



UNIVERSIDAD NACIONAL AUTÓNOMA DE MEXICO
PROGRAMA DE MAESTRÍA Y DOCTORADO EN CIENCIAS QUÍMICAS

**ESTUDIO CUANTITATIVO DE RELACIONES ESTRUCTURA ACTIVIDAD DE
SUSTANCIAS QUÍMICAS ALIMENTARIAS**

TESIS

PARA OPTAR POR EL GRADO DE

MAESTRA EN CIENCIAS

PRESENTA

QFB KARINA EURÍDICE JUÁREZ MERCADO

DR. JOSÉ LUIS MEDINA FRANCO
FACULTAD DE QUÍMICA, UNAM

CIUDAD UNIVERSITARIA, CDMX. ABRIL, 2024.



Universidad Nacional
Autónoma de México



UNAM – Dirección General de Bibliotecas
Tesis Digitales
Restricciones de uso

DERECHOS RESERVADOS ©
PROHIBIDA SU REPRODUCCIÓN TOTAL O PARCIAL

Todo el material contenido en esta tesis esta protegido por la Ley Federal del Derecho de Autor (LFDA) de los Estados Unidos Mexicanos (México).

El uso de imágenes, fragmentos de videos, y demás material que sea objeto de protección de los derechos de autor, será exclusivamente para fines educativos e informativos y deberá citar la fuente donde la obtuvo mencionando el autor o autores. Cualquier uso distinto como el lucro, reproducción, edición o modificación, será perseguido y sancionado por el respectivo titular de los Derechos de Autor.



UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

PROGRAMA DE MAESTRÍA Y DOCTORADO EN CIENCIAS QUÍMICAS

**ESTUDIO CUANTITATIVO DE RELACIONES ESTRUCTURA ACTIVIDAD DE
SUSTANCIAS QUÍMICAS ALIMENTARIAS**

**T E S I S
PARA OPTAR POR EL GRADO DE**

MAESTRA EN CIENCIAS

P R E S E N T A

QFB KARINA EURÍDICE JUÁREZ MERCADO



Ciudad de México, Abril, 2024

SITIO DONDE SE DESARROLLÓ EL TEMA:

EL PRESENTE TRABAJO SE DESARROLLÓ EN EL GRUPO DE INVESTIGACIÓN DIFACQUIM DEL DEPARTAMENTO DE FARMACIA, UBICADO EN EL CUBÍCULO 117 DEL EDIFICIO F DE LA FACULTAD DE QUÍMICA DE LA UNAM.

TUTOR:



DR. JOSÉ LUIS MEDINA FRANCO

Agradecimientos

A José Luis Medina, por ser mi guía y compañero desde antes de empezar esta travesía. Gracias, por tu paciencia, comprensión y apoyo incondicional, por cada enseñanza, y momento compartido a tu lado.

A mis padres, Karina Mercado y Víctor Juárez, por su constante apoyo mientras seguía un camino incierto pero que me llenaba de felicidad. Gracias por animarme a perseguir mis sueños y a seguir mi corazón.

A mis hermanos Elizabeth Juárez y Jorge Juárez, por siempre creer en mí.

A mis abuelos Guadalupe Medrano, Lilia Velázquez y Jorge Juárez por consentirme, cuidarme, amarme y apoyarme en todo desde siempre.

A Norma Velázquez y Victoria Zurita por su cariño en cada momento vivido a su lado.

A mis amigos, en especial a Bárbara Díaz, Karen Montiel, Araceli Flores, Silvia Sánchez y René Flores. Gracias por su constante ánimo, por no juzgarme en mis locuras y apoyar mis decisiones.

A los miembros del jurado: Dr. Enrique García Hernández, Dr. Francisco Hernández Luis, Dr. Samuel Enoch Estrada Soto, Dra. Carmina Montiel Pacheco y Dr. Luis Jesús Córdova Bahena por su disposición para participar en la revisión y por sus valiosos comentarios sobre este trabajo de tesis.

A la Universidad Nacional Autónoma de México por ser más que una institución educativa y convertirse en un segundo hogar para mí durante más de 10 años. Un lugar dónde adquirí mucho conocimiento, pero también y me dio la oportunidad de crecer como persona en muchos aspectos y conocer personas increíbles.

Al Programa de Maestría y Doctorado en Ciencias Químicas por darme la oportunidad de formarme como Maestra en Ciencias Químicas.

Al grupo DIFACQUIM por recibirme en repetidas ocasiones con sus puertas abiertas, por compartir su conocimiento y permitirme conocer a personas que, además de compañeros, se convirtieron en amigos. Especialmente a Fernanda Saldívar, Diana Prado, Alejandro Gómez, Naicolette Agudo, Edgar López y Juan Felipe Avellaneda por su lealtad, consejos y todos los momentos compartidos.

A Juan Felipe Avellaneda y Hassan Villegas por sus contribuciones académicas en este proyecto.

Al Programa de Apoyo a Proyectos de Investigación e Innovación Tecnológica (PAPIIT) de la Dirección General de Asuntos del Personal Académico de la UNAM, proyecto PAPIIT IN201321.

Al Consejo Nacional de Humanidades, Ciencias y Tecnologías (CONAHCyT) por la beca con número 893849.

*Phantasie ist wichtiger als Wissen, denn Wissen ist begrenzt,
und Phantasie umfasst die ganze Welt.*

Albert Einstein.

Los resultados de este trabajo fueron presentados en:

➤ *Fall meeting 2023 of the American Chemical Society (ACS).* San Francisco California, Estados Unidos de América. Del 13 al 17 de agosto. Presentación oral en la división *Agricultural and Food Chemistry* durante el simposio *Advances in Food Chemical Informatics, Knowledge Bases and Databases*. Título de la presentación: “*Food Chemicals in Epigenetic Targets: Towards an Epi Food Chemical Database.*”

➤ Juárez-Mercado, K. E.; Avellaneda-Tamayo, J. F.; Villegas-Quintero, H.; Chávez-Hernández, A. L.; López-López, C. D.; Medina-Franco, J. L. *Food Chemicals and Epigenetic Targets: Assembling an Epi Food Chemical Database.* *ChemRxiv* **2023**. DOI: 10.26434/chemrxiv-2023-drxvm-v2 (and under peer-review).

