



Facultad de Ciencia y Tecnología

Ingeniería en Alimentos

**Desarrollo de una base de datos de gustos de diversos
compuestos para estudios *in silico* en química de los
alimentos**

**Trabajo de graduación previo a la obtención del título de:
INGENIERA EN ALIMENTOS**

Autor:

Karen Elizabeth Pacheco Sarmiento

Director:

Dr. Cristian Rojas Villa

Codirector:

Dr. Davide Ballabio

CUENCA, ECUADOR

2021

DEDICATORIA

Mi trabajo de titulación lo dedico principalmente a Dios, por darme vida para culminar con esta meta. Y a mi familia, por siempre apoyarme y creer en mí.

AGRADECIMIENTOS

Agradezco a la Universidad del Azuay, por abrirme las puertas y permitirme crecer profesionalmente; por dejarme ser parte de experiencias estudiantiles inolvidables.

A los profesores de la escuela de Ingeniería en Alimentos, por todas las enseñanzas a lo largo de la carrera. Agradezco también al Dr. Davide Ballabio, y de manera especial al Dr. Cristian Rojas Villa por el trabajo, paciencia y dedicación durante todo este trayecto.

A mi familia, por ser parte del camino de mi formación personal y profesional, por siempre apoyarme y darme su cariño, por saber estar para mí de una u otra manera en todos los aspectos de mi vida.

A Mario, porque de inicio a fin ha estado para los momentos buenos y malos, ha tenido una palabra de aliento, me ha entregado todo su apoyo, ha creído en mis capacidades y me ha entendido cuando más lo he necesitado. Gracias por ser un pilar fundamental e incondicional en mi vida.

A mis grandes amigos del colegio: Cris, Carlitos, Jois, Pao A, Gracia, Michelle, Dome, Pao V. gracias por su amistad tan sincera durante todos estos años, por siempre estar pendiente de mí, sacándome las mejores risas en mis momentos de estrés, por su apoyo y cariño.

Finalmente, pero no menos importante, quiero agradecerle a Josselyn por haber compartido tantas experiencias durante la carrera, porque a pesar de lo bueno y lo malo siempre nos hemos apoyado. Y a Elisa, por su ayuda y respaldo tan importante durante el desarrollo de mi tesis.

Son algunas personas que me falta por mencionar, pero a todas ellas quiero agradecerles por haber sido parte de este trayecto, y por haber aportado algo en mi vida personal como profesional.

Desarrollo de una base de datos de gustos de diversos compuestos para estudios *in silico* en química de los alimentos

RESUMEN

Esta investigación describe el desarrollo de una base de datos de gustos para compuestos orgánicos e inorgánicos. Inicialmente, se recopilaban 4484 estructuras de diversas fuentes bibliográficas, las cuales incluyen los cinco gustos principales: dulce, amargo, ácido, salado y umami; así como otras moléculas etiquetadas como insípidas, no dulce y multi gusto. Las moléculas fueron modeladas en el programa HyperChem y sus estructuras se estandarizaron en alvaMolecule. Entonces, se usó el programa KNIME para curar la base de datos y controlar la consistencia del nombre, número CAS, PubChem CID o notación SMILES de los compuestos, así como dividir la base de datos de acuerdo a las diversas clases de gustos. Luego del curado se retuvieron 3589 compuestos, los cuales se usaron para analizar el espacio químico mediante las huellas dactilares moleculares de conectividad ampliada, que son descriptores moleculares binarios que codifican la estructura molecular. Las similitudes entre los compuestos se calcularon a través de la distancia de Jaccard-Tanimoto y se utilizaron para proyectar cada clase en las coordenadas más adecuadas del escalado multidimensional.

Palabras clave: base de datos, gustos básicos, receptor del gusto, ECFP, espacio químico, escalado multidimensional.



Dr. Cristian Rojas Villa
Director de tesis



Dr. Davide Ballabio
Codirector de tesis



Mgtr. María Fernanda Rosales
**Coordinadora de la Escuela Ingeniería
en Alimentos**



Karen Elizabeth Pacheco Sarmiento
Autora

Development of a database of tastes of diverse compounds for in silico studies in food chemistry

ABSTRACT

This research describes the development of a taste database for both organic and inorganic chemicals. Initially, 4484 compounds were collected from several external sources, which cover the five basic classes of tastes: sweet, bitter, sour, salty and umami, as well as other molecules labeled as tasteless, non-sweet and multitaste. Molecules were designed and optimized in the HyperChem software and their structures were then standardized in the alvaMolecule program. Therefore, the KNIME software was used to curate the database and check name consistency, CAS number, PubChem CID or SMILES notation of the collected chemicals, as well as to split the database according to the taste classes. After the database curation, 3589 molecules were retained in the database and were used to analyze the chemical space by means of their extended-connectivity fingerprints, which are binary molecular descriptors encoding the structure of chemicals. Similarities between chemicals were calculated through the Jaccard-Tanimoto binary distance and used to project each class in suitable coordinates of the corresponding multidimensional scaling.

Keywords: database, basic tastes, taste receptor, ECFPs, chemical space, multidimensional scaling



Dr. Cristian Rojas Villa
Director de tesis



Dr. Davide Ballabio
Codirector de tesis



Mgtr. María Fernanda Rosales
**Coordinadora de la Escuela Ingeniería
en Alimentos**



Karen Elizabeth Pacheco Sarmiento
Autora



Translated by



Karen Pacheco

INDICE DE CONTENIDO

DEDICATORIA	ii
AGRADECIMIENTOS	iii
RESUMEN	iv
ABSTRACT	v
INDICE DE CONTENIDO	vi
INDICE DE FIGURAS	viii
INDICE DE TABLAS	ix
INTRODUCCIÓN	1
Aspectos generales sobre el gusto	1
Definición de diversos gustos	2
Gusto dulce	2
Gusto amargo	3
Gusto ácido	4
Gusto salado	4
Gusto umami	5
Insípido	6
Análisis sensorial	7
Bases de datos de gustos de moléculas	8
SuperSweet	8
BitterDB	9
TasteDB	9
TastesDB	9
SweetenersDB	10
BitterSweetForest	10
BitterSweet	11
Espacio químico basado en similitud molecular	11
CAPÍTULO I	13
MATERIALES Y MÉTODOS	13
1.1. Generación de la base de datos	13
1.2. Diseño y representación de las estructuras moleculares	13

Formas de representación de estructura molecular	13
1.3. Curado de las estructuras moleculares	14
1.3.1 alvaMolecule	15
1.4. Fusión y filtrado de las bases de datos.....	15
1.5. Espacio químico del gusto	16
1.5.1. Huellas dactilares moleculares de conectividad ampliada	16
Generación de patrones.....	16
Longitud del patrón	16
Función Hash	17
alvaDesc.....	17
1.5.2. Escalado multidimensional	17
Medidas de distancia para datos binarios.....	18
PCA toolbox para MATLAB.....	20
CAPÍTULO II	21
RESULTADOS Y DISCUSIÓN	21
2.1. Generación de la base de datos	21
2.2. Diseño y representación de las estructuras moleculares	21
2.3. Curado de las estructuras moleculares	21
2.4. Fusión y filtrado de las bases de datos.....	22
2.5. Huellas dactilares moleculares de conectividad ampliada	25
2.6. Espacio químico	25
2.6.1. Espacio químico para el gusto dulce	25
2.6.2. Espacio químico para el gusto amargo.....	27
2.6.3. Espacio químico para el gusto umami	28
2.6.4. Espacio químico para el gusto ácido	30
2.6.5. Espacio químico para el gusto salado	31
2.6.6. Espacio químico para la clase insípida.....	32
2.6.7. Otros espacios químicos.....	34
CONCLUSIONES	37
REFERENCIAS BIBLIOGRÁFICAS	38

INDICE DE FIGURAS

Figura 1. Percepción de los distintos gustos en la lengua.....	2
Figura 2. Representación esquemática del receptor de dulzor humano T1R2 y T1R3.....	3
Figura 3. Representación esquemática de una célula receptora del gusto umami.	6
Figura 4. Diagrama de flujo KNIME para el filtrado de la base de datos.	24
Figura 5. a) Espacio químico definido mediante el MDS para el gusto dulce. b) Proyección de las moléculas dulces y no dulces en el espacio químico.....	26
Figura 6. a) Espacio químico definido mediante el MDS para el gusto amargo. b) Proyección de las moléculas amargo y no amargo en el espacio químico.	27
Figura 7. a) Espacio químico definido mediante el MDS para el gusto umami. b) Proyección de las moléculas umami y no umami en el espacio químico.....	29
Figura 8. a) Espacio químico definido mediante el MDS para el gusto ácido. b) Proyección de las moléculas del gusto ácido y no ácido en el espacio químico.	30
Figura 9. a) Espacio químico definido mediante el MDS para el gusto salado. b) Proyección de las moléculas de gusto salado y no salado en el espacio químico.	32
Figura 10. a) Espacio químico definido mediante el MDS para la clase insípida. b) Proyección de las moléculas de clase insípida y no insípida en el espacio químico.	33
Figura 11. Espacio químico para los compuestos multi gusto definido por las puntuaciones del MDS de las dimensiones 2 y 4.....	35
Figura 12. Espacio químico para los compuestos no dulces definido por las puntuaciones del MDS de las dimensiones 2 y 3.....	36
Figura 13. Espacio químico para los compuestos que presentan otro gusto diferente a los primarios definido por las puntuaciones del MDS de las dimensiones 3 y 5.....	36

INDICE DE TABLAS

Tabla 1. Gustos básicos con sus respectivos receptores y clase de gusto.	6
Tabla 2. Estándares para los cinco gustos básicos.	7
Tabla 3. Medidas de distancia para datos binarios.	19
Tabla 4. Tabla de frecuencias de las cuatro combinaciones posibles de valores 0 y 1 para muestras con variables binarias.	19

Karen Elizabeth Pacheco Sarmiento

Trabajo de Titulación

Dr. Cristian Rojas Villa

Dr. Davide Ballabio

Marzo, 2021

Desarrollo de una base de datos de gustos de diversos compuestos para estudios *in silico* en química de los alimentos

INTRODUCCIÓN

Aspectos generales sobre el gusto

Dentro de la química de los alimentos, el gusto se ha convertido en un campo de investigación que ha llamado la atención de diversos investigadores alrededor del mundo. El gusto es la combinación de sensaciones químicas percibidas por las papilas gustativas de la lengua y se asocia a compuestos solubles con diferentes propiedades osmóticas, endotérmicas y exotérmicas, los cuales interactúan con las membranas biológicas (receptores moleculares) ubicadas en las papilas gustativas. Sin embargo, muchas de las sustancias no tienen un sabor único, sino una sensación compleja que comprende más de uno de los cinco gustos básicos (Wong, 2018). Los gustos se clasifican en cinco grupos básicos: dulce, amargo, salado, ácido y umami (Damodaran, Parkin, & Fennema, 2008). La apreciación del gusto no es igual en todas las personas, esto se debe a varios factores internos tanto del alimento como de quien lo consume; por ejemplo, la concentración del compuesto, psicología y anatomía del humano, entre otros (Shallenberger, 1993). En el proceso de ingestión y masticación de los alimentos, las sustancias químicas de los alimentos entran en los poros de las papilas gustativas en la lengua, donde interactúan con las células epiteliales especializadas, conocidas como células receptoras de gusto (TRCs, taste receptor cells). Cada papila gustativa tiene alrededor de 50-150 células receptoras del gusto, cada una de ellas contiene quimiorreceptores que detectan una de las cinco sensaciones gustativas. De esta manera, cuando se produce la unión de un constituyente químico alimentario a un receptor específico se desencadena una reacción que envía señales a la médula del cerebro (Wong, 2018).

En la Figura 1 se muestra el lugar de percepción de cada gusto en la lengua. Se observa que el gusto dulce se percibe en la punta de la lengua, el gusto amargo en la parte posterior y los gustos ácido y salado en los laterales (Manfugás, 2020).

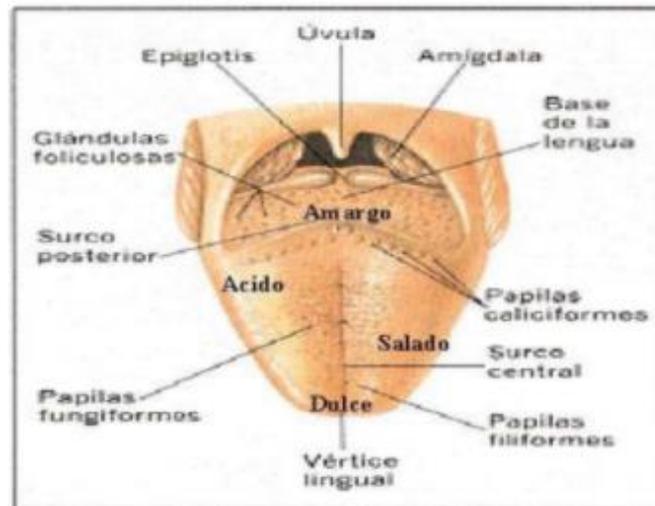


Figura 1. Percepción de los distintos gustos en la lengua (Manfugás, 2020).

Definición de diversos gustos

Gusto dulce

Cuando se habla de los distintos gustos que las personas pueden percibir, el gusto dulce resulta el más significativo, debido a que produce una sensación muy agradable en diversos alimentos. La sacarosa es utilizada como estándar internacional para definir el dulzor de otras moléculas debido a que, aún en altas concentraciones, evoca un gusto dulce limpio y no produce retrogustos (Damodaran et al., 2008). Así, el dulzor relativo resulta de la relación entre la concentración de una solución de sacarosa (estándar de dulzor) y la concentración de otro endulzante; es decir, el estándar tiene un valor de 1 (o 100) y el dulzor del edulcorante será comparado con el de la sacarosa (Bassoli et al., 2001).

Una de las características particulares del sistema receptor del gusto dulce, es su capacidad para reconocer moléculas que pertenecen a clases de compuestos muy diferentes, que van desde azúcares hasta aminoácidos, péptidos, proteínas y otras clases de compuestos orgánicos (Morini, Bassoli, & Borgonovo, 2011). El receptor del dulzor consta de un heterodímero de dos subunidades relacionadas con la secuencia que pertenece a la familia de receptores acoplados a proteínas G (GPCR, Protein-Coupled Receptor) de clase C. Así, existe el receptor 2 tipo 1 (T1R2, Type 1 Receptor 2) y receptor 3 tipo 1 (T1R3, Type 1 Receptor 3), donde cada subunidad está formada por tres dominios estructurales que pueden ser identificados en la Figura 2: un dominio extracelular amino-terminal parecido a una trampa para moscas de Venus (VFTD, Venus flytrap domain), formado por los lóbulos 1 y 2; un dominio transmembrana heptahelical (TMD, heptahelical transmembrane) y un dominio rico en cisteína (CRD, cysteine-rich domain) que conecta los dos dominios anteriormente mencionados. Los lóbulos de T1R2 y T1R3 cambian su disposición de forma flexible para crear una conformación “abierta” o “cerrada”. Existen dos formas activas cerrado-abierto:

cuando T1R2 está abierto y T1R3 cerrado o viceversa. Por el contrario, existe inactividad cuando ambos lóbulos están abiertos. Cuando un edulcorante de bajo peso molecular se une al receptor, este se activa al cerrar los dos lóbulos, transformando la forma inactiva en forma activa. El aspartamo, sacarina sódica, acesulfamo K, ciclamato y sucralosa son reconocidos en la hendidura de los lóbulos 1 y 2 de T1R2, provocando interacciones de unión multipunto. Por otro lado, las proteínas dulces parecen unirse a un sitio de la superficie secundaria pero sin entrar en la hendidura del lóbulo 1-2 que pertenece principalmente al receptor T1R3 (Wong, 2018).

