheahan, M.B.; Ryder, C.A. Synthesis and Taste Properties of Sodium Disubstituted Phenylsulfamates. Structure–Taste Relationships for Sweet and Bitter/Sweet Sulfamates. *Food Chemistry* **1993**, *47*, 363–369.
68. Bassoli, A.; Laureati, M.; Borgonovo, G.; Morini, G.; Servant, G.; Pagliarini, E. Isovanillic Sweeteners: Sensory Evaluation and in vitro Assays with Human Sweet Taste Receptor. *Chemosensory Perception* **2008**, *1*, 174–183.
69. Kar, S.; Roy, K.; Leszczynski, J. On Applications of QSARs in Food and Aagricultural Sciences: History and Critical Review of Recent Developments. In *Advances in QSAR Modeling: Applications in Pharmaceutical, Chemical, Food, Agricultural and Environmental Sciences*, Roy, K., Ed.; Springer: Cham (Switzerland), **2017**; pp. 203–302.
70. Spillane, W.J.; McGlinchey, G. Structure–Activity Studies on Sulfamate Sweeteners II: Semiquantitative Structure–Taste Relationship for Sulfamate ($RNHSO_3^-$) Sweeteners–The Role of R. *Journal of Pharmaceutical Sciences* **1981**, *70*, 933–935.

71. Takahashi, Y.; Miyashita, Y.; Tanaka, Y.; Abe, H.; Sasaki, S. A Consideration for Structure–Taste Correlations of Perillartines Using Pattern–Recognition Techniques. *Journal of Medicinal Chemistry* **1982**, *25*, 1245–1248.
72. Spillane, W.J.; McGlinchey, G.; Muircheartaigh, I.Ó.; Benson, G.A. Structure–Activity Studies on Sulfamate Sweetners III: Structure–Taste Relationships for Heterosulfamates. *Journal of Pharmaceutical Sciences* **1983**, *72*, 852–856.
73. Takahashi, Y.; Abe, H.; Miyashita, Y.; Tanaka, Y.; Hayasaka, H.; Sasaki, S.I. Discriminative Structural Analysis Using Pattern Recognition Techniques in the Structure–Taste Problem of Perillartines. *Journal of Pharmaceutical Sciences* **1984**, *73*, 737–741.
74. Miyashita, Y.; Takahashi, Y.; Takayama, C.; Sumi, K.; Nakatsuka, K.; Ohkubo, T.; Abe, H.; Sasaki, S.i. Structure–Taste Correlation of L–Aspartyl Dipeptides Using the SIMCA Method. *Journal of Medicinal Chemistry* **1986**, *29*, 906–912.
75. Miyashita, Y.; Takahashi, Y.; Takayama, C.; Ohkubo, T.; Funatsu, K.; Sasaki, S.i. Computer–Assisted Structure/Taste Studies on Sulfamates by Pattern Recognition Methods. *Analytica Chimica Acta* **1986**, *184*, 143–149.
76. Okuyama, T.; Miyashita, Y.; Kanaya, S.; Katsumi, H.; Sasaki, S.i.; Randić, M. Computer Assisted Structure–Taste Studies on Sulfamates by Pattern Recognition Method Using Graph Theoretical Invariants. *Journal of Computational Chemistry* **1988**, *9*, 636–646.
77. Spillane, W.J.; Sheahan, M.B. Semi–Quantitative and Quantitative Structure–Taste Relationships for Carboand Hetero–Sulphamate ($RNHSO_3^-$) Sweeteners. *Journal of the Chemical Society, Perkin Transactions 2* **1989**, 741–746.
78. Spillane, W.J.; Sheahan, M. Structure–Taste Relationships for Sulfamate Sweeteners ($RNHSO_3^-$). *Phosphorus, Sulfur, and Silicon and the Related Elements* **1991**, *59*, 255–258.
79. Drew, M.G.B.; Wilden, G.R.H.; Spillane, W.J.; Walsh, R.M.; Ryder, C.A.; Simmie, J.M. Quantitative Structure–Activity Relationship Studies of Sulfamates $RNHSO_3Na$: Distinction between Sweet, Sweet–Bitter, and Bitter Molecules. *Journal of Agricultural and Food Chemistry* **1998**, *46*, 3016–3026.