Índice

1. Abreviaturas.....	10
2. Resumen.....	15
3. Antecedentes.....	17
3.1 Epigenética.....	17
3.2 Mecanismos de regulación.....	17
3.2.1 Metilación del ADN.....	18
3.2.1.1 Silenciamiento génico.....	19
3.2.1.2 Estabilidad genómica.....	19
3.2.2 Modificación de aminoácidos en las histonas.....	19
3.2.3 Por moléculas de ARN no codificante (ncRNA).....	21
3.2.3.1 ARN no codificante largo (lncRNA).....	21
3.2.3.2 ARNs no codificantes pequeños (siRNA y miRNA).....	22
3.2.3.3 ARN circular (circRNA).....	24
3.2.4 Modificaciones postraduccionales.....	25
3.2.4.1 Acetilación de histonas.....	25
3.2.4.2 Metilación de histonas.....	26
3.2.4.3 Fosforilación de histonas.....	27
3.2.4.4 Ubiquitinación de histonas.....	27
3.3 Dianas epigenéticas.....	28
3.3.1 Escritoras.....	29
3.3.1.1 DNMTs.....	29
3.3.1.2 EZH2.....	30
3.3.1.3 PRMT1.....	30
3.3.1.4 HATs.....	31
3.3.2 Borradoras.....	32
3.3.2.1 HDACs.....	32
3.3.2.2 SIRTs.....	33
3.3.2.3 LSD1.....	34
3.3.2.4 KDMs.....	35
3.3.3 Lectoras.....	36
3.3.3.1 BET/BRD4.....	36
3.4 Enfermedades asociadas.....	37
3.5 Nutri epigenética.....	44
3.6 Quimiinformática.....	45
3.6.1 Bases de datos moleculares.....	46
3.6.2 Núcleos estructurales base (<i>molecular scaffolds</i>).....	47
3.6.3 Espacio químico y su representación visual.....	48

3.6.4 Relaciones cuantitativas estructura-actividad.....	49
3.6.4.1 Panoramas de actividad.....	52
3.6.4.2 Acantilados de actividad (<i>activity cliffs</i>).....	54
4. Planteamiento del problema.....	56
5. Hipótesis.....	58
6. Objetivos.....	59
6.1 General.....	59
6.2 Particulares.....	59
7. Metodología.....	60
7.1 Búsqueda en la literatura de artículos indizados.....	60
7.2 Creación de la base de datos de compuestos químicos presentes en alimentos y productos naturales con actividad epigenética.....	61
7.3 Análisis quimioinformático de la base de datos <i>Epi Food Chemical Database</i>	62
7.3.1 Contenido y diversidad química.....	62
7.3.2 Visualización del espacio químico.....	62
7.3.3 Perfil de relación estructura-actividad epigenética.....	63
8. Resultados y discusión.....	67
8.1 Análisis de la literatura.....	67
8.2 <i>Epi Food Chemical Database</i>	67
8.3 Análisis quimioinformático.....	70
8.3.1 Contenido y diversidad química.....	71
8.3.1.1 Análisis de núcleos base (<i>molecular scaffolds</i>).....	71
8.3.1.2 Curva de recuperación cíclica del sistema.....	72
8.3.2 Visualización de espacio químico.....	73
8.3.3 Perfil epigenético.....	75
8.3.3.1 ECFP4.....	76
8.3.3.2 ECFP6.....	82
8.3.3.3 MACCS Keys.....	88
8.3.3.4 RDKit.....	94
8.3.3.5 Pares de compuestos consenso con relación estructura-actividad epigenética.....	101
8.3.3.6 Pares de compuestos consenso potenciales acantilados de actividad (<i>activity cliffs</i>).....	106
9. Conclusiones.....	111
10. Perspectivas.....	112
11. Referencias.....	113
12. Anexos.....	129

1. Abreviaturas

-CH₃: grupo metilo

-CH₃CO: grupo acetilo

-PO₄: grupo fosfato

ABC: área bajo la curva

AcCoA: acetil coenzima A

ADN: ácido desoxirribonucleico

AKT1: gen de la AKT serina/treonina quinasa 1

AMI-1: gen de la amidasa 1

APC: gen del regulador APC de la vía de señalización WNT

ARN: ácido ribonucleico

ATP: adenosin trifosfato

ATP2A3: gen de la ATPasa del retículo sarcoplásmico/endoplásmico Ca²⁺ transportador 3

BET/BRD4: proteína 4 de la familia del dominio y bromodominio extra-terminal

Brca1: gen del cáncer de mama 1, aparición temprana

Brca2: gen del cáncer de mama 2, aparición temprana

CAV1: gen de la caveolina 1

CBP: gen de la proteína sarcoplásmica fijadora de calcio

CCND1: gen de la ciclina D1

cdk4: gen de la quinasa 4 dependiente de ciclina

CDKN2A: gen del inhibidor de la quinasa 4 dependiente de ciclina 2A

CDK2AP1: gen de la proteína 1 asociada a la quinasa 2 dependiente de ciclina

CDK6: gen de la quinasa 6 dependiente de ciclina

CDKN1A: gen inhibidor de la quinasa dependiente de ciclina 1A

CDKN1B: gen inhibidor de la quinasa dependiente de ciclina 1B

CDKN1C: gen inhibidor de la quinasa dependiente de ciclina 1B

CEACAM5: gen de la molécula 5 de adhesión celular CEA

CI50: concentración inhibitoria 50

circRNA: molécula de ARN que forma una estructura circular cerrada

COCONUT: COllection of Open Natural ProDUcTs

CoREST: correpresor del factor de transcripción silenciador

CpG: citosinas seguidas de una guanina

CSR: recuperación del sistema cíclico

DAPK1: gen de la proteína quinasa 1 asociada a la muerte

DHRS3: gen de la deshidrogenasa/reductasa 3

DNMTs: metiltransferasas de ADN

DOI: números de identificación de objetos digitales

DUB: enzimas deubiquitinasas

E2F1: gen del factor de transcripción E2F 1

E3: ligasas de ubiquitina

ECFP: *Extended Connectivity Fingerprint*

ECFP4: huella digital molecular *Extended Connectivity Fingerprint* de radio 2

ECFP6: huella digital molecular *Extended Connectivity Fingerprint* de radio 3

EIF4E: gen del factor de iniciación de la traducción eucariota 4E

EP300: gen de la proteína p300 de unión a E1A

Epi Food Chemical Database: base de datos de compuestos químicos presentes en alimentos y productos naturales con actividad epigenética reportada en la literatura del 2017-2023.

ER: gen de la proteína rica en leucina de la familia de la proteína quinasa tipo-receptor

ERK1: gen de la quinasa 1 regulada por señales extracelulares

ER α : gen del receptor de estrógeno alfa

ER β : gen del receptor de estrógeno beta

EZH2: potenciador del homólogo 2 de zeste

EZH2: gen del potenciador de la subunidad 2 del complejo represivo polycomb zeste 2

FAD: dinucleótido de flavina-adenina

FooDB: base de datos que contiene información sobre compuestos químicos presentes en los alimentos, por el Instituto de Investigación de Informática Médica de la Universidad de Alberta

FOXO1: gen de la caja de horquilla O1

G9a: gen de la histona eucromática lisina metiltransferasa 2

GATA4: gen de la proteína 4 de unión a GATA

H3K27: residuo de lisina 27 de las histonas H3
H3K27me3: metilación del residuo de lisina 27 de las histonas H3
H3K4: lisina 4 en la histona H3
HAI-2/PB: gen del inhibidor de la serina peptidasa, Kunitz tipo 2
HATs: acetiltransferasas de histonas
HBA: átomos aceptores de hidrógeno
HBD: átomos donadores de hidrógeno
HDAC: deacetilasas de histonas
HDM: demetilasas de histonas
HER2: gen de la subunidad de glutamil-ARNt(Gln) amidotransferasa HER2
HIF-1 α : gen de la subunidad alfa del factor 1 inducible por hipoxia
HIN-1: gen del miembro 1 de la familia de la secretoglobina
HMT: metiltransferasas de histonas
HOXB6: gen de la proteína homeobox B6
IDs: identificadores
I κ B α : gen del inhibidor alfa de NF κ B
IMPDH1: gen de la inosina monofosfato deshidrogenasa 1
InChI: Identificador Químico Internacional
InChIKey: versión hash del Identificador Químico Internacional
KAT2B: gen de la acetiltransferasa de lisina 2B
KDMs: desmetilasas de lisina
KRAS: protooncogén KRAS, GTPasa
LANaPDB: *Latin American Natural Product Database*
lncRNA: ARN largos no codificantes
logP: coeficiente de partición octanol-agua
LSD1: desmetilasa específica de lisina 1
MACCS Keys: huella digital molecular *Molecular ACCess System keys*
MAPK1: gen de la proteína quinasa 1 activada por mitógenos
MICA: gen de la secuencia A relacionada con el polipéptido MHC de clase I
miRNA: micro ARNs
MITF-M: gen del factor de transcripción inductor de melanocitos
MKRN3: gen de la proteína 3 del dedo angular de makorin