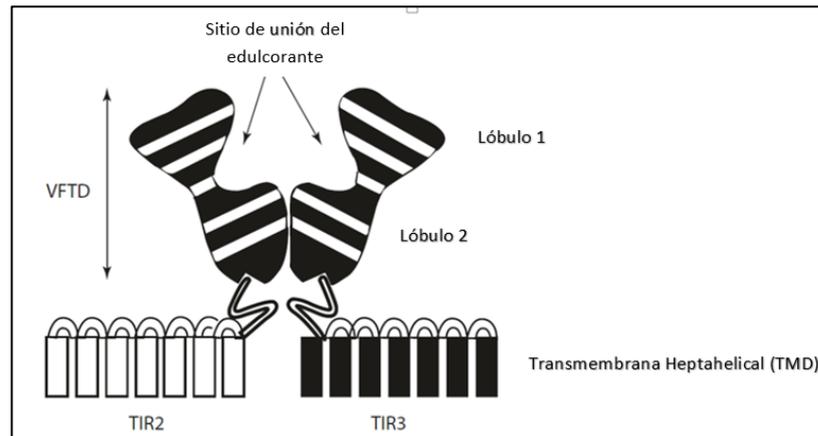


Figura 2. Representación esquemática del receptor de dulzor humano T1R2 y T1R3 (Wong, 2018).

Gusto amargo

El gusto amargo proviene principalmente de los alcaloides (por ejemplo, la quinina) y generalmente se percibe como un gusto desagradable en algunos alimentos. No obstante, en productos como el café, cerveza, aceitunas y otros, la presencia de este gusto es fundamental (placentero) para definir la identidad de los mismos (Damodaran et al., 2008). Por otro lado, el sabor amargo de un alimento o de una sustancia está vinculado a que el hombre se proteja contra el consumo de alcaloides perjudiciales (venenos). Sin embargo, no todos los compuestos amargos son tóxicos, pues existen fito nutrientes que se encuentran comúnmente en frutas y verduras, así como fármacos a base de hierbas que provocan beneficio en el cuerpo humano (Banerjee & Preissner, 2018).

Para los seres humanos existen 25 miembros, denominados T2R, de la familia de receptores acoplados a la proteína G (GPCR), que captan el sabor amargo y tienen secuencias de entre 291 y 334 aminoácidos. Un único compuesto amargo es capaz de activar múltiples receptores T2R, pero también cada unidad T2R puede ser activada por varias moléculas amargas. Los receptores T2R constan de siete transmembranas, tres bucles extracelulares (ECLs, three extracellular loops) y tres bucles intracelulares (ICLs, intracellular loops), con un extremo N extracelular corto y un extremo C intracelular. La vía de transducción de la señal canónica T2R empieza con la unión de un compuesto amargo (agonista) y la superficie extracelular de

un T2R, lo cual genera cambios conformacionales en dicho receptor, y a su vez se activa el complejo de proteína G heterotrimérica (α -gustducina, β 1/3 y γ 13) en la superficie intracelular del receptor. Las subunidades $\beta\gamma$ activan la enzima fosfolipasa C β 2 (PLC β 2) que hidroliza el fosfolípido de inositol (PIP2) dando lugar a la producción de 1, 4,5-trifosfato (IP3) y diacilglicerol (DAG). Al generarse IP3 se activan los receptores IP3 en la membrana del retículo endoplásmico abriendo los canales de liberación del calcio, causando un aumento transitorio en el calcio intracelular. Este proceso abre los canales potenciales de receptor transitorio de melastatina subtipo 5 TRPM5 (Transient receptor potential melastatin subtype 5 channel) selectivos monovalentes, provocando una afluencia de sodio, despolarización de la membrana y liberación de ATP como neurotransmisor para activar las aferencias gustativas. La G α -gustducina activa las fosfodiesterasas (PDEs, phosphodiesterases) que provocan una disminución en la producción de adenosín monofosfato cíclico (cAMP, cyclic Adenosine MonoPhosphate) (Upadhyaya, Singh, Bhullar, & Chelikani, 2017).

Gusto ácido

La característica común de todos los ácidos es la disociación de protones (H^+) cuando se disuelven en agua. Es por eso que además de los iones de hidrógeno, los aniones y/o especies ácidas no disociadas son determinantes en la intensidad del sabor ácido (da Conceicao Neta, Johanningsmeier, Drake, & McFeeters, 2007). La acidez es el gusto más simple entre los existentes porque el ion hidrógeno es la única especie química que causa un gusto ácido, es así que el sabor ácido está en función de la concentración potencial de iones hidrógeno y no solo del ion hidrógeno (pH) per se. Datos basados en algunas publicaciones proponen una nueva hipótesis para la base química de la percepción del gusto ácido: la intensidad de la percepción del sabor ácido en soluciones ácidas o alimentos acidificados está linealmente relacionada con la concentración molar de todas las especies de ácidos orgánicos con al menos un grupo carboxilo protonado más la concentración molar de iones de hidrógenos libres; es decir, un ácido orgánico será igual de ácido solo si al menos un grupo carboxilo es protonado (Johanningsmeiner, McFeeters, & Drake, 2005).

Gusto salado

El cloruro de sodio (NaCl) proporciona un sabor salado clásico y limpio, pero también lo proporciona el cloruro de litio (LiCl), aunque no se lo puede utilizar en alimentos por contener propiedades tóxicas. Se ha establecido que los cationes son los responsables de producir el gusto salado básico, pero los aniones modifican dicho gusto. Los cationes de sodio y litio producen la sensación de salinidad, mientras que el potasio y otros cationes alcalinos evocan gustos salados y amargos. El gusto salado es modificado por acción de los cationes que con frecuencia aportan su propio gusto, aunque el anión cloruro es el que menos inhibe el gusto salado del compuesto o alimento (Damodaran et al., 2008).

Gusto umami

El gusto umami fue descrito por primera vez por Ikeda en el año de 1908, y ha sido estudiado por un gran número de investigadores desde diferentes puntos de vista. La mayoría de las sustancias umami se dividen en dos series de compuestos. El primer grupo está conformado por L- α -aminoácidos representados por el glutamato mono sódico (MSG, monosodium glutamate), y un segundo grupo que incluye a los 5'-ribonucleótidos y sus derivados, representados por el 5'-inosinato disódico (IMP, inosine monophosphate) o 5'-guanilato disódico (GMP, guanosine monophosphate) (Yamaguchi, 1979). Umami es el término usado para identificar el gusto de diversas sustancias, tales como las sales del L-glutamato. El L-glutamato y los 5'-ribonucleótidos poseen un sinergismo que mejora el gusto entre dos compuestos umami, además de poseer un regusto prolongado. El umami desempeña un papel predominante en el gusto de alimentos como carnes, aves, pescado y otros mariscos, productos lácteos o verduras (Yamaguchi & Ninomiya, 2000). Las sustancias umami aportan un gusto “delicioso” a los alimentos cuando se usan en niveles superiores al de su umbral de detección, y modifican o mejoran el gusto cuando se utilizan a niveles inferiores (Johanningsmeiner et al., 2005).

Al hablar de los receptores del gusto umami es necesario mencionar que el compuesto inosina monofosfato (IMP, inosine monophosphate) por sí solo no activa los receptores T1R1/T1R3, pero potencia en gran medida la respuesta de T1R1/T1R3 para el L-glutamato. Por otro lado, varios experimentos en modelos transgénicos sugieren que la detección del gusto umami se basa en sistemas de múltiples receptores que comprenden T1R1/T1R3, “taste-mGluR4” y el “truncate-mGluR1”. Así, la señal mediada por la vía que involucra los receptores T1R1/T1R3 puede realizar un papel diferente al derivado de los mGluR. En la Figura 3 se indica de manera más detallada la conformación de una célula receptora del gusto umami. El primer receptor ocurre principalmente en la parte de adelante de la lengua; por el contrario, el último receptor tiene lugar en la parte posterior de la lengua y ayuda a diferenciar entre umami y otros gustos. De este modo, el heterodímero T1R1/T1R3 está acoplado a una proteína G heterotrimerica que tiene una subunidad $G\alpha$ que modula los niveles de cAMP y una subunidad $G\beta\gamma$ que estimula la fosfolipasa C, produciendo de esta manera la movilización del calcio mediado por el trifosfato de inositol, modulación del canal catiónico y despolarización de la membrana. Haciendo referencia a mGluR, al activarse mGluR1 se produce la hiperpolarización de las células receptoras del gusto (TCRs, taste receptor cells) mediante la movilización del calcio intracelular y un aumento de conductancia del anión Cl^- ; no obstante, la hiperpolarización causada por mGluR4 se produce por la disminución de AMPc en niveles intracelulares provocando el cierre de canales cíclicos regulados por nucleótidos (Ghirri & Bignetti, 2012).

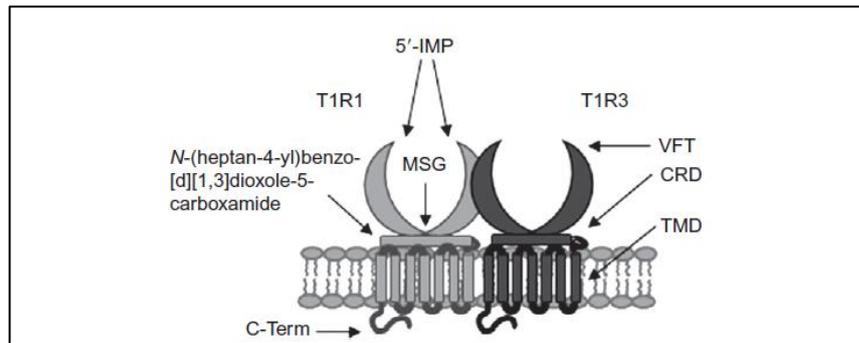


Figura 3. Representación esquemática de una célula receptora del gusto umami (Suess, Festring, & Hofmann, 2015).

Insípido

Este término no define a una sensación provocada por un alimento, por el contrario, la insipidez hace referencia a la ausencia o pérdida de sensación de uno de los cinco gustos básicos: dulce, amargo, ácido, salado o umami (Damodaran et al., 2008).

En la Tabla 1 se sintetiza información sobre cuatro de los gustos básicos (umami, dulce, amargo, ácido); donde constan receptores; clase de gusto que pertenecen y ciertos ejemplos de estructuras moleculares.

Tabla 1. Gustos básicos con sus respectivos receptores y clase de gusto.

Cualidad del gusto	Receptor (es)	Clase del gusto	Ejemplos del gusto
Umami	T1R1+T1R3	Amino ácidos	L-Glutamato, L-AP4, glicina, L-amino ácidos
		Potenciadores de nucleótidos	IMP, GMP, AMP
Dulce	T1R2+T1R3	Azúcares	Sacarosa, fructosa, glucosa, maltosa
		Edulcorantes artificiales	Sacarina, acesulfamo-K, ciclamato, aspartame
		D- amino ácidos	D-Fenilalanina, D-alanina, D-serina (también algunos L-amino ácidos selectivos)

		Proteínas dulces	Monelina, taumatina, curculina
Amargo	T2R5		Cicloheximida
	T2R8, T2R4, T2R44		Denatonio
	T2R16		Salicina
	T2R38		-
	T2R43, T2R44		Sacarina
	No conocidas		Quinina, estricnina, atropina
Ácido	PKD2L1	Ácidos	Ácido cítrico, ácido tartárico, ácido acético, ácido clorhídrico

(Chandrashekar, Hoon, Ryba, & Zuker, 2006)

Por otra parte, la Tabla 2 muestra los estándares básicos del gusto, proporcionando un criterio bajo el cual se podría realizar la catación de alimentos, evitando puntos de vista subjetivos al momento de otorgar una calificación.

Tabla 2. Estándares para los cinco gustos básicos.

Estándares básicos del gusto		
Gusto	Químico usado para gustos básicos	Escala de intensidad
Dulce	Azúcar (sacarosa)	1
Ácido	Ácido cítrico	1
Salado	Cloruro de sodio	1
Amargo	Sulfato de quinina	1
Umami	Ácido glutámico	1

Adaptado de (Ghirri & Bignetti, 2012; Kelly, Spillane, & Newell, 2005)

Análisis sensorial

La evaluación sensorial (o análisis sensorial) es una disciplina científica que se utiliza para evocar, medir, analizar e interpretar reacciones a productos o materiales que son percibidos por los sentidos de la vista, el olfato, el gusto, el tacto y el oído. Inicialmente se desarrolló la evaluación sensorial para estudiar las reacciones a determinadas características de los productos alimenticios. Estas reacciones se miden generalmente en forma de puntuaciones otorgadas a los atributos percibidos de los estímulos alimentarios (Ruan & Zeng, 2013). En el

caso particular de compuestos químicos puros, también se realiza un análisis sensorial mediante comparación con un estándar, en base al cual se definirá el gusto de un compuesto.

Por ejemplo, para determinar el gusto de 19 compuestos sintetizados derivados del sulfamato, se utilizó un panel de cinco personas y se empleó la metodología de “sorber y escupir”. La evaluación del gusto se llevó a cabo a $18 \pm 0,5$ °C. Las soluciones para degustación se prepararon en matraces aforados con agua destilada y se degustaron dentro de las 24 h de su preparación. Todos los sulfamatos de sodio se probaron como soluciones 0,01 M, una concentración que se encuentra para dar sabores detectables. Cada solución de sulfamato se presentó a un grupo experimentado de panelistas en vasos de plástico blancos y limpios. El panel recibió un máximo de cinco muestras por sesión de degustación y se utilizaron alícuotas fijas de 8 mL. Se han utilizado soluciones estándar de sacarosa, ácido cítrico, sulfato de quinina y cloruro de sodio para identificar los gustos dulce, ácido, amargo y salado, respectivamente. Para permitir a los panelistas detectar un gusto definido, cada uno de los estándares se preparó en una concentración igual o superior a su umbral de reconocimiento. A los catadores se les dio primero una muestra de cada uno de los estándares primarios y una muestra de agua destilada pura. Posteriormente, se les entregó las soluciones de prueba de sulfamato y se les pidió que identificaran uno o más de los estándares que describían mejor el gusto. También se les pidió que registraran la presencia de regusto. Cada uno de los 19 sulfamatos sintetizados en este trabajo se probó una única vez. En las pruebas para establecer la confiabilidad y obtener una desviación estándar de los procedimientos de degustación empleados, a los mismos cinco evaluadores se les dio un compuesto dulce (S), uno no dulce (N) y uno dulce/no dulce (N/S) y cada uno fue probado cinco veces por los evaluadores. La solución de prueba se administró de forma aleatoria intercalada con los cuatro estándares y agua destilada. Se ignoró el porcentaje de insipidez al momento de evaluar si el compuesto era predominantemente dulce, no dulce o mostraba cantidades sustanciales de ambos, en cuyo caso la muestra se etiquetó como N/S (Kelly et al., 2005).

Bases de datos de gustos de moléculas

SuperSweet

La base de datos SuperSweet fue publicada en el año 2011 y ofrece información de alrededor 8000 compuestos dulces, tales como hidratos de carbono, proteínas, D-aminoácidos y edulcorantes artificiales. Cada molécula exhibe sus propiedades fisicoquímicas, calorías, estructura y su respectivo dulzor. Posiblemente, a futuro, puedan incluirse los compuestos en mapas de vías bioquímicas para evitar ciertos riesgos; además podrían incluir el receptor del gusto dulce. Para la búsqueda de similitudes en SuperSweet, se implementó una “huella dactilar estructural”, la misma que codifica las características tanto químicas como topológicas de un compuesto. La huella dactilar se pre-calcula para las moléculas de SuperSweet y el compuesto de consulta, para así compararlo con las entradas de la base de datos. Se

incluyeron cuatro huellas dactilares (FP, fingerprints) diferentes (FP2, FP3, FP4 y MACCS) mediante el programa Open Babel. Debido a que FP2 no es la huella dactilar capaz de distinguir entre estructuras con anillos y por tanto entre carbohidratos, no es la más apropiada en SuperSweet. Para superar este inconveniente, se combinó las huellas dactilares FP2 y FP4. La huella dactilar FP4 se basa en un conjunto de patrones SMARTS y considera grupos funcionales. Para la búsqueda de similitudes moleculares se utiliza el coeficiente para datos binarios de Tanimoto (cero indica que no hay bits en común y uno que todos los bits son iguales) (Ahmed et al., 2010).