80. Spillane, W.J.; Ryder, C.A.; Curran, P.J.; Wall, S.N.; Kelly, L.M.; Feeney, B.G.; Newell, J. Development of Structure–Taste Relationships for Sweet and Non–Sweet Heterosulfamates. *Journal of the Chemical Society, Perkin Transactions 2* **2000**, 1369–1374.
81. Spillane, W.J.; Feeney, B.G.; Coyle, C.M. Further Studies on the Synthesis and Tastes of Monosubstituted Benzenesulfamates. A Semi–Quantitative Structure–Taste Telationship for the meta–Compounds. *Food Chemistry* **2002**, *79*, 15–22.
82. Spillane, W.J.; Kelly, L.M.; Feeney, B.G.; Drew, M.G.; Hattotuwagama, C.K. Synthesis of Heterosulfamates. Search for Structure–Taste Relationships. *Arkivoc* **2003**, *7*, 297–309.
83. Kelly, D.P.; Spillane, W.J.; Newell, J. Development of Structure–Taste Relationships for Monosubstituted Phenylsulfamate Sweeteners Using Classification and Regression Tree (CART) Analysis. *Journal of Agricultural and Food Chemistry* **2005**, *53*, 6750–6758.
84. Rodgers, S.; Glen, R.C.; Bender, A. Characterizing Bitterness: Identification of Key Structural Features and Development of a Classification Model. *Journal of Chemical Information and Modeling* **2006**, *46*, 569–576.
85. Spillane, W.J.; Kelly, D.P.; Curran, P.J.; Feeney, B.G. Structure–Taste Relationships for Disubstituted Phenylsulfamate Tastants Using Classification and Regression Tree (CART) Analysis. *Journal of Agricultural and Food Chemistry* **2006**, *54*, 5996–6004.
86. Spillane, W.J.; Coyle, C.M.; Feeney, B.G.; Thompson, E.F. Development of Structure–Taste Relationships for Thiazolyl–, Benzothiazolyl–, and Thiadiazolylsulfamates. *Journal of Agricultural and Food Chemistry* **2009**, *57*, 5486–5493.
87. Rojas, C.; Ballabio, D.; Consonni, V.; Tripaldi, P.; Mauri, A.; Todeschini, R. Quantitative Structure–Activity Relationships to Predict Sweet and Non–Sweet Tastes. *Theoretical Chemistry Accounts* **2016**, *135:66*, 1–13.
88. Huang, W.; Shen, Q.; Su, X.; Ji, M.; Liu, X.; Chen, Y.; Lu, S.; Zhuang, H.; Zhang, J. BitterX: A Tool for Understanding Bitter Taste in Humans. *Scientific Reports* **2016**, *6*, 23450.

89. Chéron, J.B.; Casciuc, I.; Golebiowski, J.; Antonczak, S.; Fiorucci, S. Sweetness Prediction of Natural Compounds. *Food Chemistry* **2017**, *221*, 1421–1425.
90. Zheng, S.; Jiang, M.; Zhao, C.; Zhu, R.; Hu, Z.; Xu, Y.; Lin, F. e–Bitter: Bitterant Prediction by the Consensus Voting From the Machine–Learning Methods. *Frontiers in Chemistry* **2018**, *6*, 82.
91. Zheng, S.; Chang, W.; Xu, W.; Xu, Y.; Lin, F. e–Sweet: A Machine–Learning Based Platform for the Prediction of Sweetener and Its Relative Sweetness. *Frontiers in Chemistry* **2019**, *7*, 35.
92. Charoenkwan, P.; Nantasenamat, C.; Hasan, M.M.; Moni, M.A.; Lio', P.; Shoombuatong, W. iBitter–Fuse: A Novel Sequence–Based Bitter Peptide Predictor by Fusing Multi–View Features. *International Journal of Molecular Sciences* **2021**, *22*, 8958.
93. Shalev–Shwartz, S.; Ben–David, S. *Understanding Machine Learning: From Theory to Algorithms*; Cambridge University press: New York (USA), **2014**.
94. Kubat, M. *An Introduction to Machine Learning*, Second ed.; Springer: Cham (Switzerland), **2017**.
95. Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Second ed.; Springer: New York (USA), **2017**.
96. James, G.; Witten, D.; Hastie, T.; Tibshirani, R. *An Introduction to Statistical Learning: with Applications in R*, Second ed.; Spinger: New York (USA), **2021**.
97. Bro, R.; Smilde, A.K. Principal Component Analysis. *Analytical Methods* **2014**, *6*, 2812–2831.
98. Jolliffe, I.T. *Principal Component Analysis*, Second ed.; Springer–Verlag: New York (USA), **2002**.
99. Wold, S.; Esbensen, K.; Geladi, P. Principal Component Analysis. *Chemometrics and Intelligent Laboratory Systems* **1987**, *2*, 37–52.
100. Kruskal, J.B. Multidimensional Scaling by Optimizing Goodness of Fit to a Nonmetric Hypothesis. *Psychometrika* **1964**, *29*, 1–27.
101. Winsberg, S.; Carroll, J.D. A Quasi–Nonmetric Method for Multidimensional Scaling Via an Extended Euclidean Model. *Psychometrika* **1989**, *54*, 217–229.