MMP-2: gen de la metalopeptidasa de matriz 2
mRNA: ARN mensajero
MYC: protooncogén MYC, factor de transcripción bHLH
Myc ZNF278: gen coactivador de Myc
NAD⁺: nicotinamida adenina dinucleótido
ncRNA: ARN no codificante
NM1: gen interactador de N-myc y STAT
NQO1: gen de NAD(P)H quinona deshidrogenasa 1
NR4A1: gen de la subfamilia de receptores nucleares 4 grupo A miembro 1
NTRK2: gen del receptor neurotrófico tirosina quinasa 2
p-TEFb: complejo de Elongación de la Transcripción Positivo del Factor b
CDKN1A: gen inhibidor de la quinasa dependiente de ciclina 1A
PCA: análisis de componentes principales
PCD4: supresor de tumores de la muerte celular programada 4
PDX1: gen de la proteína homeobox 1 pancreático y duodenal
PHD2: gen del factor 1 inducible por hipoxia familiar egl-9
PI3K: gen fosfatidilinositol 3-quinasa, putativa
PR: receptores de progesterona
pRB: gen del correpresor transcripcional RB1
PRC2: complejo polimerasa asociado a la metiltransferasa 2
PRMT1: proteína arginina metiltransferasa 1
PTEN: gen homólogo de la fosfatasa y tensina
QSAR: relaciones cuantitativas estructura-actividad
RARB: gen del receptor beta del ácido retinoico
Ras: protooncogén de la familia RAS (HRAS, KRAS y NRAS)
RASSF1A: gen del miembro 1 de la familia de las asociación Rassf1 Ras
RELA: protooncogén RELA, subunidad NF-kB
RDKit: huella digital molecular *RDKit Fingerprint*
RDKit: biblioteca de código abierto para quimioinformática
RECK: gen de la proteína rica en cisteína que induce reversión con motivos kazal
RGS16: gen del regulador de la señalización de la proteína G 16
RUNX3: gen del factor de transcripción 3 de la familia RUNX

SAH: S-adenosilhomocisteína

SALL3: gen del factor de transcripción tipo spalt 3

SAM: S-adenosilmetionina

SAR: relación estructura-actividad

SAS: similitud estructura-actividad

SE: similitud estructural

siRNA: ARN interferente pequeño

SIRT6: sirtuinas

SLIT2: gen supresor de tumor slit 2

SMILES: *Simplified Molecular Input-Line Entry System*

SOM: mapas autoorganizados

SPA: similitud del perfil de actividad

SUV39: gen de la histona lisina metiltransferasa SET/SUV39

t-SNE: *T-Distributed Stochastic Neighbor Embedding*

TIMP-3: gen del inhibidor de la metalopeptidasa TIMP 3

TMAP: gráficos de mapas de árbol

TMPRSS2:ERG: gen del factor de transcripción ERG de la proteasa transmembranal de serina

TPSA: área de superficie polar topológica

TP53: gen de la proteína tumoral p53

UNC13B: gen homólogo B de unc-13

WIF-1: gen del factor inhibidor WNT 1

WWOX: gen del dominio WW que contiene oxidoreductasa

2. Resumen

La epigenética se ha convertido en un campo de gran importancia para la comprensión de la patogénesis de enfermedades y el desarrollo de terapias innovadoras. El incremento en el número de publicaciones recientes refleja el interés en los procesos epigenéticos y su relación con compuestos químicos presentes en los alimentos. No obstante, hasta este momento, había faltado una investigación que recopilara los avances más recientes en esta área, enfocada en la interacción entre las estructuras químicas de los compuestos químicos alimentarios y su actividad biológica.

Este proyecto de investigación examina y discute la relación entre los compuestos químicos que se encuentran en los alimentos y en productos naturales que tienen reportada actividad epigenética. Se emplean herramientas quimioinformáticas para llevar a cabo una comparación cuantitativa a partir de análisis de diversidad estructural y espacio químico.

Un componente esencial de esta investigación fue la creación de la primera versión de una base de datos llamada *Epi Food Chemical Database*, que contiene los compuestos químicos encontrados en alimentos y productos naturales que tienen actividad epigenética reportada en artículos en revistas indizadas del 2017-2023. *Epi Food Chemical Database* contiene: información detallada de estos compuestos como su estructura química en notación SMILES, las dianas epigenéticas vs. las cuales presentan actividad biológica, la fuente natural en dónde se encuentran presentes, las respectivas notaciones lineales InChI e INChIKey, los identificadores (IDs) respectivos en bases de datos públicas (*COCONUT* y *FoodDB*), así como las referencias (usando DOIs) de los artículos indizados asociados.

El archivo respectivo de la base de datos *Epi Food Chemical Database* creada se encuentra disponible en libre acceso en el repositorio en línea: https://github.com/DIFACQUIM/Epi_food_Chemical_Database.



Figura 1. Resumen gráfico

3. Antecedentes

3.1 Epigenética

El concepto de epigenética ha experimentado una evolución desde que se introdujo por primera vez por Conrad Waddington en la década de 1940. En sus inicios se definía como “la rama de la biología que se ocupa de las interacciones causales entre los genes y sus productos, que dan lugar al fenotipo”.¹ En la actualidad la epigenética se entiende como el estudio de las alteraciones hereditarias en el perfil de expresión génica que no conllevan modificaciones en la secuencia de ácido desoxirribonucleico (ADN), pero que afectan la accesibilidad del código genético a través de diferentes mecanismos de regulación. Esto incluye cambios que ocurren durante toda la vida de un organismo a través de varios mecanismos que pueden activar o inactivar la expresión génica y así influir en el fenotipo.¹⁻³

Los cambios epigenéticos se pueden clasificar en tres categorías: la epigenética directa, que ocurre a lo largo de la vida de un individuo; la epigenética indirecta, que se refiere a las modificaciones que acontecen en el útero debido a eventos durante la gestación; y, por último, la epigenética transgeneracional, que abarca aquellos cambios que afectaron a generaciones anteriores y que de alguna manera se transmiten, posiblemente a través de modificaciones en los gametos o en el entorno intrauterino.²

Existen diversos mecanismos por los cuales se pueden regular las modificaciones epigenéticas, los cuales se describen en la siguiente sección.