BitterDB

La base de datos BitterDB fue publicada en el año 2012. Contiene aproximadamente 1000 compuestos, de los cuales en su mayoría son de gusto amargo, pero también existen moléculas declaradas con un gusto diferente. Los compuestos que forman parte de BitterDB, se incluyeron solo si es que cumplían con el criterio de ser amargos o si experimentalmente demuestran ser capaces de activar al menos un receptor humano de amargor. La base de datos muestra información acerca de las principales propiedades moleculares, amargor, SMILES, número de registro CAS, entre otras. De manera adicional, BitterDB provee un enlace a la librería PubChem, e incluye enlaces a las publicaciones donde se explica las interacciones ligando-receptor, concentración efectiva para la activación del receptor y/o el valor de la concentración efectiva media (EC_{50}). Además, se puede conocer sobre mutaciones en los receptores que influyen en la activación por parte de moléculas de sabor amargo (Dagan-Wiener et al., 2019).

TasteDB

En un trabajo posterior, las bases de datos SuperSweet y BitterDB se fusionaron para formar un conjunto de datos de 806 compuestos, el cual se denominó TasteDB. Paralelamente, se fusionaron 1760 compuestos orgánicos volátiles (VOCs, volatile organic compounds) de las bases de datos SuperScent y Flavornet, para obtener una segunda base de datos a la que se denominó FragranceDB. Posteriormente, con los compuestos de las dos bases de datos (TasteDB y FragranceDB) se definió el espacio químico mediante análisis de componentes principales, usando diversos descriptores moleculares y huellas dactilares moleculares (Ruddigkeit & Reymond, 2014).

TastesDB

En el año 2017 se publicó TastesDB, una base de datos constituida por 727 estructuras moleculares asociadas a los gustos dulce (435 moléculas), amargo (81 compuestos) e insípido (133 estructuras). Debido a que el interés del modelado era la clase dulce, los gustos amargo e insípido se fusionaron en una sola clase etiquetada como no dulce. Cada

compuesto fue representado por huellas dactilares moleculares de conectividad ampliada (ECFPs) y descriptores moleculares independientes de la conformación, que se utilizaron para definir una relación cuantitativa predictiva estructura-actividad (QSAR) basada en un sistema experto. Este modelo integra un escalado multidimensional (espacio químico) obtenido con las ECFPs y los modelos de clasificación de análisis discriminante de mínimos cuadrados parciales (PLSDA, Partial Least Squares Discriminant Analysis) y el método de los *N*-vecinos más cercanos (*N*3, *N*- Nearest Neighbors), desarrollados con los descriptores moleculares (Rojas et al., 2017).

SweetenersDB

En el mismo año 2017 se publicó la base de datos SweetenersDB, la cual reporta 316 edulcorantes con su respectivo dulzor relativo (RS, Relative Sweetness). Con esta base de datos se desarrollaron modelos QSAR basados en máquinas de aprendizaje automático supervisado de bosques aleatorios (RF, Random Forest) y regresión de soporte vectorial (SVR, Support Vector Regression). Posteriormente, se utilizó un conjunto de moléculas naturales tomadas de la base de datos Supernatural II para realizar un cribado virtual y discriminar las estructuras dulces de aquellas con características tóxicas y de sabor amargo. SweetenersDB fue actualizada en el año 2020 (segunda versión) mediante la inclusión de nuevas moléculas y un proceso de curado. Con esta base de datos se realizaron diversos modelos de regresión basados en RF, SVM, AdaBoost Tree, y el método de los *k*-vecinos más cercanos (*k*NN). Posteriormente, se hizo una exploración virtual de un conjunto de aproximadamente 4800 moléculas naturales, lo que permitió identificar tres edulcorantes potenciales que fueron analizados mediante acoplamiento molecular con el receptor de dulzor humano T1R2/T1R3. Para el desarrollo de los modelos, la base de datos se dividió en conjuntos de calibración (252 compuestos) y predicción (64 compuestos) mediante un algoritmo de agrupación de exclusión de esferas, usando descriptores de Dragon normalizados entre 0 y 1. Para el espacio químico se realizó con un análisis de inclusión de vecinos estocásticos con la distribución *t* en el programa python (Bouysset, Belloir, Antonczak, Briand, & Fiorucci, 2020).

BitterSweetForest

La base de datos BitterSweetForest, está conformada por aproximadamente 1000 compuestos artificiales y naturales, recopilados de publicaciones previas; dichos compuestos se estandarizaron mediante el software Instant JChem, el cual identificó los errores presentes en las moléculas. El objetivo de este trabajo fue desarrollar y validar un modelo de aprendizaje automático basado en huellas moleculares para discriminar los compuestos dulces y amargos, así como fármacos y moléculas de toxicidad aguda. BitterSweetForest es el primer modelado de acceso abierto basado en un diagrama de flujo programado en KNIME que muestra una precisión del 95% y un área bajo la curva (AUC) de 0.98 en validación cruzada.

Además, en una prueba independiente alcanzó una precisión del 96% y un AUC de 0.98 para la predicción del sabor amargo y dulce. Se utilizaron cuatro distintas FP calculadas usando el nodo RDKit en KNIME: Huellas de Morgan (2048 bits), pares de átomos (1024 bits), huella dactilar de torsión (1024 bits) y FP de Morgan Feat (2048 bits)(Banerjee & Preissner, 2018).

BitterSweet

BitterSweet es una recopilación de 1205 compuestos dulces y 1171 no dulces, así como 918 amargos y 1510 no amargos. Los compuestos insípidos y de gusto contrastante se incluyeron como controles importantes para la predicción del dulzor y amargor. Las moléculas fueron recopiladas de diferentes publicaciones científicas y otras fuentes, y en el caso de las estructuras que mostraron información no exacta o conflictiva en cuanto a su amargor/dulzor, estas fueron excluidas. Las características moleculares para cada estructura se han descrito mediante el algoritmo de Boruta, con los cuales se han desarrollado modelos de aprendizaje automático de Bosques Aleatorios, Regresión Ridge Logística y AdaBoost (árboles de decisión). Los modelos desarrollados tienen un rendimiento que iguala o supera a modelos previamente reportados en la literatura para la clasificación de compuestos dulces y amargos (Tuwani, Wadhwa, & Bagler, 2019).

Espacio químico basado en similitud molecular

El espacio químico (CS, Chemical Space) se define como el conjunto de todas las posibles estructuras moleculares, el cual puede variar con el conjunto particular de descriptores considerados y parámetros que se utilizan para definir la ubicación de las estructuras moleculares (Medina-Franco, Martinez-Mayorga, Giulianotti, Houghten, & Pinilla, 2008). Una forma de definir el espacio químico se fundamenta en el análisis de componentes principales (PCA, Principal Component Analysis) (Jolliffe, 2002; Wold, Esbensen, & Geladi, 1987). El PCA proyecta los datos en un hiperespacio ortogonal reducido y definido por las componentes principales significativas, las cuales son combinaciones lineales de las variables originales, donde la primera componente tiene la máxima varianza, la segunda componente tiene la segunda máxima varianza y así sucesivamente. Este método permite observar los objetos en el gráfico de puntuaciones (scores) y la proyección de las variables en el gráfico de cargas (loadings).

El PCA se ha usado para definir el espacio químico para las bases de datos TasteDB y FragranceDB, mediante el cálculo de 42 números cuánticos moleculares (MQN, molecular quantum numbers) y 34 huellas dactilares de conteo del sistema de especificación de introducción lineal molecular simplificada (SMIfp, SMILES fingerprint). El espacio químico se definió por el gráfico de puntuaciones de las dos primeras componentes principales PC1 (67.97 % de varianza) y PC2 (15.54 % varianza), lo que se puede considerar como un mapa

general 2D de su espacio químico. En este espacio químico se evidencia el aumento de tamaño de los compuestos en la PC1, mientras que la PC2 separa las moléculas por rigidez estructural. Por otro lado, para el espacio SMIfp, la PC1 cubre el 66.9% de la variabilidad de datos y la PC2 cubre el 18.97%. La primera componente separa las moléculas según el número de carbonos no aromáticos, mientras que la segunda componente lo hace según el número de carbonos aromáticos. Es así que, al comparar el mapa de categorías con los valores de propiedad, se logra apreciar que FragranceDB en su mayoría tiene moléculas no aromáticas. En general, el espacio químico definido por los MQN y las SMIfp ilustran la amplia gama de moléculas que se encuentran en los gustos. Además, el plano bidimensional (PC1-PC2) no refleja ninguna distribución de propiedades de polaridad porque estos se encuentran por lo general en la tercera componente PC3, por lo que se requiere de representaciones adicionales (Ruddigkeit & Reymond, 2014).

Otro enfoque para definir el espacio químico es el uso de las huellas dactilares moleculares de conectividad ampliada (ECFPs, Extended Connectivity Fingerprints) con el escalado multidimensional (MDS, Multidimensional Scaling). Este enfoque se utilizó en la base de datos TastesDB, donde las ECFPs ayudaron a describir 649 moléculas dulces y no dulces mediante una exploración de las similitudes existentes entre compuestos químicos por medio del escalado multidimensional. Así fue posible identificar dos grupos importantes de moléculas dulces y un tercer grupo donde se encuentran las moléculas dulces y no dulces, con los cuales se aplicaron métodos de clasificación de los N-vecinos más cercanos (N3) (Todeschini, Ballabio, Cassotti, & Consonni, 2015) y análisis discriminante de mínimos cuadrados parciales (PLSDA) (Wold, Sjöström, & Eriksson, 2001).

CAPÍTULO I

MATERIALES Y MÉTODOS

1.1. Generación de la base de datos

Después de realizar una revisión y recopilación exhaustiva de información, obtenida de diferentes fuentes científicas, se ha logrado compilar 4358 compuestos a partir de 1 base de datos (Dagan-Wiener et al., 2019), 36 artículos científicos, 4 libros y 46 capítulos de libros (3 libros). Para cada molécula ha sido reportado uno de los cinco gustos básicos: dulce, amargo, salado, ácido, umami, y en ciertos casos las moléculas presentaban insipidez, o multi gustos. Además, para cada compuesto se incluyó el número de registro CAS, notación SMILES (canónico o isomérico), PubChem CID, y su respectivo código de identificación en su fuente bibliográfica primaria. La información antes mencionada de cada compuesto se corroboró en diferentes bibliotecas químicas de acceso libre como PubChem (Kim et al., 2016), ChemSpider (Pence & Williams, 2010) y NIST Chemistry WebBook (Linstrom & Mallard, 2001).

1.2. Diseño y representación de las estructuras moleculares

Los compuestos recopilados han sido diseñados en el programa HyperChem (Hypercube Inc.), un modelador y editor molecular versátil y poderoso paquete computacional. En algunos casos se graficó las moléculas a partir del SMILES canónico, y en otras ocasiones del SMILES isomérico. Es así que primero se diseñó la estructura en 2D y con la opción Model Builder se consiguió la estructura en 3D. Por otro lado, ciertas estructuras fueron diseñadas utilizando el programa Open Babel GUI, donde se colocó el SMILES como entrada, y automáticamente se generó el compuesto en formato HyperChem. En caso de ser necesario se realizaron ciertos ajustes a algunos compuestos (por ejemplo quiralidades), y se optimizó mediante campos de fuerza de la mecánica molecular (MM+).

Formas de representación de estructura molecular

El número de registro del Servicio Químico de Resúmenes (CAS, Chemical Abstracts Service) de la Sociedad Química Americana, es una denominación numérica de registro otorgada a cada molécula para poder identificarla, sin considerar la nomenclatura utilizada para la estructura. Los números CAS en ocasiones son modificados debido a que el compuesto ha sido identificado de mejor manera, y es necesario actualizar su información (Voigt, 2003).

Otra forma de representación de la estructura molecular es la notación lineal de cadena SMILES, un sistema de notación química diseñado para el procesamiento moderno de información química. La notación SMILES permite, basado en los principios de la teoría de grafos moleculares, una especificación de cadena de texto. El uso de la notación SMILES ha sido de gran importancia en la predicción de propiedades bioquímicas, tales como toxicidad

y biodegradabilidad, basándose en el principio fundamental de la química informática de congenericidad (moléculas similares tienen propiedades similares) (Weininger, 1988; Weininger, Weininger, & Weininger, 1989). Existen dos notaciones lineales SMILES, el canónico e isomérico. El canónico no es generado por un algoritmo universalmente aceptado para producirlo, pero esto no representa un problema siempre y cuando se utilice el mismo software para crear, almacenar y buscar dicha notación (O'Donnell, 2008). Por otra parte, el isomérico incluye extensiones para admitir la especificación de isótopos, quiralidad y configuración sobre dobles enlaces (Marquardt, Morbach, Wiesner, & Yang, 2009).

Otra opción que permite buscar y reconocer un compuesto químico es PubChem CID (CompoundID), un identificador propio de la base de datos PubChem. En ciertos casos se obtiene el PubChem SID (SubstanceID), que de igual manera es un identificador pero en este caso para sustancias químicas (Kim et al., 2016).

1.3. Curado de las estructuras moleculares

Debido a que los compuestos químicos disponibles en publicaciones científicas no están exentos de errores, es necesario realizar una verificación para garantizar que las estructuras moleculares sean correctas. El curado se aplica a la base de datos completa y se compone de 5 etapas fundamentales (Fourches, Muratov, & Tropsha, 2010):

1. Eliminación de estructuras químicas que no pueden ser procesadas por algunos programas de cálculo de descriptores: por ejemplo, compuestos inorgánicos, organometálicos y las mezclas de sustancias químicas. Esta exclusión se realiza cuando el programa utilizado para el cálculo de descriptores moleculares no admite este tipo de compuestos.
2. Conversión y limpieza de las estructuras químicas: se relaciona a la preparación de la base de datos, donde las estructuras químicas deben ser codificadas en formatos computacionales específicos. Esta conversión se puede realizar con el diseño de la estructura química en un editor molecular, por ejemplo, HyperChem (Hypercube Inc.). Una forma de verificar que las estructuras químicas modeladas sean correctas es a través de diversas bibliotecas electrónicas.
3. Estandarización y normalización de quimiotipos específicos: una molécula, fragmento o grupo funcional puede ser representada de diferentes formas; por ejemplo, el anillo aromático puede representarse en el formato de Kekulé o en el formato aromático. Es importante elegir la misma forma de representación para aplicarla a todas las estructuras moleculares.
4. Eliminación de compuestos repetidos: es común que algunos compuestos se encuentren repetidos en diversas bases de datos, por ejemplo, en el caso de este estudio, la sacarosa, aspartame, ciclamato, entre otros. Por este motivo es importante identificarlos para fusionarlos, pues su repetición incluye redundancia durante el desarrollo de los

modelos QSAR. Las formas más apropiadas para identificar compuestos repetidos es aplicar filtros por nombre, número de registro CAS, PubChem CID y SMILES canónico.

5. Control manual final: es un paso importante para identificar moléculas no diseñadas correctamente y para corregir cualquier problema que haya aparecido en los cuatro pasos anteriores.