102. Hinton, G.; Roweis, S.T. Stochastic Neighbor Embedding. In Proceedings of the Proceedings of the 15th International Conference on Neural Information Processing Systems, **2002**; pp. 833–840.
103. van der Maaten, L.; Hinton, G. Visualizing Data Using t–SNE. *Journal of Machine Learning Research* **2008**, *9*, 2579–2605.
104. Linderman, G.C.; Steinerberger, S. Clustering with t–SNE, provably. *SIAM Journal on Mathematics of Data Science* **2019**, *1*, 313–332.
105. Ballabio, D.; Consonni, V.; Mauri, A.; Claeys–Bruno, M.; Sergent, M.; Todeschini, R. A Novel Variable Reduction Method Adapted from Space–Filling Designs. *Chemometrics and Intelligent Laboratory Systems* **2014**, *136*, 147–154.
106. Santiago, J.; Claeys–Bruno, M.; Sergent, M. Construction of Space–Filling Designs using WSP Algorithm for High Dimensional Spaces. *Chemometrics and Intelligent Laboratory Systems* **2012**, *113*, 26–31.
107. Breiman, L.; Friedman, J.; Olshen, R.A.; Stone, C.J. *Classification and Regression Trees*; Chapman & Hall/CRC: USA, 1984.
108. Breiman, L. Random Forests. *Machine Learning* **2001**, *45*, 5–32.
109. James, G.M. Majority Vote Classifiers: Theory and Applications. Stanford University, **1998**.
110. Baurin, N.; Mozziconacci, J.C.; Arnoult, E.; Chavatte, P.; Marot, C.; Morin–Allory, L. 2D QSAR Consensus Prediction for High–Throughput Virtual Screening. An Application to COX–2 Inhibition Modeling and Screening of the NCI Database. *Journal of Chemical Information and Computer Sciences* **2004**, *44*, 276–285.
111. Cover, T.; Hart, P. Nearest Neighbor Pattern Classification. *IEEE Transactions on Information Theory* **1967**, *13*, 21–27.
112. Todeschini, R.; Ballabio, D.; Cassotti, M.; Consonni, V. N3 and BNN: Two New Similarity Based Classification Methods in Comparison with Other Classifiers. *Journal of Chemical Information and Modeling* **2015**, *55*, 2365–2374.
113. Varmuza, K. Methods for Multivariate Data Analysis. In *Chemoinformatics: Basic Concepts and Methods*, Engel, T., Gasteiger, J., Eds.; Wiley–VCH Verlag GmbH & Co. KGaA: Weinheim (Germany), **2018**; pp. 399–437.
114. Leardi, R. Genetic Algorithms in Chemistry. In *Comprehensive Chemometrics: Chemical and Biochemical Data Analysis*, Second ed.;

Brown, S., Tauler, R., Walczak, B., Eds.; Elsevier: Amsterdam (Netherlands), **2020**; Volume 1, pp. 617–634.

115. Ballabio, D.; Grisoni, F.; Todeschini, R. Multivariate Comparison of Classification Performance Measures. *Chemometrics and Intelligent Laboratory Systems* **2018**, *174*, 33–44.
116. Varmuza, K.; Filzmoser, P. *Introduction to Multivariate Statistical Analysis in Chemometrics*; CRC press: Boca Raton (USA), **2009**.
117. Hawkins, D.M. The Problem of Overfitting. *Journal of Chemical Information and Computer Sciences* **2004**, *44*, 1–12.
118. Arlot, S.; Celisse, A. A Survey of Cross–Validation Procedures for Model Selection. *Statistics Surveys* **2010**, *4*, 40–79.
119. Ballabio, D.; Consonni, V. Classification Tools in Chemistry. Part 1: Linear Models. PLS–DA. *Analytical Methods* **2013**, *5*, 3790–3798.
120. Krakowska, B.; Custers, D.; Deconinck, E.; Daszykowski, M. 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* **2016**, *141*, 1060–1070.
121. Efron, B. *The Jackknife, the Bootstrap and Other Resampling Plans*; Society for Industrial and Applied Mathematics: USA, **1982**.
122. Hongmao, S. *A Practical Guide to Rational Drug Design*; Elsevier: Cambridge (UK), **2016**.
123. Suess, B.; Festring, D.; Hofmann, T. Umami compounds and taste enhancers. In *Flavour development, analysis and perception in food and beverages*, Parker, J.K., Elmore, J.S., Methven, L., Eds.; Woodhead Publishing: **2015**; pp. 331–351.
124. Fourches, D.; Muratov, E.; Tropsha, A. Trust, but Verify: on the Importance of Chemical Structure Curation in Cheminformatics and QSAR Modeling Research. *Journal of Chemical Information and Modeling* **2010**, *50*, 1189–1204.
125. Alvascience *alvaMolecule (software to View and Prepare Chemical Datasets) version 1.0.4,* https://www.alvascience.com, **2020**.
126. Berthold, M.R.; Cebron, N.; Dill, F.; Gabriel, T.R.; Kötter, T.; Meinl, T.; Ohl, P.; Sieb, C.; Thiel, K.; Wiswedel, B. KNIME: The Konstanz Information Miner. In *Data Analysis, Machine Learning and Applications*, Preisach, C., Burkhardt, H., Schmidt–Thieme, L., Decker, R., Eds.; Studies in Classification, Data Analysis, and

Knowledge Organization; Springer: Heidelberg (Germany), 2008; pp. 319–326.

127. Todeschini, R.; Ballabio, D.; Consonni, V. Distances and Other Dissimilarity Measures in Chemometrics. In *Encyclopedia of Analytical Chemistry: Applications, Theory and Instrumentation*, Meyers, R.A., Ed.; John Wiley & Sons, Ltd: 2015; pp. 1–34.

128. The MathWorks Inc. *MATLAB,* http://www.mathworks.com.