3.2 Mecanismos de regulación

A continuación, se presentan en detalle los mecanismos epigenéticos clave que desempeñan un papel esencial en la regulación de la expresión génica. Algunos de estos mecanismos incluyen la metilación del ADN, la modificación de aminoácidos en las histonas, la metilación y acetilación de proteínas, la regulación a través de moléculas de ácido ribonucleico (ARN) no codificante, entre otras modificaciones postraduccionales.⁴

3.2.1 Metilación del ADN

La metilación de ADN es un proceso epigenético fundamental que implica la adición de grupos metilo (-CH₃) a las bases de ADN, especialmente a las citosinas que suelen estar seguidas de una guanina (CpG). Este proceso se encuentra regulado por enzimas específicas conocidas como metiltransferasas de ADN (DNMTs, por sus siglas en inglés), las cuales añaden grupos metilo en la posición 5 del anillo de citosina. En esta reacción, la S-adenosilmetionina (SAM) sirve como donador de los grupos metilos, y posteriormente a S-adenosilhomocisteína. La metilación del ADN desempeña un papel crucial en la regulación de la expresión génica, ya que afecta la estructura de la cromatina y, en consecuencia, la accesibilidad de las proteínas reguladoras al ADN.⁵⁻⁷ La Figura 2 muestra el mecanismo de metilación del ADN.

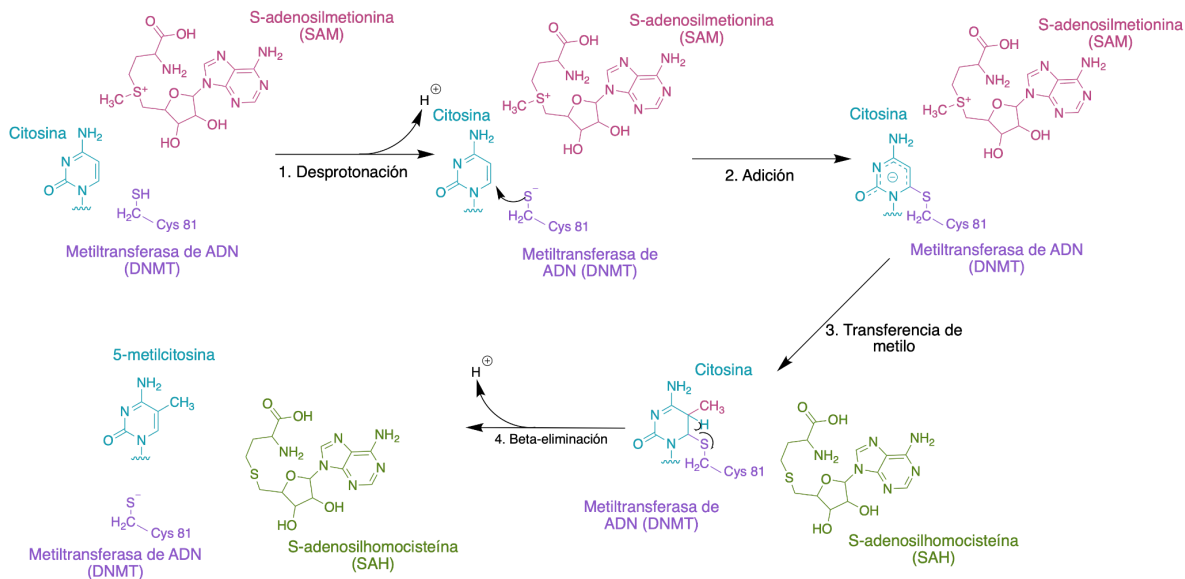


Figura 2. Mecanismo de metilación del ADN.⁵⁻⁷

La metilación del ADN es un proceso dinámico y reversible. Por lo que las marcas de metilación pueden agregarse o eliminarse a lo largo del tiempo en respuesta a cambios ambientales o señales intracelulares. Este proceso puede tener dos efectos principales en la expresión génica; el silenciamiento génico y la estabilidad genómica.⁸

3.2.1.1 Silenciamiento génico

Uno de los principales efectos de la metilación del ADN es el silenciamiento génico. Este efecto ocurre cuando ciertas regiones del ADN, como los promotores de los genes, se metilan, por lo que las proteínas reguladoras que normalmente se unirían a estas regiones presentan dificultades para hacerlo. Esto conlleva a tener la inhibición o represión de un gen, lo que significa que no podrá expresarse. Debido a esto, este mecanismo es indispensable para controlar la activación e inactivación de genes en un tejido o células específicas.⁹

3.2.1.2 Estabilidad genómica

La metilación del ADN también desempeña un rol importante en la estabilidad del genoma. Esta función incluye la supresión de elementos genéticos móviles, como los transposones que pueden causar daños en la secuencia de ADN y como consecuencia dañar la integridad del genoma. Por lo que al metilar ciertas regiones del ADN, se evita que pueda ocurrir la transposición de estos elementos y por consiguiente se mantiene la integridad del genoma.¹⁰

3.2.2 Modificación de aminoácidos en las histonas

Las histonas son proteínas altamente conservadas que desempeñan un papel fundamental en la estructura y regulación del ADN en las células. Estas proteínas actúan como “bobinas” alrededor de las cuales se enrolla la cadena de ADN, permitiendo que pueda ser compactado. Las cinco principales familias de histonas son H1 (~200 aminoácidos), H2A (~130 aminoácidos), H2B (~125 aminoácidos), H3 (~135 aminoácidos) y H4 (~102 aminoácidos). Estas histonas interactúan entre sí y con el ADN para formar el nucleosoma, la unidad básica de la cromatina.¹¹

Cada histona tiene una región rica en aminoácidos con carga positiva, lo que facilita su interacción con el ADN, el cual tiene una carga negativa debido a la presencia de grupos fosfatos. La unión de las histonas al ADN se produce a partir

de interacciones electrostáticas. Este arreglo compacto entre las histonas y el ADN es esencial para poder llevar a cabo el empaquetamiento del material genético en el núcleo de las células de manera eficiente y regular su acceso a este durante procesos como la replicación y transcripción.¹²⁻¹³ La Figura 3 ilustra en forma esquemática cómo se llevan a cabo las interacciones entre las histonas y el ADN para formar el nucleosoma.