1.3.1 alvaMolecule

alvaMolecule (Alvascience, 2020b) es un programa de nueva generación cuya finalidad es la visualización y curado de bases de datos de estructuras moleculares. Este programa incluye 10 verificadores predefinidos para identificar estructuras erróneas o para filtrar características estructurales específicas: estructuras múltiples, valencia inusual, enlaces covalentes/iónicos, carga total, isótopos, átomos cargados, moléculas sin átomos de carbono, presencia de átomos no estándar (H, C, N, O, P, S, F, Cl, Br, I), aromaticidad y átomos radicales. Por otra parte, permite utilizar diversos criterios de estandarización para corregir representaciones erróneas de las estructuras, para eliminar características específicas de las moléculas o para estandarizar características estructurales específicas: estandarización de anillos aromáticos (bencénicos), convertir enlaces covalentes inusuales en enlaces iónicos, agregar cargas a nitrógenos cuaternarios, eliminar (o agregar) átomos de hidrógeno en exceso (o faltantes), eliminar fragmentos monoatómicos, retener el fragmento más grande, estandarizar diversos grupos funcionales (nitro, azida, diazo), eliminar isótopos, eliminar quiralidad y dirección de enlaces.

1.4. Fusión y filtrado de las bases de datos

Es importante realizar una fusión de los datos obtenidos porque permite obtener información unificada procedente de fuentes de datos heterogéneas. Entonces, el primer paso es identificar los atributos que se utilizarán para describir los elementos de la información. En segundo lugar, se debe identificar los diferentes compuestos que se describen en las fuentes de los datos, de esta manera se puede detectar estructuras duplicadas. Finalmente, las moléculas duplicadas serán fusionadas en una sola representación (Bleiholder & Naumann, 2009). En otras palabras, se puede decir que la fusión y filtrado de datos es una combinación de múltiples fuentes para obtener información mejorada, relevante y de una manera menos compleja.

Para lograr el proceso de fusión y filtrado se utiliza KNIME (Constan Information Miner) (Berthold et al., 2009) un programa de acceso libre que permite realizar minería de datos. Ofrece un espacio de trabajo gráfico para el ensamblaje visual donde se integra de manera sencilla nuevos algoritmos y herramientas, así como métodos de manipulación o visualización de datos en forma de nodos. Estos nodos se encuentran instalados y organizados por categorías, las mismas que sirven para programar el diagrama de flujo. Los módulos (nodos)

que se encuentran en el repositorio se deben arrastrar hacia el área de trabajo para enlazarlos mediante puertos de entrada y salida, y posterior a esto se configuran para que se lleve a cabo las operaciones requeridas.

1.5. Espacio químico del gusto

Para la definición del espacio químico del gusto se usarán las huellas dactilares moleculares de conectividad ampliada, con las cuales se desarrollará un escalado multidimensional.

1.5.1. Huellas dactilares moleculares de conectividad ampliada

Las huellas dactilares moleculares de conectividad ampliada (ECFPs, Extended Connectivity Fingerprints) (Rogers & Hahn, 2010) son FPs topológicas circulares que codifican la presencia/ocurrencia de fragmentos centrados en el átomo identificados mediante una variante del algoritmo de conectividad ampliado de Morgan. Este tipo de representación molecular también se denomina huellas dactilares moleculares de Morgan o huellas dactilares moleculares circulares (Mauri, Consonni, & Todeschini, 2017).

Las huellas dactilares moleculares fueron inicialmente desarrolladas para la búsqueda rápida de bases de datos, debido a que son capaces de identificar subestructuras comunes. Sin embargo, las FPs también se las aplica en estudios QSAR, particularmente en el desarrollo de modelos basados en similitudes locales (Cassotti, Consonni, Mauri, & Ballabio, 2014). Los parámetros generales que se deben considerar para la generación de las huellas dactilares moleculares de trayecto y de conectividad ampliada son:

Generación de patrones

Constituye el primer paso para el cálculo de las FPs y consiste en la identificación de todas las subestructuras posibles en una molécula. Para las ECFPs, los fragmentos centrados en el átomo constituyen subestructuras circulares que generan “árboles” de una longitud establecida, fijada en un vértice particular del grafo molecular y que dependerá de la longitud del patrón.

Longitud del patrón

La longitud del patrón indica el radio máximo a ser explorado desde el átomo centrado; es decir, el número de enlaces máximos a recorrer. La longitud del patrón indica la forma en que se explora el entorno circundante al átomo central. Por ejemplo, a radio 0 se obtienen subestructuras representadas únicamente por los átomos individuales, a radio 1 se obtiene la información de todos los átomos directamente enlazados al átomo centrado; mientras que a

radio 2 se obtiene la información de todos los átomos dentro de un diámetro de 4 enlaces químicos. Seguidamente, cada fragmento se codifica mediante la notación lineal SMARTS (Daylight Chemical Information Systems Inc., 2011), para luego eliminar fragmentos iguales o equivalentes; es decir, aquellos fragmentos que contienen exactamente el mismo grupo de enlaces. Generar mayor cantidad de subestructuras permite obtener mayor información estructural; pero contemporáneamente aumenta el porcentaje de bits igual a 1 (opacidad).

Función Hash

Es un algoritmo matemático que permite llevar varios vectores de entrada con dimensión variable a una dimensión fija, de tal forma que sea posible el tratamiento matricial de los datos. En el caso de las ECFPs, esta función se usa para asignar datos a una subestructura dada dentro de un vector booleano de una longitud fija (dimensión final de la huella dactilar molecular) y se usa el operador lógico “o” para generar una huella dactilar molecular de toda la molécula. Es decir, cada subestructura es mapeada a la FP final mediante un cierto número de bits. Esta función es determinística, lo que permite que cada subestructura sea siempre mapeada por el mismo conjunto de bits. El número de bits por patrón corresponde al número de bits que se usa para codificar con la función de resumen, una misma subestructura. La probabilidad de colisión se reduce al aumentar el número de bits por patrón, pero al mismo tiempo incrementa la opacidad.

alvaDesc

El programa alvaDesc (Alvascience, 2020a; Mauri, 2020) fue lanzado a inicios del año 2019 y es un programa de nueva generación para el cálculo de descriptores moleculares y huellas dactilares. La versión más reciente permite calcular 5666 descriptores moleculares y 166 claves estructurales MACCS. También calcula dos tipos de huellas dactilares moleculares: 1) de conectividad ampliada (ECFPs) y 2) de trayecto (PFPs), para lo cual utiliza la notación de cadena SMARTS como identificador de los fragmentos. Este programa permite trabajar con moléculas desconectadas mediante diferentes enfoques teóricos. Adicionalmente, permite la visualización y filtrado de las estructuras químicas, así como su verificación en la biblioteca química PubChem. Por otra parte, alvaDesc permite eliminar descriptores constantes y casi constantes, descriptores con valores faltantes e implementa diversas estrategias para realizar una reducción no supervisada de descriptores, tales como el análisis de componentes principales, análisis de correlación entre pares de descriptores y el método V-WSP.

1.5.2. Escalado multidimensional

El escalado multidimensional (MDS, Multidimensional Scaling) (Kruskal, 1964; Winsberg & Carroll, 1989) es una técnica que permite reconstruir las similitudes/disimilitudes entre pares de objetos mediante el uso de distancias, para proyectarlas en un número pequeño de

dimensiones. Esto significa que el MDS busca acomodar los objetos en un espacio definido por un cierto número de dimensiones de forma tal que se reproduzcan las distancias observadas. El escalado multidimensional se basa en el análisis de la matriz de distancias \mathbf{D} , obtenida a partir de la matriz original de los datos \mathbf{X} ; donde D_{st} representa la distancia entre los objetos s y t en el espacio p -dimensional y D_{st} representa la distancia entre s y t en el subespacio M -dimensional ($M < p$). De esta forma, el análisis de las coordenadas principales (ejes del MDS) en el subespacio M -dimensional brinda una solución para la cual es mínima la siguiente función:

$$L^* = \min \left[\sum_s \sum_t w_{st} (D_{st} - D_{st})^2 \right] \quad (0.1)$$

Donde w_{st} son los elementos de una matriz de pesos estadísticos. Al igual que en el PCA, se debe considerar la selección del número apropiado de dimensiones. Generalmente, en el MDS se buscan proyecciones en 2 o 3 dimensiones ($M = 2, 3$). El número de dimensiones se puede lograr en función de los residuos entre las distancias originales y las distancias representadas en las dimensiones seleccionadas; es decir, mientras más bajo es el residuo, mejor es la aproximación de las distancias en las proyecciones. Por consiguiente, el algoritmo del MDS usado con mayor frecuencia asume el caso particular donde $w_{st} = 1/D_{st}$ y se basa en la minimización de la siguiente expresión:

$$\min \left[\sum_{s < t} \frac{(D_{st} - D_{st})^2}{D_{st}} \right] \quad (0.2)$$

De esta manera, el MDS permite obtener un gráfico de dispersión de los objetos en un espacio reducido, en el cual se observa una reproducción de las distancias originales y que permite un análisis rápido y fácil de las relaciones entre los objetos. A diferencia del PCA, los ejes coordenados en el subespacio M no corresponden a combinaciones lineales de las variables originales por lo que no se les puede atribuir una interpretación.

Medidas de distancia para datos binarios

El concepto de similitud se usa para reconocer, distinguir o clasificar objetos. Para medir la similitud (o su complemento la disimilitud) se recurre al uso de medidas de distancia. Si los datos son representados por variables binarias o dicotómicas (0 y 1), por ejemplo, la ausencia/presencia de grupos funcionales o fragmentos moleculares en una estructura química; se debe recurrir al uso de medidas para datos binarios (Tabla 3). A continuación se describirán algunas distancias (Todeschini, Ballabio, & Consonni, 2015):

Tabla 3. Medidas de distancia para datos binarios.

Distancia	Fórmula
Jaccard-Tanimoto	$D_{st} = \frac{a}{a + b + c}$
Rogers-Tanimoto	$D_{st} = \frac{a + d}{p + b + c}$
Sokal-Michener	$D_{st} = \frac{a + d}{p}$
Gleason-Dice-Sorensen	$D_{st} = \frac{2a}{2a + b + c}$
Consonni-Todeschini (CT4)	$D_{st} = \frac{\ln(1 + a)}{\ln(1 + a + b + c)}$
Austin-Colwell	$D_{st} = \frac{2}{\pi} \arcsen \sqrt{\left(\frac{a + d}{p}\right)}$

Estas medidas de distancia se calculan a partir de una tabla de frecuencias, tal como se presentan en la Tabla 4.

Tabla 4. Tabla de frecuencias de las cuatro combinaciones posibles de valores 0 y 1 para muestras con variables binarias.

	t = 1	t = 0	
s = 1	a	b	a + b
s = 0	c	d	c + d
	a + c	b + d	p

Los elementos a , b , c y d son las frecuencias de ocurrencia de las cuatro combinaciones posibles al comparar los dos vectores \mathbf{s} y \mathbf{t} representados por p variables binarias ($p = a + b + c + d$). Particularmente, a es el número de “presencias comunes” de los atributos y d el número de “ausencias comunes”; mientras que $a + b$ es el número de atributos presentes en \mathbf{s} y $a + c$ el número de atributos presentes en \mathbf{t} . Los elementos diagonales a y d brindan información de la similitud entre el par de vectores; mientras que los elementos no diagonales b y c proveen información acerca la disimilitud entre los mismos.

PCA toolbox para MATLAB

El PCA toolbox para MATLAB (Ballabio, 2015) es una colección de módulos para calcular modelos no supervisados para el análisis de la estructura de los datos, por ejemplo, el análisis de componentes principales (PCA), el escalado multidimensional (MDS) y el análisis de conglomerados. Este toolbox permite el uso de una interfaz gráfica (GUI) que permite cargar los datos, realizar los modelos no supervisados y visualizar los resultados de forma amigable. También permite grabar los modelos e importar los gráficos desarrollados con mucha facilidad.

CAPÍTULO II

RESULTADOS Y DISCUSIÓN

2.1. Generación de la base de datos

Para generar la base de datos se realizó una búsqueda y recopilación extensa y cautelosa de diversas fuentes bibliográficas (Tabla 1A de los anexos), de donde se obtuvieron 4484 compuestos orgánicos e inorgánicos. Cada una de las estructuras químicas obtenidas muestra su código de identificación, PubChem CID, número de registro CAS, notación SMILES, gusto (clase) y su referencia bibliográfica. Esta base de datos recopila moléculas de los cinco gustos básicos: dulce, amargo, ácido, salado, umami, así como compuestos insípidos, no dulces, multi gustos y un grupo con gustos astringente y picante, que han sido englobados en la categoría “otros”. Durante la generación de la base de datos, se han excluido proteínas que presentan gusto, por ejemplo, la curculina, monelina, miraculina, entre otras; así como enzimas (pepsina, α -quimotripsina, molsin y pronasa). Es importante hacer énfasis que en bases de datos previamente publicadas, por ejemplo, SweetenersDB (Bouysset et al., 2020) y BitterDB (Dagan-Wiener et al., 2019), se centran en estudiar únicamente compuestos dulces y amargos, respectivamente.

2.2. Diseño y representación de las estructuras moleculares

Para poder representar cada una de las 4484 moléculas, se utilizó el programa HyperChem (Hypercube Inc.). Con ayuda de los SMILES se logró modelar las estructuras, aunque en algunas de ellas no se pudo identificar de manera clara sus quiralidades por tener un SMILES canónico, a diferencia del isomérico que muestra más información acerca de la molécula. Después, las geometrías se optimizaron mediante los campos de fuerza de la mecánica molecular (MM+) usando el algoritmo de gradiente conjugado en la versión Polak-Ribiere. Las moléculas se consideraron como optimizadas cuando la desviación estándar del vector gradiente es menor a $0.01\text{kcal}\times (\text{\AA}\times\text{mol})^{-1}$. Al momento de diseñar las estructuras, los monosacáridos D-galactosa, fructosa y D-tagatosa se representaron mediante la proyección de Haworth, pues es muy común su uso para azúcares cíclicos como las hexosas (Dickerson, 1992).

2.3. Curado de las estructuras moleculares

Con la finalidad de poder estandarizar la forma de representación de los compuestos químicos, se utilizó el programa alvaMolecule (Alvascience, 2020a) para otorgar a cada una de las moléculas su respectiva cadena SMILES (canónico), de esta manera todas las estructuras podrían ser identificadas bajo la misma forma de notación.

2.4. Fusión y filtrado de las bases de datos

Para esta etapa, se agruparon en una única base de datos todas las estructuras pertenecientes a todas las clases, y una vez fusionada se trabajó únicamente sobre la clase de los compuestos, pues existía gran variedad de gustos y algunos de ellos se podían concatenar o agrupar en una sola clase. Mediante un diagrama de flujo programado en KNIME (Figura 4), se aplicaron ciertos filtros para agrupar a los compuestos de acuerdo a su nombre, número de registro CAS, PubChem CID y SMILES canónico. Además, se realizó un curado de forma manual que nos ayudó a identificar compuestos con ciertas singularidades. En el grupo del gusto dulce se incluyeron todas aquellas moléculas que estaban declaradas como apenas dulce, muy dulce, intensamente dulce, carente de dulzura, menos dulce, levemente dulce, dulzón, muy dulce, dulce/fresco y dulce. Por otro lado, en la clase amargo se incluyó: totalmente amargo, extremadamente amargo, levemente amargo, débilmente amargo, muy amargo, intensamente amargo, amargor moderado, moderadamente amargo, algo amargo, amargo fuerte, fuertemente amargo, muy amargo, amargo-levemente amargo y amargo. Para el grupo de gusto ácido se juntaron aquellos compuestos con gusto agrio y ácido. En el caso del gusto salado, se consideraron los compuestos descritos solo como salado. Para el gusto umami se abarcaron las clases: salado-umami, más umami, débilmente umami y umami. Para la clase “insípido” se han incluido compuestos denominados como: neutro e insípido. Otro grupo, pero no menos importante, es la clase “multi gusto”, que comprende los compuestos etiquetados como: amargo/dulce, amargo/salado, amargo/astringente, amargo/picante, amargo/ardiente, amargo/astringente/dulce, dulce/astringente, dulce/insípido, no amargo/ardiente, ácido/amargo, dulce/amargo/ácido, dulce/no dulce, amargo/ardiente/picante, amargo/repugnante, amargo/insípido, dulce/salado, no dulce/dulce, ácido/amargo/dulce, umami/ácido, umami/amargo. Además, en la base de datos existen compuestos no dulces. Asimismo, se incluyó un grupo de compuestos que son identificados con la clase “otros” por no ser comunes o abundantes en comparación a los gustos antes mencionados. Dentro de este grupo se incluyen los gustos quemante y astringente. Por otra parte, se decidió excluir el compuesto limonina monolactona (Belitz, Grosch, & Schieberle, 2009) debido a que es el único compuesto que fue reportado como no amargo.