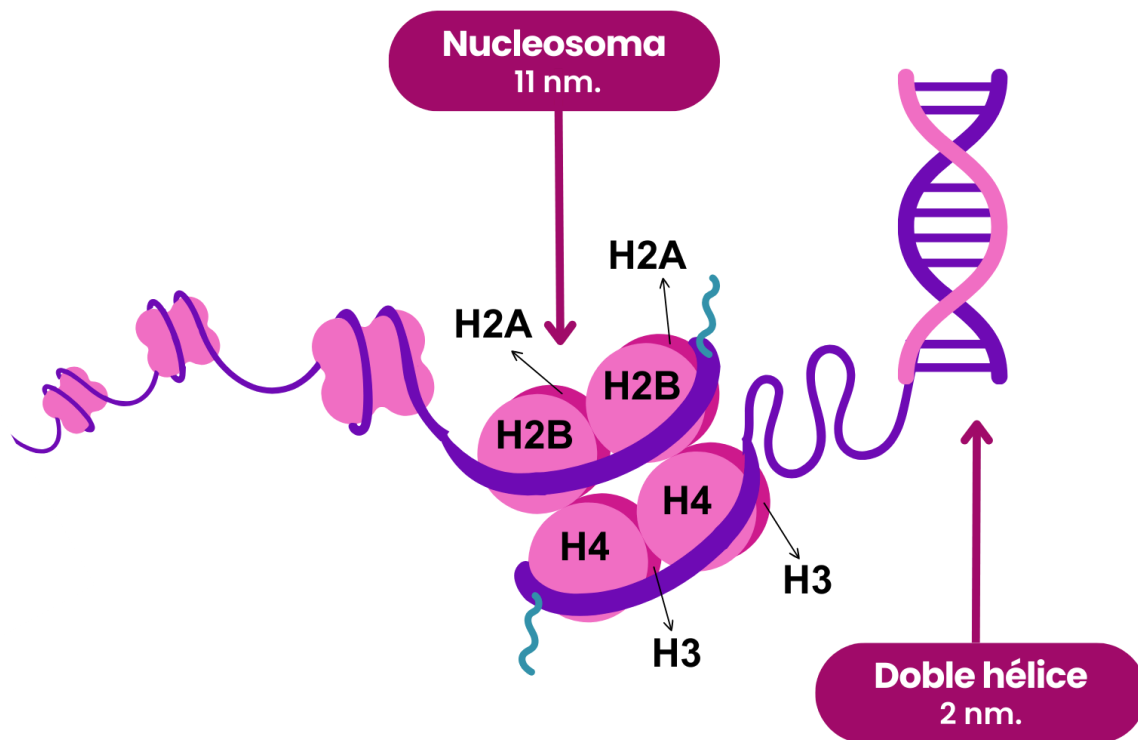


Figura 3. Esquema de la formación del nucleosoma a partir de la interacción entre histonas y el ADN.

La modificación de aminoácidos en las histonas es un proceso epigenético fundamental que desempeña un papel crucial en la regulación de la expresión génica y la estabilidad de la estructura de la cromatina. Las histonas pueden experimentar varios tipos de modificaciones químicas incluyendo la acetilación, metilación, fosforilación y ubiquitinación.¹⁴⁻¹⁶

La combinación de diferentes modificaciones sobre las histonas en una región específica del genoma puede determinar si un gen se expresa o se silencia. Por ejemplo, una región del ADN con histonas altamente acetiladas y metiladas en

patrones específicos generalmente indica un estado activo de la transcripción. En cambio, una región con histonas fuertemente metiladas y desprovistas de acetilaciones puede estar asociada con la represión génica.¹⁷

3.2.3 Por moléculas de ARN no codificante (ncRNA)

El mecanismo de regulación epigenética a través de moléculas de ARN no codificante (ncRNA, por sus siglas en inglés) constituye un campo bastante interesante en el estudio de la biología molecular y la genética. A pesar de que durante mucho tiempo prevaleció la idea de que el ARN ejercía predominantemente una función de intermediario en el proceso de síntesis de proteínas, su función en la regulación génica se encontraba en segundo plano. No obstante, investigaciones recientes han revelado que otra función con mucha importancia desempeñado por las moléculas de ncRNA es la modulación de la expresión génica sin provocar alteraciones en la secuencia de ADN. Este mecanismo, de naturaleza epigenética, confiere a las células la capacidad de controlar cuando y en qué cantidad se expresan los genes, lo que es esencial para diversos procesos como el desarrollo embrionario, la diferenciación celular y la adaptación a cambios en el microambiente celular y extracelular.¹⁸

Las moléculas de ncRNA se dividen en diversas categorías, incluyendo los ARN largos no codificantes (lncRNA, por sus siglas en inglés), el ARN interferente pequeño (siRNA, por sus siglas en inglés), los micro ARNs (miRNA, por sus siglas en inglés), el ARN circular (cirRNA, por sus siglas en inglés), entre otros. Estas moléculas pueden influir en la expresión génica mediante la degradación del ARN mensajero (mRNA, por sus siglas en inglés), la modulación de la cromatina, la regulación de la metilación del ADN, etc. Además, su capacidad para interactuar con genes específicos les permite desarrollar un papel crucial en procesos biológicos y en la adaptación de las células a su entorno.¹⁹⁻²⁰

3.2.3.1 ARN no codificante largo (lncRNA)

El mecanismo de regulación epigenética mediado por moléculas de lncRNA es un proceso interesante que involucra moléculas de ARN que no codifican

proteínas, pero que desempeñan un rol crítico en la modulación de la expresión génica y en la regulación epigenética. Las moléculas de lncRNA son transcritas a partir de ADN, y en lugar de ser traducidas por proteínas, interactúan con diversos componentes del genoma y la cromatina para influir en la actividad génica y la estructura de la cromatina. Algunos de los mecanismos de acción de la regulación génica mediada por lncRNAs son la interacción con complejos de remodelación de la cromatina, la modulación de la actividad de los mRNAs, la guía de complejos proteicos hacia blancos específicos y la regulación de la localización subcelular.¹⁹

Algunos lncRNAs son capaces de unirse a complejos de remodelación de la cromatina, como los complejos de metiltransferasa y desmetilasa. Estas interacciones pueden llevar a cambios en la estructura de la cromatina, lo que a su vez afecta la estabilidad del ADN y, en última instancia, la expresión de genes específicos.²¹

Por otra parte, otros lncRNAs pueden a su vez interactuar con mRNAs específicos, ya sea directamente o a través de proteínas. Esto puede resultar en la estabilización o degradación del mRNA, o en la inhibición de la traducción, lo que afecta la cantidad de proteínas que se sintetizan a partir de estos mRNAs.²²

A sí mismo, los lncRNAs también son capaces, como guías de reclutar complejos de proteínas hacia regiones específicas del genoma. Estos complejos pueden incluir proteínas encargadas de modular la metilación del ADN, la acetilación de histonas o la remodelación de la cromatina en sitios específicos.²³

Además, algunos lncRNAs pueden influir en la localización subcelular de proteínas y otros ARNs, lo que resulta en múltiples implicaciones en la regulación de la expresión génica.²⁴

3.2.3.2 ARNs no codificantes pequeños (siRNA y miRNA)

El mecanismo de regulación epigenética por ARNs no codificantes pequeños como los siRNA y miRNA, es un proceso fundamental en la regulación de la expresión génica a nivel postraducciona.²⁵

Los siRNAs son moléculas de ARN no codificantes de aproximadamente 20 a 25 nucleótidos de longitud. Su principal función en la regulación epigenética es

dirigirse de manera altamente específica a secuencias complementarias de mRNA. Una vez que el siRNA se une a su mRNA objetivo mediante la complementariedad de secuencias de bases nitrogenadas, pueden inducir la represión de genes a nivel traduccional. Esto se logra principalmente a través de dos mecanismos en los que los siRNA pueden impedir la traducción de proteínas a partir de ese gen o guiar a la degradación del mRNA. Además de estas funciones, los siRNAs también pueden influir en otros mecanismos de regulación epigenética como la metilación del ADN y las modificaciones de histonas.²⁶⁻²⁷

Los miRNAs, con una longitud de aproximadamente 22 nucleótidos, son también moléculas de ARN no codificantes. De manera similar a los siRNAs, los miRNAs establecen interacciones con secuencias complementarias de mRNAs, lo que puede conllevar al silenciamiento de genes. Estas interacciones suelen resultar en el bloqueo de traducción de proteínas o la degradación parcial del mRNA. A diferencia de los siRNAs, los miRNAs tienen la capacidad de tener múltiples objetivos y de regular varios genes. Además de su función directa en la represión génica, los miRNAs pueden influir en otros mecanismos como el desencadenamiento de modificaciones en la estructura de la cromatina o activando rutas de señalización que impactan en la regulación epigenética, lo que contribuye a la complejidad de su papel en la regulación epigenética.²⁸⁻²⁹

La Figura 4 muestra un esquema con las interacciones que se establecen entre secuencias complementarias de siRNAs con mRNAs (Figura 4a) y miRNAs con mRNAs (Figura 4b).