Después de realizar una primera clasificación de los compuestos, existieron algunos de ellos que estaban repetidos y presentaban distintos gustos, entonces se optó por aplicar ciertos criterios para el filtrado de la base de datos:

1. Aquellos compuestos que presenten dos gustos distintos, serán colocados en la base de datos “multi gusto”.
2. Para aquellas moléculas que presenten los tres gustos distintos, serán asignadas a la base de datos “multi gusto”, caso contrario, cuando las moléculas muestren dos gustos iguales y uno distinto, serán designadas a la base de datos del gusto predominante.

3. Para aquellos compuestos que muestren de 4 gustos en adelante: se colocará la molécula en la base datos "multi gusto" si todas las clases son distintas o cuando exista un empate entre las clases. Para el resto de casos donde si existe una clase dominante se aplicará el criterio del voto mayoritario.

Al final del tratamiento de los datos se han identificado 3589 estructuras moleculares divididas de la siguiente manera: 1521 dulces (Tabla 2A), 1185 amargos (Tabla 3A), 38 compuestos ácidos (Tabla 4A), 10 salados (Tabla 5A), 99 umami (Tabla 6A), 247 insípidos (Tabla 7A), 301 no dulce (Tabla 8A), 180 multi gusto (Tabla 9A), y 8 compuestos de la clase otros (astringente y picante) (Tabla 10A).

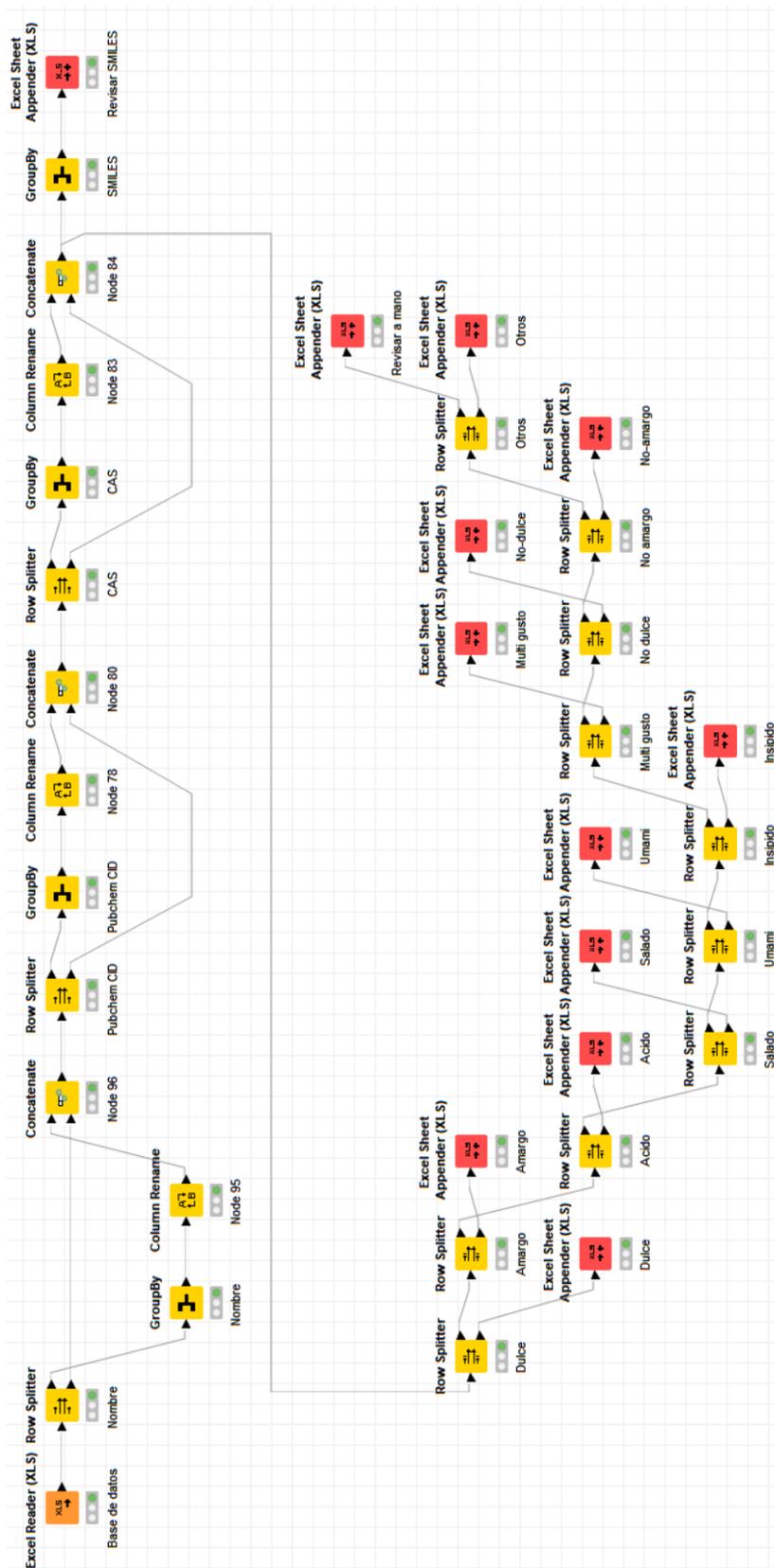


Figura 4. Diagrama de flujo KNIME para el filtrado de la base de datos.

2.5. Huellas dactilares moleculares de conectividad ampliada

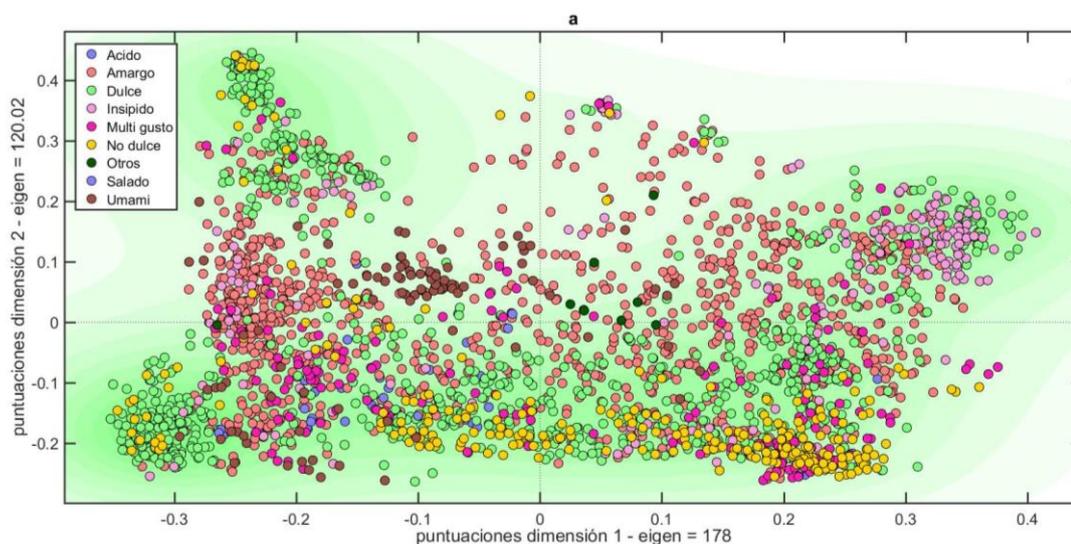
Para cada estructura molecular, utilizando los SMILES canónicos, se han calculado las huellas dactilares moleculares de conectividad ampliada (ECFPs) (vectores booleanos) en el programa alvaDesc (Alvascience, 2020a; Mauri, 2020). La dimensión de las ECFPs es de 1024 variables obtenidas usando 2 bits por patrón, con un radio mínimo de 0 y máximo de 2. Se consideraron las opciones atómicas: contar fragmentos, tipo de átomo, aromaticidad, conectividad total, carga y orden de enlace.

2.6. Espacio químico

La proyección del espacio químico se obtuvo mediante el MDS programado en el PCA toolbox para MATLAB. Se introdujo la matriz de las ECFPs obtenidas previamente, y de manera adicional se agregó una columna donde se importó las clases de los compuestos. Al ser las ECFPs datos binarios (valores 0 - 1), se utilizó la distancia de Jaccard-Tanimoto sin la opción de auto escalado de los datos. De esta manera se obtienen las gráficas de las puntuaciones (scores) que definen el espacio químico. Aquí es factible visualizar las similitudes o disimilitudes que existen en el espacio multidimensional.

2.6.1. Espacio químico para el gusto dulce

El espacio químico para el gusto dulce se definió mediante la proyección de las dos primeras dimensiones del MDS (Figura 5a), en el cual se observa que existe sobre posición de las moléculas de los distintos gustos. Sin embargo, también se evidencian algunos grupos consistentes de compuestos dulces que se distinguen de los otros gustos. Para identificar estos grupos se ha realizado una gráfica en la que se observan únicamente las clases dulce y no dulce en estas dos dimensiones (Figura 5b).



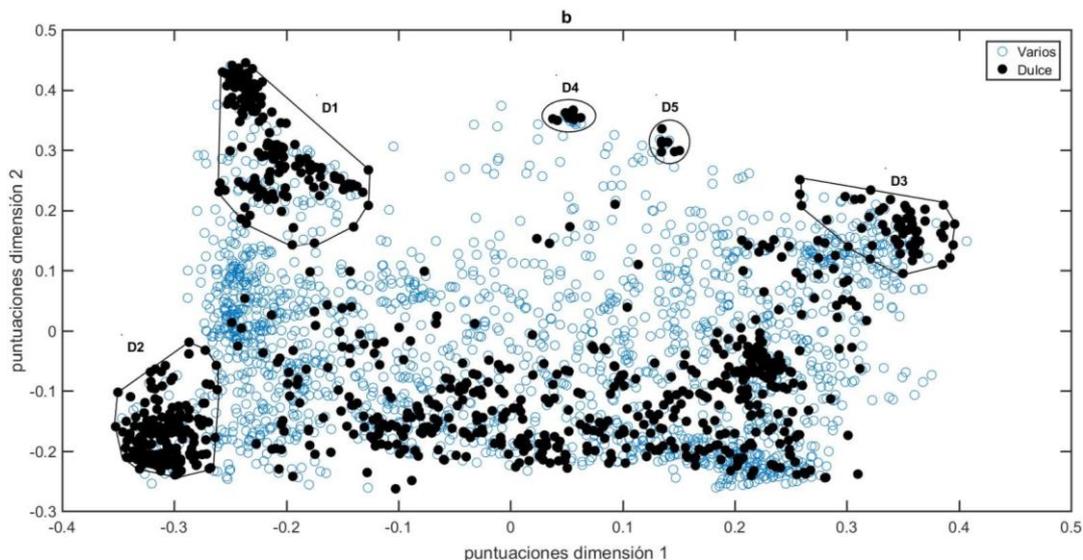


Figura 5. a) Espacio químico definido mediante el MDS para el gusto dulce. b) Proyección de las moléculas dulces y no dulces en el espacio químico.

En el grupo de moléculas dulces D1, a pesar de existir una cercanía entre los compuestos, se logra diferenciar dos conjuntos, uno consistente donde ciertas estructuras están más cercanas entre sí, y otro grupo donde las moléculas están un poco dispersas. En términos generales, estos compuestos resultan ser derivados de la sacarosa y otros azúcares como la fructopiranosas. También, al analizar el conjunto se visualiza una ligera sobre posición de la clase dulce con compuestos amargos y no dulces. Las moléculas del conjunto D1 están presentes en el segundo cuadrante de la Figura 5a.

El conjunto de compuestos D2 es más consistente, es decir son muy cercanas entre sí. Son muy pocas las moléculas del conjunto que están dispersas; sin embargo, existe una sobre posición con compuestos pertenecientes a la clase no dulce. Aquellas moléculas que forman parte de este conjunto tienen la característica común de ser derivados del aminoácido aspartato. El grupo D2 se encuentra claramente en el tercer cuadrante del gráfico, donde las moléculas se evidencian en la parte negativa de la primera y segunda dimensión.

El grupo D3 también es un grupo consistente y muestra una notable sobre posición con compuestos insípidos e incluso con ciertos compuestos de gusto amargo. La mayoría de las estructuras tienen en común la presencia de los fragmentos 2-methoxyphenol y benceno. El conjunto de compuestos químicos pertenecientes a D3 se observan en el primer cuadrante del plano.

Finalmente, los grupos D4 y D5 son pequeños en comparación a los descritos anteriormente, y a pesar de eso son grupos consistentes que presentan una ligera sobre posición con compuestos amargos e insípidos. Las moléculas pertenecientes a estos dos grupos, mayoritariamente se muestran como derivados de dihidrochalconas. Las estructuras de los conjuntos D4 y D5 están ubicadas en el primer cuadrante del gráfico.

2.6.2. Espacio químico para el gusto amargo

El espacio químico para el gusto amargo se definió mediante la proyección de las dimensiones 1 y 3 del MDS (Figura 6a), en el cual se observa la sobre posición con moléculas de los distintos gustos. Para este gusto también se evidencian algunos grupos consistentes de compuestos que se distinguen de mejor manera en la Figura 6b.

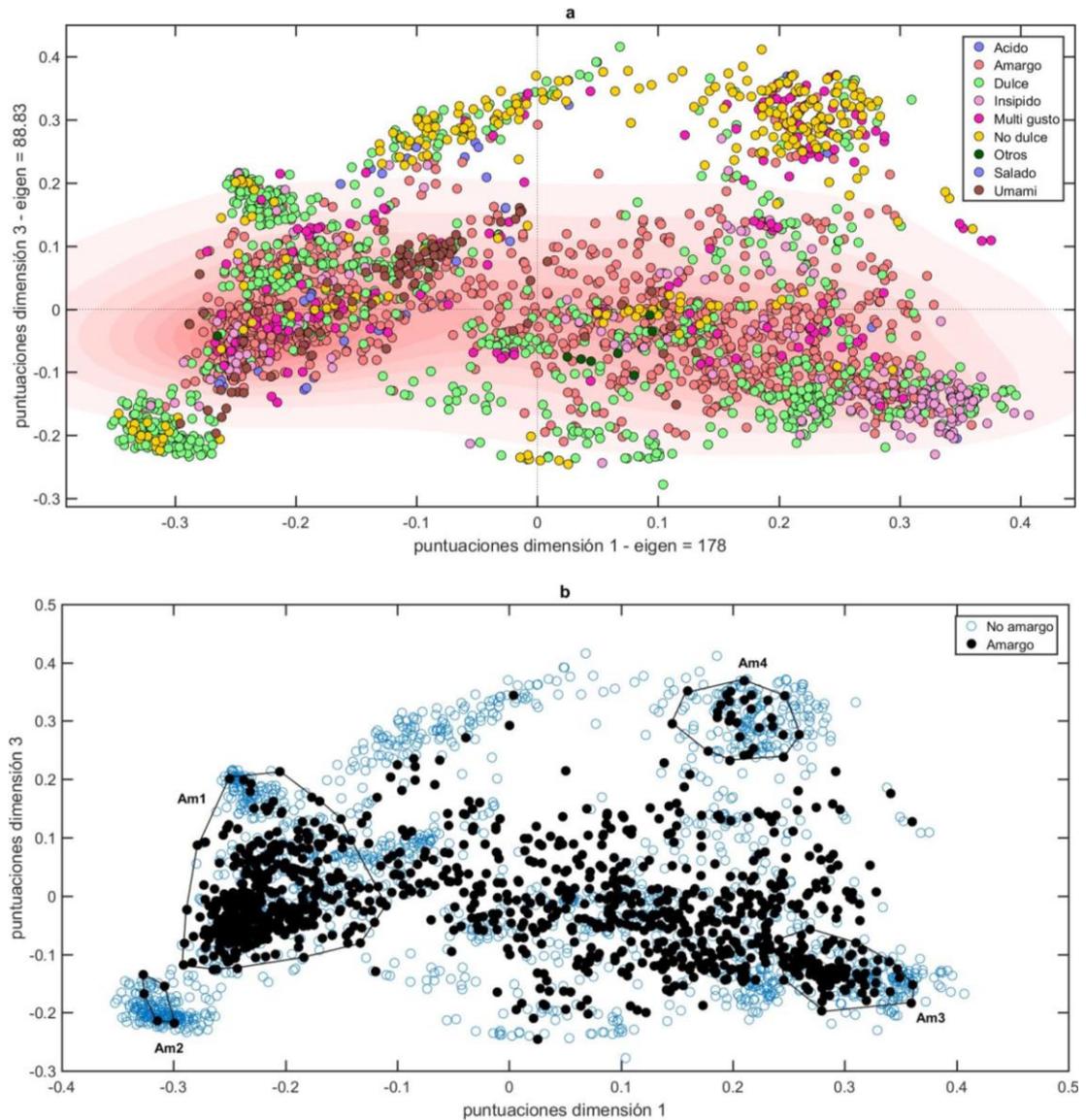


Figura 6. a) Espacio químico definido mediante el MDS para el gusto amargo. b) Proyección de las moléculas amargo y no amargo en el espacio químico.

El grupo de moléculas Am1 es consistente, con algunas de ellas algo dispersas. Este grupo muestra una considerable superposición con compuestos de la clase dulce, umami, multi gusto y algunas estructuras insípidas. La mayoría de moléculas de este grupo son aminoácidos y algunos derivados del sulfamato. El conjunto Am1 se encuentra distribuido claramente entre los cuadrantes 2 y 3 del plano, hacia las puntuaciones negativas de la primera dimensión.

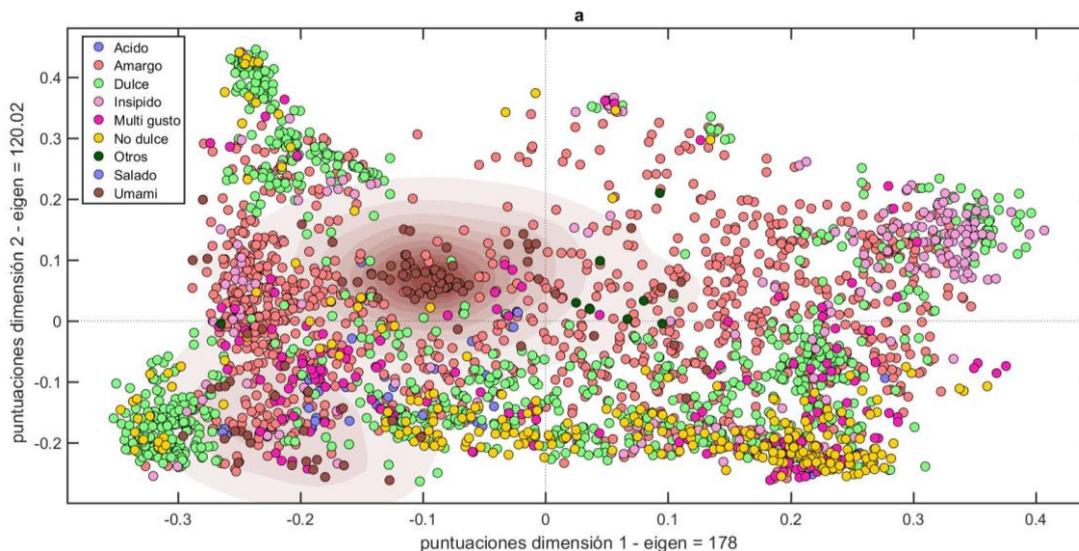
Por otra parte, el grupo Am2 contiene pocas moléculas y muestra alta similitud entre ellas (consistencia). De hecho, estos compuestos son derivados del ácido aspártico. En la Figura 6a se identifica al grupo Am2 con una gran superposición con la clase dulce y no dulce. Se puede observar que las moléculas de Am2 están concentradas en el tercer cuadrante del plano, de manera alejada del punto central.

En el caso del grupo Am3, se puede identificar que es un conjunto consistente, aunque ciertas moléculas están un poco dispersas en relación a las otras. Una característica común que tiene los compuestos de Am3 es que son derivados de la chalcona. La Figura 6a evidencia una notable superposición de estos compuestos con moléculas de clase dulce e insípidos. Los compuestos pertenecientes al conjunto Am3 están ubicados en el cuarto cuadrante del gráfico, hacia el lado positivo de las puntuaciones de la primera componente y negativo para la segunda coordenada.

Finalmente, el grupo Am4 evidencia un conjunto de estructuras químicas que, si bien son similares estructuralmente, también están dispersas entre ellas. En este conjunto se superponen mayoritariamente compuestos no dulces y moléculas multi gusto. Las estructuras de este grupo son mayoritariamente derivados del sulfamato y fenil sulfamato. En el caso de Am4, las moléculas están presentes en el primer cuadrante del plano.

2.6.3. Espacio químico para el gusto umami

El espacio químico para el gusto umami se definió mediante la proyección de las dimensiones 1 y 2 del MDS (Figura 7a), donde se observa que estas moléculas también se superponen con compuestos de los otros gustos. Sin embargo, la Figura 7b permite identificar algunos grupos consistentes de estructuras.



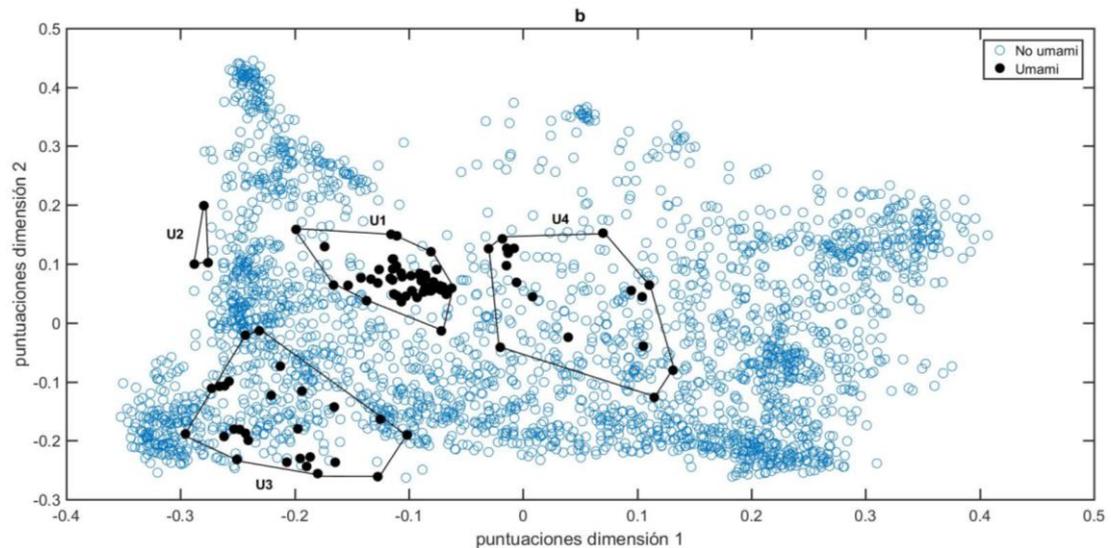


Figura 7. a) Espacio químico definido mediante el MDS para el gusto umami. b) Proyección de las moléculas umami y no umami en el espacio químico.

El grupo U1 es el más compacto, a pesar de presentar una ligera sobre posición con estructuras moleculares del gusto dulce y amargo. La mayoría de los compuestos tienen en común la presencia de grupos imidazol y fosfato disódico/ácido fosfórico. Se puede observar que el conjunto U1 está ubicado mayoritariamente en el segundo cuadrante, muy cerca del origen.

En el caso de U2, se identifica un grupo de tres moléculas (S-morelid, ácido N-(1-desoxi-D-fructos-1-il)-L-glutámico, y ácido N-(1-desoxi-D-fructos-1-il) piroglutámico) que tienen una leve dispersión, pero a pesar de eso es un conjunto, que si bien está cerca de otras moléculas, no representa superposición con las mismas. Las estructuras pertenecientes a este grupo tienen en común fragmentos de pentosa y ácido carboxílico. En la Figura 6a se evidencia la presencia del grupo U2 en el segundo cuadrante del gráfico, en la parte negativa de la componente 1.

En el grupo U3 se puede observar un conjunto de moléculas algo más dispersas entre sí. El conjunto exhibe una superposición mayoritariamente con moléculas amargas, dulces y multi gusto. A pesar de la distancia que existe entre los compuestos, la mayoría de ellos tienen en común ser derivados del ácido glutámico. Además, se puede observar que este conjunto está ubicado, aunque de manera dispersa, en el tercer cuadrante del plano, en la parte negativa de las dos dimensiones.

Finalmente, en el grupo U4 se observa una mayor dispersión entre las moléculas en comparación a los grupos mencionados anteriormente. También existe una superposición considerable con compuestos de la clase dulce, amargo y multi gusto. La similitud entre este grupo se debe a que la mayoría de estructuras se derivan del inosinato disódico. El grupo U4 ocupa regiones cercanas al origen en los cuatro cuadrantes del plano.

2.6.4. Espacio químico para el gusto ácido

El espacio químico para el gusto ácido se definió mediante la proyección de la segunda y quinta dimensión del MDS (Figura 8a). En dicho gráfico se puede observar sobre posición con compuestos de las distintas clases. Por otra parte, en la Figura 8b se puede identificar de mejor manera los grupos consistentes de moléculas que se diferencian de los otros gustos.

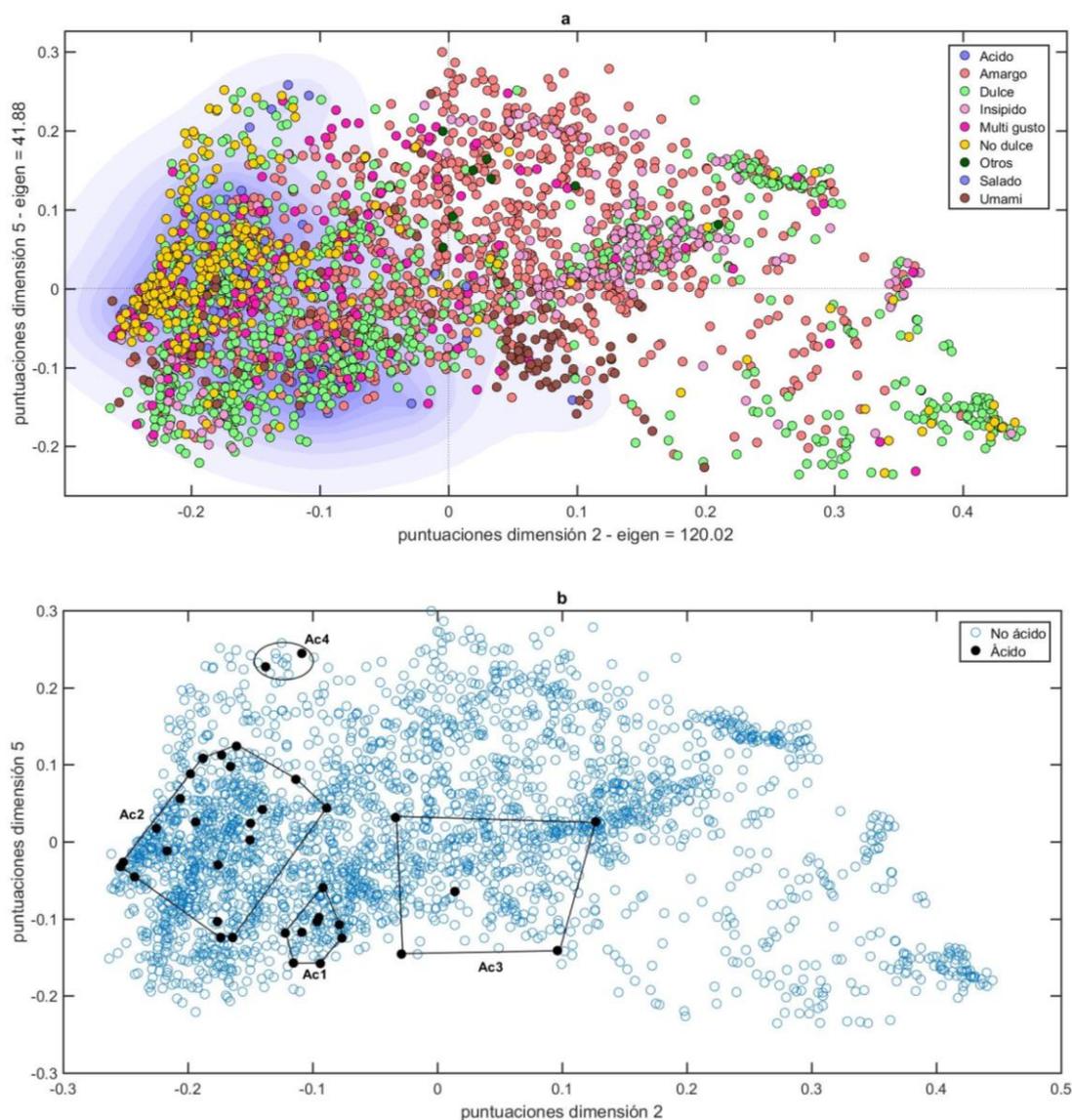


Figura 8. a) Espacio químico definido mediante el MDS para el gusto ácido. b) Proyección de las moléculas del gusto ácido y no ácido en el espacio químico.

El grupo Ac1 es un conjunto de estructuras químicas que muestran buena compactación. Estos compuestos tienen el grupo carboxílico (COOH) como fragmento común. Es un grupo que tiene gran superposición con otros compuestos, en su mayoría, dulces y amargos. Además, este conjunto se ubica de manera clara en el tercer cuadrante del plano, en las puntuaciones negativas de la dimensión 2 y 5.

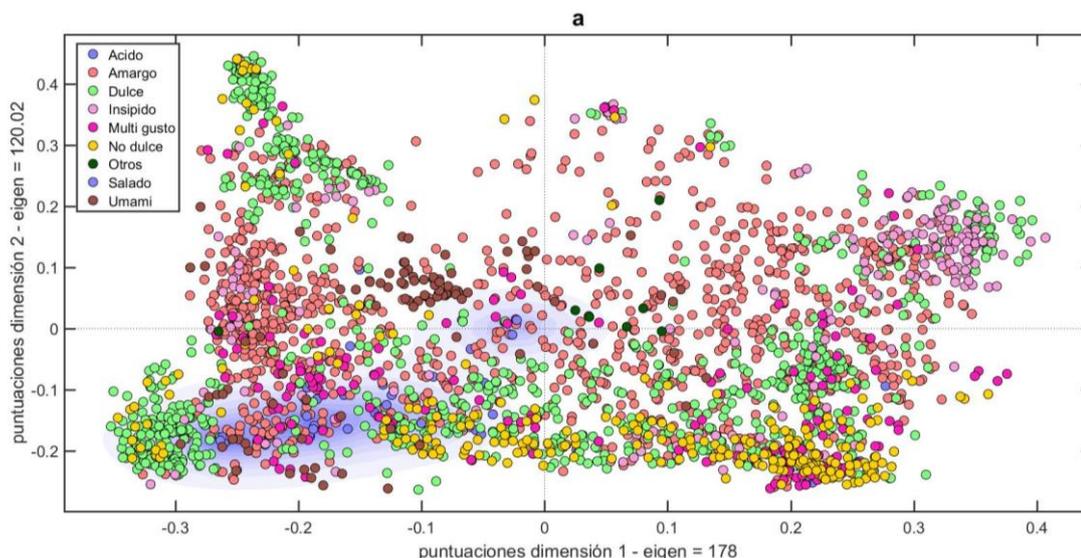
El grupo Ac2 está conformado por moléculas que exhiben una gran dispersión entre ellas. A pesar de la distancia existente, se encuentra que son compuestos que en su gran mayoría resultan ser derivados del sulfamato y fenil sulfamato. Este grupo presenta superposición con estructuras moleculares principalmente dulces, amargas y no dulces, así como algunos compuestos de gusto umami y multi gusto. Se evidencia que este grupo Ac2 está distribuido entre los cuadrantes 2 y 3 del gráfico.