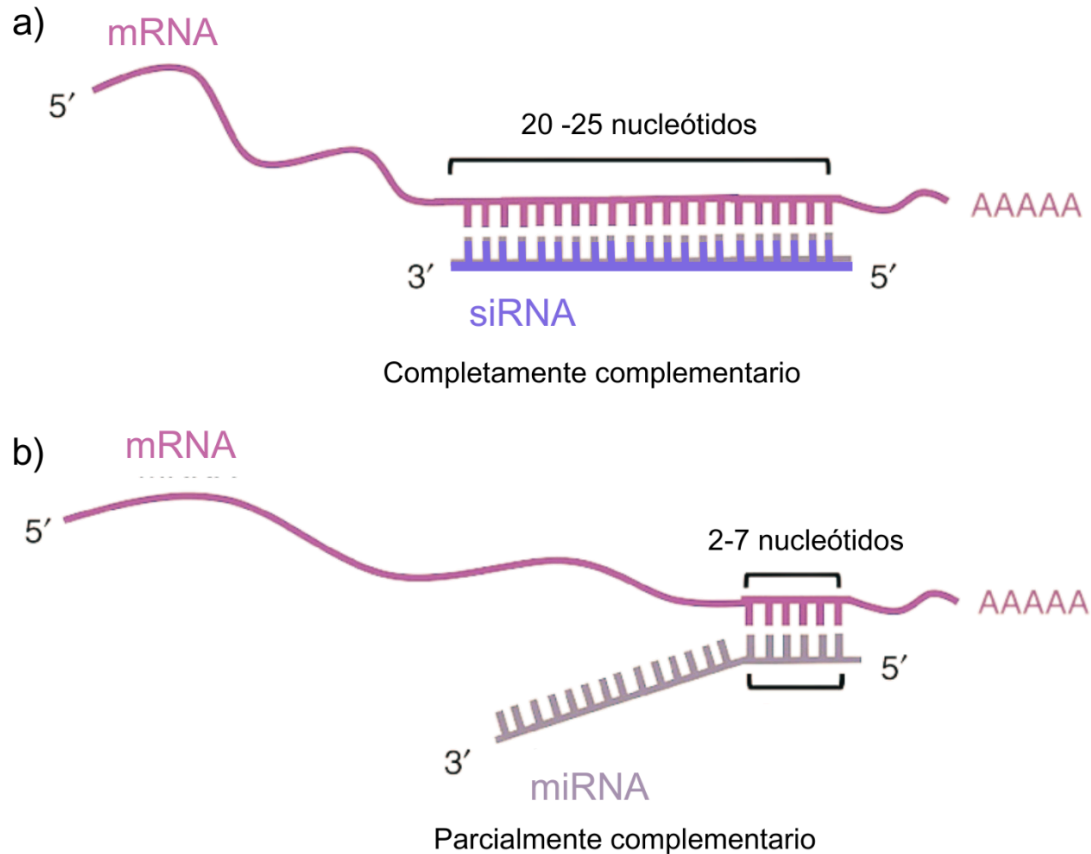


Figura 4. Reconocimiento de secuencias complementarias de mRNA a) por siRNA y b) por miRNA.

3.3.3.3 ARN circular (circRNA)

Un circRNA, es una molécula de ARN que forma una estructura circular cerrada en lugar de seguir la estructura lineal típica de las moléculas de ARN. Estas moléculas se generan a partir de la unión covalente de los extremos 3' y 5' de una secuencia de ARN precursora. Los circRNA se consideran una clase especial de ARN no codificantes y desempeñan un papel importante en la regulación de la expresión génica y la regulación epigenética.³⁰

Los circRNAs pueden actuar como cebadores para interactuar con proteínas y miRNAs. Algunos circRNAs tienen secuencias que se unen específicamente a proteínas reguladoras epigenéticas o a miRNAs (Figura 5). Esta interacción permite que los circRNAs modulen la actividad de proteínas y regulen la disponibilidad de miRNAs para regular genes específicos.³¹

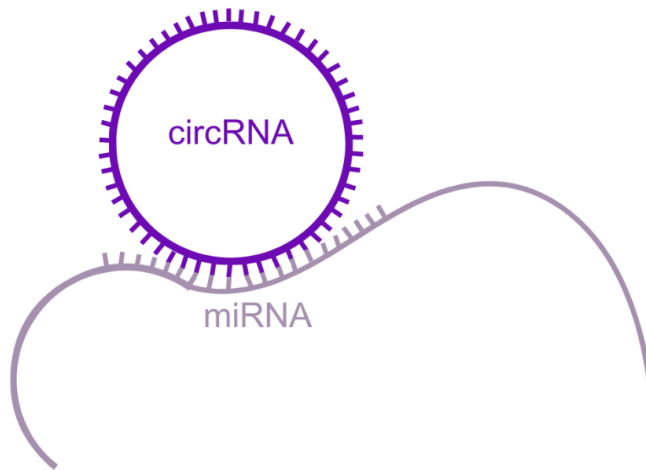


Figura 5. Interacción entre secuencias complementarias de circRNAs y miRNAs.

Algunos circRNAs han demostrado influir en la metilación del ADN y las modificaciones de histonas, procesos epigenéticos claves. Así mismo, pueden guiar complejos proteicos hacia regiones específicas del genoma, afectando así la cromatina y la accesibilidad de factores reguladores al ADN.³²

3.2.4 Modificaciones postraduccionales

El mecanismo de regulación epigenética por modificaciones postraduccionales se refiere a cambios químicos que ocurren en las proteínas después de que se han sintetizado a partir de la información genética. Estas modificaciones afectan la actividad de las proteínas, y en última instancia, pueden influir en la regulación epigenética y la expresión génica. Algunos de los tipos más comunes de modificaciones postraduccionales involucradas en la regulación epigenética incluyen la acetilación, metilación, fosforilación y ubiquitinación de histonas.¹⁴

3.2.4.1 Acetilación de histonas

La acetilación de histonas es una modificación epigenética clave que involucra la adición de grupos acetilo ($-\text{CH}_3\text{CO}$) a los residuos de Lys. Esta

modificación es mediada por enzimas específicas conocidas como acetiltransferasas de histonas (HATs, por sus siglas en inglés). Estas enzimas transfieren grupos acetilo de la acetilcoenzima A (AcCoA) a los residuos de Lys. Este proceso neutraliza la carga positiva de las histonas, lo que interfiere en su afinidad por el ADN.^{16,33}

La acetilación de histonas conduce a un estado de la cromatina más relajado y accesible, lo cual facilita el acceso de las enzimas y factores de transcripción al ADN y en consecuencia se promueve la activación de transcripción génica.

Además de las HATs, existen enzimas con la función opuesta conocidas como deacetilasas de histonas (HDACs). Las HDACs revierten la acetilación, restableciendo la carga positiva de las histonas. Esta dinámica de acetilación y deacetilación es esencial para la regulación fina de la expresión génica y la adaptación a cambios en el entorno celular.³⁴

3.2.4.2 Metilación de histonas

La metilación de histonas involucra la adición de grupos metilo (-CH₃) a residuos específicos de Lys o Arg en las histonas. Esta modificación es un proceso más complejo que la acetilación de histonas y puede tener efectos tanto activadores como represores, dependiendo de la ubicación donde ocurra.

Las enzimas responsables de la metilación de histonas se conocen como metiltransferasas de histonas (HMT, por sus siglas en inglés), las cuales transfieren grupos metilos desde el SAM a los residuos de Lys o Arg. La metilación de Lys en las histonas H3 y H4 se suelen asociar con la activación de la transcripción. Mientras que la metilación de otras Lys o Arg pueden tener un efecto represor. Debido a esta complejidad en los efectos de metilación, este es un mecanismo de regulación epigenético altamente sofisticado.^{15,17}

Así como en el caso de la acetilación de histonas, la metilación de histonas es un proceso dinámico que puede ser revertido. Las enzimas demetiladas de histonas (HDM, por sus siglas en inglés) eliminan los grupos metilo de las histonas cuando es necesario.³⁵

3.2.4.3 Fosforilación de histonas

La fosforilación de histonas implica la adición de grupos fosfato ($-\text{PO}_4$) a residuos específicos de Ser o Thr. Las enzimas responsables de esta modificación se conocen como cinasas. Estas cinasas transfieren el grupo fosfato desde el adenosín trifosfato (ATP) a las histonas. Este mecanismo desempeña un papel importante en la regulación de la transcripción génica, alterando las interacciones de las histonas con el ADN y proteínas asociadas, lo que puede tanto potenciar como reprimir la expresión de los genes específicos. Además, está estrechamente relacionado con los procesos de replicación y reparación del ADN, donde su presencia es crítica para la coordinación de la maquinaria molecular implicada en estos mecanismos. No obstante, no se limita a estas funciones, ya que también se extiende a otros procesos celulares, incluyendo la segregación cromosómica y la regulación del ciclo celular.³⁶⁻³⁷

A su vez, también existen las enzimas encargadas de eliminar los grupos fosfato de las histonas, las cuales se conocen como fosfatasas. El proceso de fosforilación es un mecanismo de regulación cuidadoso, el cual responde a señales precisas intracelulares y cambios en el entorno.³³

3.2.4.4 Ubiquitinación de histonas

Esta modificación involucra la adición de proteínas ubiquitina a residuos específicos de las histonas. La ubiquitinación de histonas se lleva a cabo por enzimas llamadas ligasas de ubiquitina (E3). Algunos ejemplos de este mecanismo de regulación son la ubiquitinación de las Lys120 en H2B y las Lys13, Lys15, Lys 118 y Lys119 en H2A. La principal función de la ubiquitinación de histonas es marcar la cromatina para su degradación o para regular la transcripción génica.³⁸

Al igual que otras modificaciones epigenéticas, la ubiquitinación de histonas es un proceso regulado y reversible. Las enzimas deubiquitinasas (DUB) tienen la función de eliminar la ubiquitina de las histonas, lo que permite tener un control preciso de la regulación de este mecanismo de regulación.³⁹

3.3 Dianas epigenéticas

Las dianas o proteínas epigenéticas desempeñan un rol crítico en la regulación de la expresión génica y la modulación de la información genética. Estas proteínas son las responsables del epigenoma, un conjunto de modificaciones químicas que actúan como interruptores maestros, encendiendo y/o apagando genes en respuesta a señales internas y externas. Para comprender su función se pueden dividir en tres principales categorías: las escritoras, las borradoras y las lectoras.⁴⁰

La Figura 6 muestra de manera esquemática cómo se pueden clasificar diferentes dianas epigenéticas en las tres categorías. Así mismo, se muestra de manera general el mecanismo de acción de cada una de estas categorías, añadiendo o removiendo marcas de metilación y acetilación. La descripción y el mecanismo detallado de cada una de estas proteínas se presenta en las siguientes secciones.

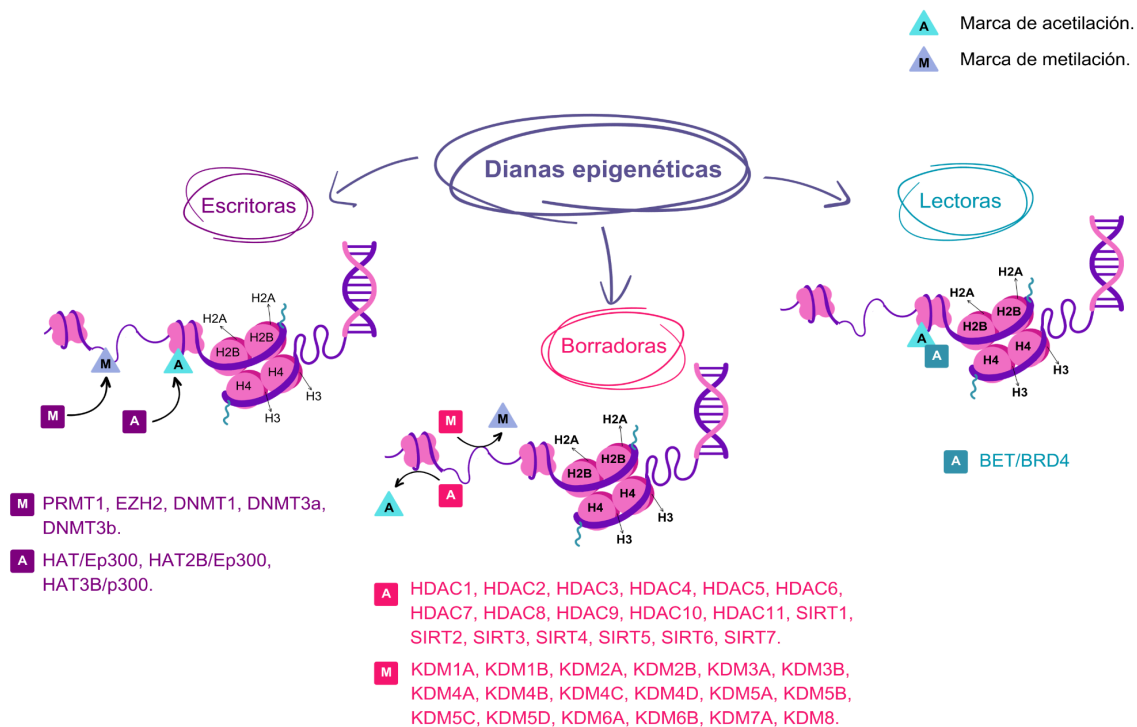


Figura 6. Representación visual de las proteínas epigenéticas y su función divididas en escritoras, borradoras y lectoras. PRMT: proteína arginina metiltransferasa; EZH2: potenciador del homólogo 2 de zeste; DNMT:

metiltransferasas de ADN; HAT: acetiltransferasas de histonas; HDAC: deacetilasas de histonas; SIRT: sirtuinas; KDM: desmetilasas de lisina; BET/BRD4: proteína 4 de la familia del dominio y bromodominio extra-terminal.