El grupo Ac3 está conformado por cinco moléculas: ácido fosfórico, ácido ascórbico, ácido L-(+)-tartárico, derivado 15 de disulfamato y compuesto XCIV (3-hidroxi-4-etoxidihidrochalcona). Este grupo presenta una dispersión bastante alta, por lo que es difícil definir un esqueleto común para dichas estructuras. Este grupo presenta sobre posición con moléculas de gusto amargo, dulce, umami y compuestos insípidos (Figura 7a).

El grupo Ac4, ubicado en el segundo cuadrante, esta conformado por dos moléculas (derivado 6 del disulfamato y derivado 7 del disulfamato), cuya característica común es de ser compuestos derivados del disulfamato. Este pequeño grupo de estructuras moleculares presenta superposición con compuestos de clase dulce y no dulce.

2.6.5. Espacio químico para el gusto salado

El espacio químico para el gusto salado se definió mediante la proyección de las dimensiones 1 y 2 del MDS (Figura 9a), donde se observa una sobre posición con moléculas de otros gustos. Para el gusto salado también se pueden identificar varios conjuntos consistentes de compuestos que se distinguen de mejor manera en la Figura 9b.



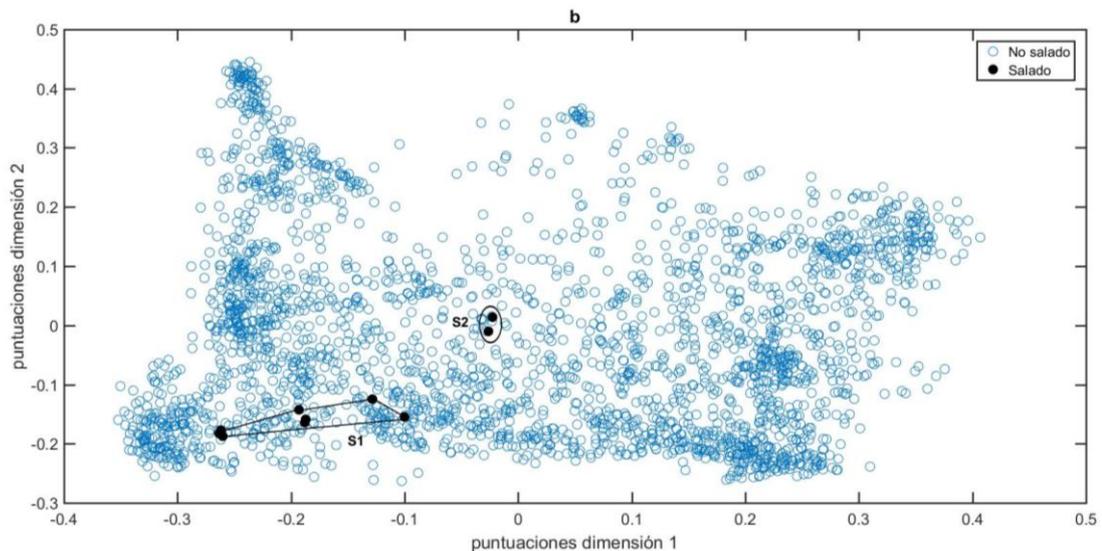


Figura 9. a) Espacio químico definido mediante el MDS para el gusto salado. b) Proyección de las moléculas de gusto salado y no salado en el espacio químico.

Para el gusto salado se han identificado únicamente dos grupos, en el S1 se observan compuestos que tienen cierta dispersión entre sí. En este grupo se encuentran los compuestos Orn- β Ala.HCl, Orn- γ Abu.HCl, ácido L-ornitil- γ -aminobutírico, Orn-Tau.HCl, Lys-Tau.HCl, L-ornitiltaurina, derivado 4 de disulfamato y derivado 8 de disulfamato. Existe una superposición con compuestos que en su mayoría son amargos, pero también unos pocos compuestos de la clase dulce y umami. Por otra parte, el grupo S2 está conformado por las sales cloruro de sodio y cloruro de potasio. Estos compuestos tienen en común la presencia del anión cloruro (Cl⁻) y presentan superposición principalmente con compuestos amargos y algunos de la clase multi gusto (Figura 8a). La superposición de moléculas saladas con compuestos dulces puede ser explicada con el caso del cloruro de sodio (NaCl), el cual a bajas concentraciones (0.009 M) no presenta un gusto y, por el contrario, a concentraciones desde 0.010 M-0.030 M el NaCl evoca un gusto dulce. Además, los dos únicos grupos presentes de esta clase se encuentran ubicados en el tercer cuadrante

2.6.6. Espacio químico para la clase insípida

Para la clase insípida, se definió el espacio químico mediante la proyección de las dimensiones 1 y 2 del MDS (Figura 10a), donde se puede observar sobre posición de moléculas pertenecientes a distintos gustos. En la figura 10b se puede identificar con más claridad algunos grupos consistentes de compuestos que se diferencian de las otras clases.

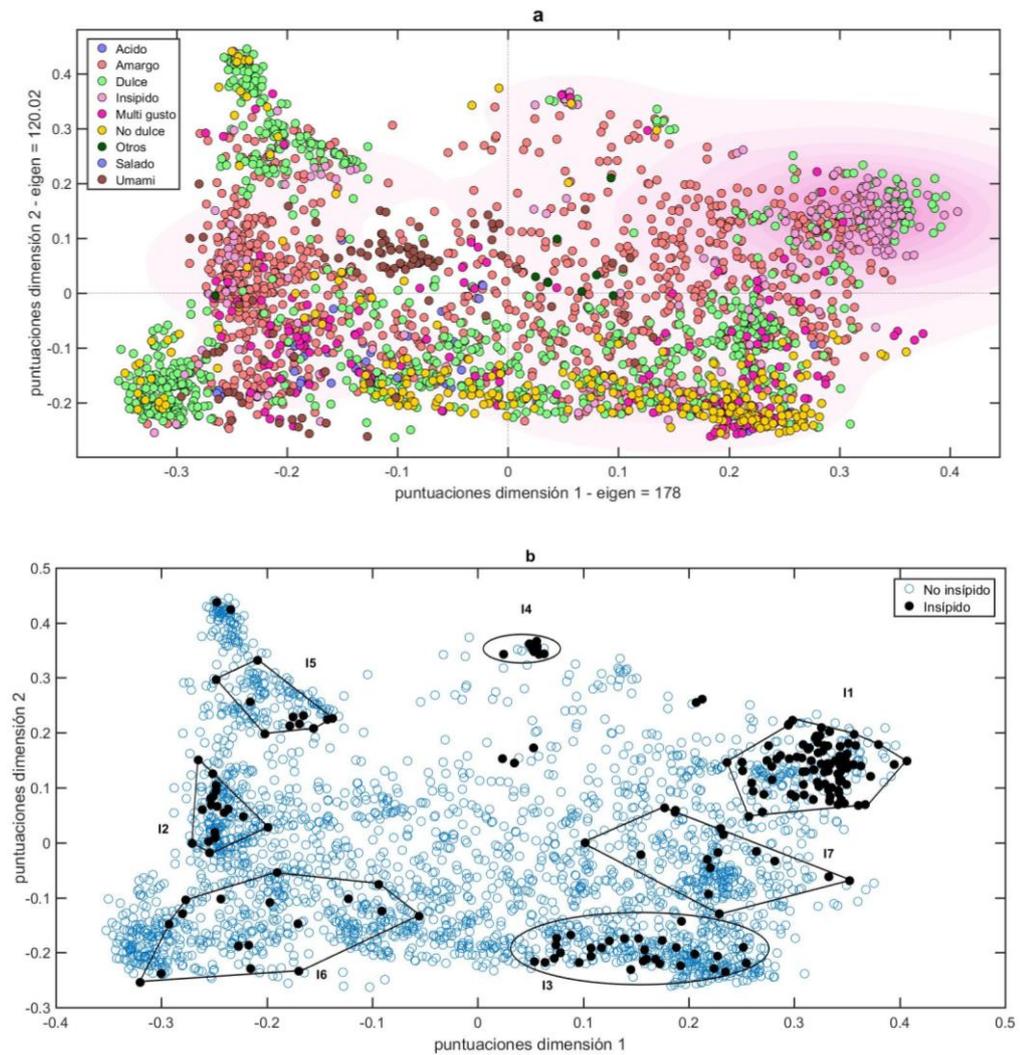


Figura 10. a) Espacio químico definido mediante el MDS para la clase insípida. b) Proyección de las moléculas de clase insípida y no insípida en el espacio químico.

El grupo I1 es un conjunto consistente de moléculas, con algunas de ellas un poco dispersas. Este grupo presenta superposición principalmente con moléculas de gusto amargo y dulce. Gran parte de las moléculas de este conjunto tienen en común la presencia del grupo 2-metoxifenol también conocido como guayaacol. El conjunto compacto de estas moléculas tiende a agruparse en el primer cuadrante del plano.

En el grupo I2 se puede observar moléculas menos dispersas en comparación a otras. La mayoría de estas estructuras tienen en común ser derivados de diterpenos. Este grupo muestra sobre posición significativa con moléculas amargas, y en menor proporción con compuestos multi gusto. Los compuestos de este grupo, en su mayoría se encuentran ubicados en el segundo cuadrante.

El grupo I3, ubicado en el cuarto cuadrante, está conformado por moléculas que presentan dispersión a lo largo de la dimensión uno, aunque entre algunas de ellas se observa cierta cercanía. Este grupo de compuestos insípidos se superpone de manera clara con estructuras

químicas pertenecientes a la clase dulce, multi gusto y amargo. Las moléculas que constituyen el conjunto I3, en su mayoría, son derivados de la sacarina y del suosan.

El conjunto I4 resulta ser pequeño y compacto en comparación con el resto de grupos, por lo que muestra consistencia entre sus moléculas. De hecho, la mayoría de las moléculas son derivados de la hesperetina dihidrocalcona y neohesperidina dihidrocalcona. Este grupo de estructuras se superponen con compuestos de clase dulce, multi gusto y no dulce.

El grupo I5 evidencia una superposición con compuestos principalmente dulces y amargos. Las estructuras pertenecientes a este conjunto tienen en común ser derivados del Telosmoside A.

Por otra parte, el grupo I6 es el más disperso de todos con superposición con la clase insípida y compuestos de los gusto amargo, dulce, umami y multi gusto. Este conjunto, que se ubica en el tercer cuadrante, tiene la característica común de ser aminoácidos (como la leucina y ácido aspártico) o derivados de ellos.

Finalmente, el grupo I7 también presenta alta dispersión y superposición con compuestos dulces, amargos y multi gustos. Este conjunto de moléculas está distribuido en el primer y cuarto cuadrante del MDS, los cuales en su mayoría presentan fragmentos de anisol y carboxibenzaldehído.

2.6.7. Otros espacios químicos

Durante el desarrollo de la base de datos se han identificado moléculas cuya etiqueta (clase) es "Multi gusto". A esta clase pertenecen los compuestos que cumplen una de las siguientes características:

1. Moléculas que en las fuentes primarias consultadas estaban etiquetadas con dos o más gustos, por ejemplo, citral (Amargo/Dulce), oxalato ácido de potasio (Ácido/Amargo), D-cisteína (Dulce/Amargo/Ácido), cloruro de bario dihidratado (Amargo/Salado).
2. Compuestos que luego de la fusión y filtrado de la base de datos, presentan dos gustos distintos en las fuentes primarias. Por ejemplo, el cloruro de calcio es amargo (Dagan-Wiener et al., 2019) y salado (Wong, 2018); mientras que el ácido láctico es ácido (Shallenberger, 1993) y umami (Suess et al., 2015).
3. Estructuras moleculares que presentan más de dos gustos distintos en diferentes fuentes primarias. Por ejemplo, el compuesto L-asparagina es insípido (Belitz et al., 2009) , no dulce (Wong, 2018) y multi gusto (Shallenberger, 1993).

Para definir el espacio químico para este tipo de compuestos se ha usado la proyección de las puntuaciones de las dimensiones 2 y 4 del escalado multidimensional (Figura 11). Se observa que la mayoría de compuestos multi gusto se agrupan en la región de puntuaciones negativas para la segunda componente (cuadrantes 2 y 3). Este tipo de moléculas se

superponen con compuestos de los gustos primarios, así como compuestos no dulces e insípidos. La superposición de estas estructuras con compuestos de los gustos primarios se puede deber a que durante las evaluaciones sensoriales mediante paneles de catación, los panelistas registran el sabor predominante, así como otros gustos primarios que pueden ser percibidos, o gustos remanentes (regustos).

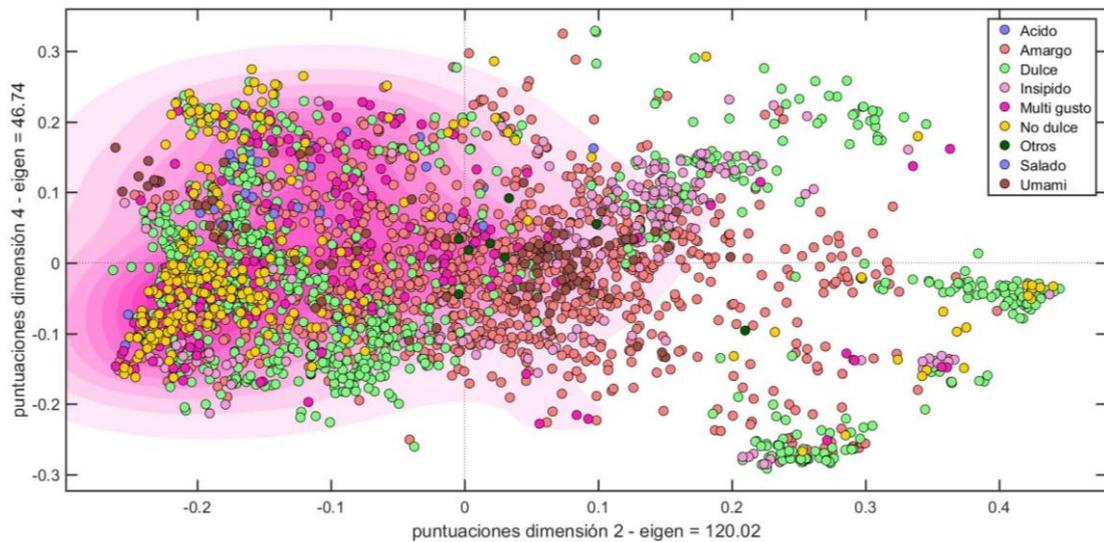


Figura 11. Espacio químico para los compuestos multi gusto definido por las puntuaciones del MDS de las dimensiones 2 y 4.

Por otra parte, otras moléculas se han asignado a la clase “No dulce”. En la Figura 12 se muestra la proyección de las puntuaciones de las dimensiones 2 y 3 donde se evidencia que estas moléculas tienden a agruparse mayoritariamente en el segundo cuadrante. Se evidencia que este grupo de moléculas es bastante compacto y presenta superposición con algunos compuestos de otros gustos. Adicionalmente, existen otras moléculas no dulces que se encuentran dispersas a lo largo del espacio químico definido, igualmente superpuestas con compuestos de los otros cinco gustos básicos.

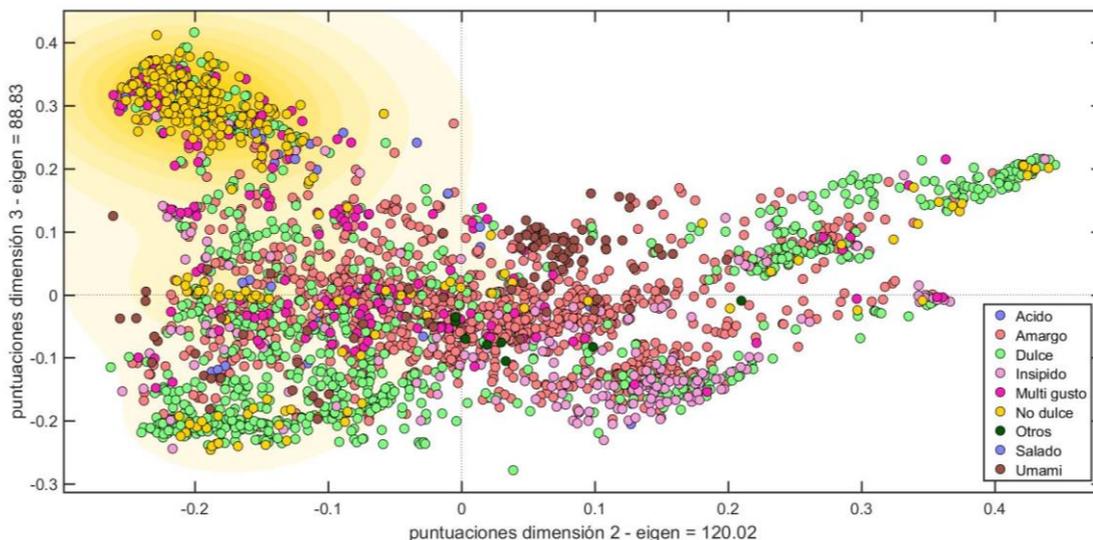


Figura 12. Espacio químico para los compuestos no dulces definido por las puntuaciones del MDS de las dimensiones 2 y 3.

Finalmente, en la base de datos desarrollada se encuentran ocho compuestos que han sido catalogados como “otros” (ácido eláidico, piperanina, piperilina, gingerol, shogaol, dihidrocapsaicina, nordihidrocapsaicina, epímero de selliguaeína A). En la Figura 13 se muestra la proyección de las puntuaciones de las dimensiones 3 y 5 donde se observa que los compuestos de esta clase tienden a agruparse en el segundo cuadrante. Este grupo de moléculas es poco numeroso y muestra cierta dispersión, además presenta superposición con compuestos mayoritariamente amargos, y otros dulces y multi gusto en menor proporción. En esta clase se han incluido aquellas estructuras químicas que presenten un gusto quemante o astringente.

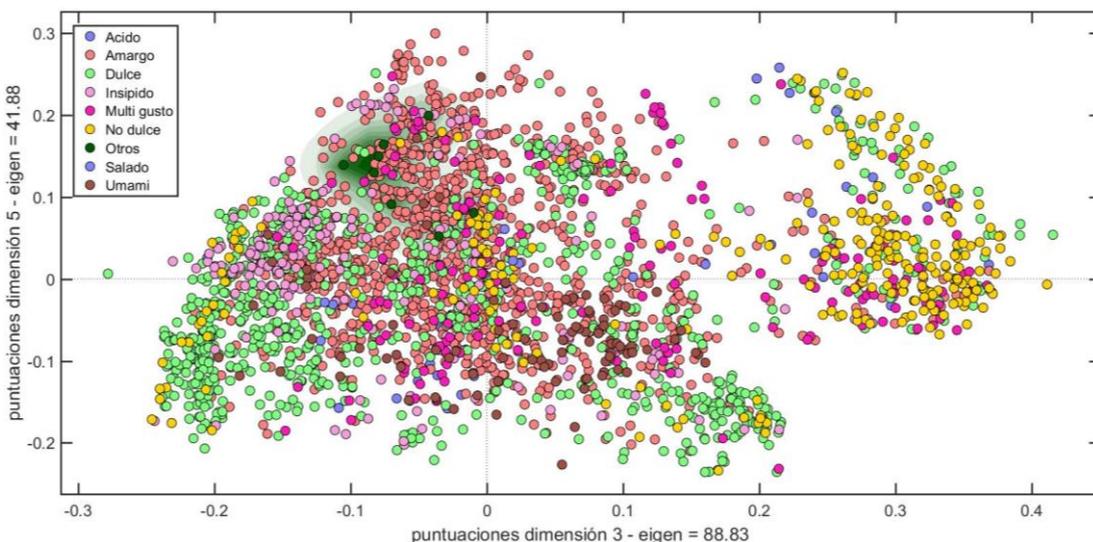


Figura 13. Espacio químico para los compuestos que presentan otro gusto diferente a los primarios definido por las puntuaciones del MDS de las dimensiones 3 y 5.

CONCLUSIONES

En este estudio se ha logrado construir una base de datos conformada por compuestos orgánicos e inorgánicos obtenidos de diferentes fuentes bibliográficas. La base de datos está constituida por 3589 compuestos químicos que representan a los gustos básicos: dulce, amargo, ácido, salado y umami, pero también está presente la clase insípida, multi gusto, no dulce y otros (astringente y quemante). El curado de las moléculas y el filtrado de la base de datos han sido aspectos fundamentales para obtener la base de datos depurada. Para cada gusto, el espacio químico ha sido definido mediante el escalado multidimensional. Los diferentes mapas químicos mostraron superposición entre los diferentes gustos, y esto puede ser causado cuando el compuesto es evaluado en distintas concentraciones, lo que hace variar el gusto. No obstante, ha sido posible identificar algunos grupos de alta similitud molecular para cada gusto.

REFERENCIAS BIBLIOGRÁFICAS

- Ahmed, J., Preissner, S., Dunkel, M., Worth, C. L., Eckert, A., & Preissner, R. (2010). SuperSweet—a resource on natural and artificial sweetening agents. *Nucleic acids research*, 39(suppl_1), D377-D382.
- Alvascience. (2020a). alvaDesc (Software for molecular descriptors calculation) version 2.0.3, <https://www.alvascience.com>.
- Alvascience. (2020b). alvaMolecule (Software to View and Prepare Chemical Datasets) version 1.0.4, <https://www.alvascience.com>.
- Ballabio, D. (2015). A MATLAB toolbox for Principal Component Analysis and unsupervised exploration of data structure. *Chemometrics and Intelligent Laboratory Systems*, 149, 1-9.
- Banerjee, P., & Preissner, R. (2018). BitterSweetForest: a random forest based binary classifier to predict bitterness and sweetness of chemical compounds. *Frontiers in Chemistry*, 6, 93.
- Bassoli, A., Drew, M. G. B., Hattotuwigama, C. K., Merlini, L., Morini, G., & Wilden, G. R. H. (2001). Quantitative structure-activity relationships of sweet isovanillyl derivatives. *Quantitative Structure-Activity Relationships*, 20(1), 3-16.
- Belitz, H.-D., Grosch, W., & Schieberle, P. (2009). *Food Chemistry*: Springer.
- Berthold, M. R., Cebron, N., Dill, F., Gabriel, T. R., Kötter, T., Meinl, T., . . . Wiswedel, B. (2009). KNIME-the Konstanz information miner: version 2.0 and beyond. *ACM SIGKDD explorations Newsletter*, 11(1), 26-31.
- Bleiholder, J., & Naumann, F. (2009). Data fusion. *ACM computing surveys (CSUR)*, 41(1), 1-41.
- Bouysset, C., Belloir, C., Antonczak, S., Briand, L., & Fiorucci, S. (2020). Novel scaffold of natural compound eliciting sweet taste revealed by machine learning. *Food chemistry*, 324(126864), 1-6.
- Cassotti, M., Consonni, V., Mauri, A., & Ballabio, D. (2014). Validation and extension of a similarity-based approach for prediction of acute aquatic toxicity towards daphnia magna. *SAR and QSAR in Environmental Research*, 25(12), 1013-1036.
- Chandrashekar, J., Hoon, M. A., Ryba, N. J., & Zuker, C. S. (2006). The receptors and cells for mammalian taste. *Nature*, 444(7117), 288-294.
- da Conceicao Neta, E. R., Johanningsmeier, S., Drake, M., & McFeeters, R. (2007). A chemical basis for sour taste perception of acid solutions and fresh-pack dill pickles. *Journal of Food Science*, 72(6), S352-S359.
- Dagan-Wiener, A., Di Pizio, A., Nissim, I., Bahia, M. S., Dubovski, N., Margulis, E., & Niv, M. Y. (2019). BitterDB: taste ligands and receptors database in 2019. *Nucleic acids research*, 47(D1), D1179-D1185.
- Damodaran, S., Parkin, K. L., & Fennema, O. R. (2008). *Fennema's food chemistry* (Fourth ed.): CRC press.

- Daylight Chemical Information Systems Inc. (2011). *Daylight theory manual*.
- Dickerson, R. E. (1992). *Principios de química*: Reverté.
- Fourches, D., Muratov, E., & Tropsha, A. (2010). Trust, but Verify: on the Importance of Chemical Structure Curation in Cheminformatics and QSAR Modeling Research. *Journal of chemical information and modeling*, 50(7), 1189-1204.
- Ghirri, A., & Bignetti, E. (2012). Occurrence and role of umami molecules in foods. *International Journal of Food Sciences and Nutrition*, 63(7), 871-881.
- Hypercube Inc. HyperChem™ Professional version 8.0. <http://www.hyper.com>.
- Johanningsmeiner, S. D., McFeeters, R. F., & Drake, M. (2005). A hypothesis for the chemical basis for perception of sour taste. *Journal of Food Science*, 70(2), R44-R48.
- Jolliffe, I. T. (2002). *Principal component analysis* (Second ed.): Springer.
- Kelly, D. P., Spillane, W. J., & Newell, J. (2005). Development of structure– taste relationships for monosubstituted phenylsulfamate sweeteners using classification and regression tree (CART) analysis. *Journal of Agricultural and Food Chemistry*, 53(17), 6750-6758.
- Kim, S., Thiessen, P. A., Bolton, E. E., Chen, J., Fu, G., Gindulyte, A., . . . Shoemaker, B. A. (2016). PubChem substance and compound databases. *Nucleic acids research*, 44(D1), D1202-D1213.
- Kruskal, J. B. (1964). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29(1), 1-27.
- Linstrom, P. J., & Mallard, W. G. (2001). The NIST Chemistry WebBook: A chemical data resource on the internet. *Journal of Chemical & Engineering Data*, 46(5), 1059-1063.
- Manfugás, J. E. (2020). *Evaluación sensorial de los alimentos*: Editorial Universitaria (Cuba).
- Marquardt, W., Morbach, J., Wiesner, A., & Yang, A. (2009). *OntoCAPE: A re-usable ontology for chemical process engineering*: Springer Science & Business Media.
- Mauri, A. (2020). alvaDesc: A tool to calculate and analyze molecular descriptors and fingerprints. In K. Roy (Ed.), *Ecotoxicological QSARs* (pp. 801-820): Humana Press.
- Mauri, A., Consonni, V., & Todeschini, R. (2017). Molecular descriptors. In J. Leszczynski, A. Kaczmarek-Kedziera, T. Puzyn, M. G. Papadopoulos, H. Reis, & M. K. Shukla (Eds.), *Handbook of Computational Chemistry* (Second ed., Vol. 3, pp. 2065-2093). Switzerland: Springer International Publishing.
- Medina-Franco, J. L., Martinez-Mayorga, K., Giulianotti, M. A., Houghten, R. A., & Pinilla, C. (2008). Visualization of the chemical space in drug discovery. *Current Computer-Aided Drug Design*, 4(4), 322-333.
- Morini, G., Bassoli, A., & Borgonovo, G. (2011). Molecular modelling and models in the study of sweet and umami taste receptors. A review. *Flavour and fragrance journal*, 26(4), 254-259.
- O'Donnell, T. (2008). *Design and use of relational databases in chemistry*: CRC Press.
- Pence, H. E., & Williams, A. (2010). ChemSpider: an online chemical information resource. In: ACS Publications.

- Rogers, D., & Hahn, M. (2010). Extended-connectivity fingerprints. *Journal of chemical information and modeling*, 50(5), 742-754.
- Rojas, C., Todeschini, R., Ballabio, D., Mauri, A., Consonni, V., Tripaldi, P., & Grisoni, F. (2017). A QSTR-based expert system to predict sweetness of molecules. *Frontiers in Chemistry*, 5(53), 1-12.
- Ruan, D., & Zeng, X. (2013). Intelligent sensory evaluation: An introduction. In D. Ruan & X. Zeng (Eds.), *Intelligent sensory evaluation: methodologies and applications*. Berlín: Springer Science & Business Media.
- Ruddigkeit, L., & Reymond, J.-L. (2014). The chemical space of flavours. In K. Martinez-Mayorga & J. L. Medina-Franco (Eds.), *Foodinformatics: Applications of chemical information to food chemistry* (pp. 83-96): Springer.
- Shallenberger, R. S. (1993). *Taste chemistry*: Springer Science & Business Media.
- Suess, B., Festring, D., & Hofmann, T. (2015). Umami Compounds and Taste Enhancers. In J. K. Parker, J. S. Elmore, & L. Methven (Eds.), *Flavour Development, Analysis and Perception in Food and Beverages* (pp. 331-351). Cambridge, UK: Elsevier.
- Todeschini, R., Ballabio, D., Cassotti, M., & Consonni, V. (2015). N3 and BNN: Two New Similarity Based Classification Methods in Comparison with Other Classifiers. *Journal of chemical information and modeling*, 55(11), 2365-2374.
- Todeschini, R., Ballabio, D., & Consonni, V. (2015). Distances and other dissimilarity measures in chemometrics. *Encyclopedia of Analytical Chemistry*, 1-34.
- Tuwani, R., Wadhwa, S., & Bagler, G. (2019). BitterSweet: Building machine learning models for predicting the bitter and sweet taste of small molecules. *Scientific reports*, 9(1), 1-13.
- Upadhyaya, J., Singh, N., Bhullar, R., & Chelikani, P. (2017). Biochemistry of human bitter taste receptors. In M. Aliani & M. N. Eskin (Eds.), *Bitterness: perception, chemistry and food processing* (pp. 3-20). Manitoba, Canada: John Wiley & Sons.
- Voigt, K. (2003). Databases on Environmental Information. In J. Gasteiger (Ed.), *Handbook of Chemoinformatics: From Data to Knowledge in 4 Volumes* (Vol. 2, pp. 722-742). Germany: WILEY-VCH Verlag GmbH & Co. KGaA.
- Weininger, D. (1988). SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of chemical information and computer sciences*, 28(1), 31-36.
- Weininger, D., Weininger, A., & Weininger, J. L. (1989). SMILES. 2. Algorithm for generation of unique SMILES notation. *Journal of chemical information and computer sciences*, 29(2), 97-101.
- Winsberg, S., & Carroll, J. D. (1989). A quasi-nonmetric method for multidimensional scaling via an extended Euclidean model. *Psychometrika*, 54(2), 217-229.
- Wold, S., Esbensen, K., & Geladi, P. (1987). Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 2(1-3), 37-52.

- Wold, S., Sjöström, M., & Eriksson, L. (2001). PLS-Regression: A Basic Tool of Chemometrics. *Chemometrics and Intelligent Laboratory Systems*, 58(2), 109-130.
- Wong, D. W. (2018). *Mechanism and theory in food chemistry* (Vol. 2). USA: Springer.
- Yamaguchi, S. (1979). The umami taste. In J. C. Boudreau (Ed.), *Food taste chemistry*. Washington, DC: American Chemical Society.
- Yamaguchi, S., & Ninomiya, K. (2000). Umami and food palatability. *The Journal of nutrition*, 130(4), 921S-926S.