3.3.1 Escritoras

Las proteínas escritoras son las responsables de agregar modificaciones químicas específicas a las histonas o al ADN.⁴¹ Algunos ejemplos de proteínas escritoras epigenéticas son las DNMTs, el potenciador del homólogo 2 de zeste (EZH2), la proteína arginina metiltransferasa 1 (PRMT1) y las HATs.⁴²⁻⁴³

3.3.1.1 DNMTs

El mecanismo de metilación epigenética llevado a cabo por las DNMTs es un proceso fundamental en la regulación de la expresión génica y la estabilidad epigenética. Las DNMTs son enzimas que reconocen secuencias de ADN específicas que contienen citosinas no metiladas. Este proceso puede dividirse en varias etapas. En primer lugar, las DNMTs identifican secuencias de ADN que deben ser metiladas. Estas secuencias suelen ser ricas en dinucleótidos CpG, donde la citosina no está metilada. Las DNMTs se unen selectivamente a estas secuencias CpG no metiladas. Una vez que las DNMTs se han unido a las secuencias objetivo, transfieren grupos metilo (-CH₃) desde la fuente de metilo, generalmente el SAM, a las citosinas no metiladas en estas secuencias específicas. Este proceso da como resultado la formación de la 5-metilcitosina en el ADN (Figura 2).⁴⁴⁻⁴⁵

La metilación del ADN tiene un profundo impacto en la regulación de la expresión génica. Cuando una citosina en una secuencia CpG se metila, tiende a silenciar la actividad de los genes cercanos. Esto ocurre debido a la interacción de proteínas lectoras de metilación de histonas, que reconocen la metilación del ADN y modifican la estructura de la cromatina, reprimiendo la transcripción génica. Además, las DNMTs desempeñan un papel importante en el mantenimiento de patrones de metilación durante la replicación del ADN. DNMT1, en particular, se

encarga de asegurar que las nuevas hebras de ADN sean metiladas en las mismas posiciones que las hebras antiguas, lo que preserva la metilación epigenética a lo largo de las divisiones celulares.⁴⁶

3.3.1.2 EZH2

La metilación epigenética llevada a cabo por la proteína EZH2 es un proceso crucial en la regulación de la expresión génica. EZH2 es una histona metiltransferasa que pertenece al complejo polimerasa asociado a la metiltransferasa 2 (PRC2). Este mecanismo se inicia con el reconocimiento de su proteína objetivo, que son las histonas H3, específicamente en el residuo de lisina 27 (H3K27), que debe estar en su estado no metilado. A continuación, EZH2 actúa como una enzima metiltransferasa y realiza la transferencia de grupos metilo desde el SAM. La consecuencia de este proceso es la formación de una marca epigenética conocida como H3K27me3, que indica que la histona H3K27 ha sido metilada. Esta metilación tiene un impacto significativo en la regulación de la expresión génica, ya que está asociada con la represión de la transcripción génica. La metilación de H3K27 altera la estructura de la cromatina, dificultando el acceso de la maquinaria de transcripción al ADN. En última instancia, esto conduce a la supresión de la expresión de los genes relacionados con las regiones metiladas por EZH2.⁴⁷⁻⁴⁹

Este mecanismo de metilación por EZH2 desempeña un papel fundamental en la regulación de genes supresores de tumores y en el control de la diferenciación celular durante el desarrollo.

3.3.1.3 PRMT1

PRMT1 es una de las enzimas encargadas de la metilación de residuos de arginina. En primer lugar, PRMT1 identifica proteínas específicas que contienen residuos de arginina que deben ser metilados mediante interacciones con proteínas objetivo y secuencias de reconocimiento. PRMT1 selecciona residuos de arginina específicos para su acción. Una vez que se ha establecido la interacción

con la proteína objetivo, PRMT1 cataliza la transferencia de grupos metilo (-CH₃) desde la fuente de metilo SAM, a los residuos de Arg en la proteína. Esto da como resultado la formación de residuos de Arg metilada, en particular, la metilación simétrica (di-metilación) o asimétrica (mono-metilación) en la Arg.⁵⁰⁻⁵¹

La metilación de residuos de Arg en proteínas puede tener varias consecuencias funcionales. Por un lado, puede alterar la interacción de la proteína con otras moléculas o con el ADN, lo que influye en su función y localización subcelular. Por otro lado, la metilación de Arg también puede tener un impacto en la regulación de la expresión génica al modificar proteínas que interactúan con la cromatina y la maquinaria de transcripción.⁵²

3.3.1.4 HATs

Las HATs son enzimas responsables de la transferencia de grupos acetilo (-CH₃CO) a residuos de lisina en las histonas, que son proteínas alrededor de las cuales se enrolla el ADN.

En la primera etapa en el proceso de acetilación las HATs identifican residuos de lisina en las histonas que deben ser acetilados. Esto se lleva a cabo mediante el reconocimiento de secuencias específicas de aminoácidos en las histonas y la selección de las lisinas como sustratos para la acetilación. Las HATs interactúan con estas lisinas específicas y preparan el sustrato para la transferencia del grupo acetilo. Una vez que las HATs han seleccionado las lisinas objetivo, transfieren grupos acetilo desde la fuente de acetilo, generalmente AcCoA, a estas lisinas en las histonas. Esta transferencia de grupos acetilos neutralizan las cargas positivas de las lisinas y modifican la estructura de las histonas, lo que resulta en una relajación de la cromatina. Este proceso facilita el acceso de la maquinaria de transcripción al ADN, lo que influye en la regulación de la expresión génica. La acetilación de histonas tiene un impacto significativo en la regulación de la expresión génica. La relajación de la cromatina permite una mayor accesibilidad al ADN, lo que favorece la activación de la transcripción.⁵³⁻⁵⁴

Las HATs participan activamente en la activación de genes, en procesos de diferenciación celular y en la respuesta de las células a señales ambientales. La

acetilación epigenética realizada por las HATs, por ende, desempeñan un papel esencial en la modulación de la expresión génica y en la adaptación de las células a su entorno. Estas enzimas representan un componente central en la regulación de la cromatina y han sido ampliamente investigadas en el campo de la epigenética.⁵⁵

3.3.2 Borradoras

Las proteínas borradoras tienen la capacidad de agregar o quitar grupos químicos (como acetilos o metilos) de las histonas y en algunos casos de otras proteínas, lo que influye directamente en la estructura de la cromatina y, por ende, en la accesibilidad del ADN a la maquinaria de transcripción. Algunos ejemplos de estas proteínas son las HDACs, las sirtuinas (SIRTs), la desmetilasa específica de lisina 1 (LSD1) y las desmetilasas de lisina (KDMs).⁵⁶⁻⁵⁸

3.3.2.1 HDACs

Las HDACs son enzimas encargadas de eliminar grupos acetilos de las histonas, un proceso conocido como desacetilación. Este mecanismo juega un papel crucial en la regulación de la cromatina y, por ende, en la accesibilidad del ADN a la maquinaria de transcripción.⁵⁹

En primer lugar, las HDACs identifican los residuos de lisina en las histonas que contienen grupos acetilo. Estas histonas acetiladas son un componente clave en la activación de la transcripción génica, ya que la acetilación neutraliza las cargas positivas de las histonas y relaja la estructura de la cromatina. Una vez que las HDACs han reconocido las histonas acetiladas, actúan como enzimas desacetilasas y catalizan la eliminación de los grupos acetilo de las lisinas en las histonas. Este proceso tiene como resultado la formación de histonas no acetiladas, lo que conlleva a una cromatina más compacta y a la represión de la transcripción génica. La desacetilación permite que las histonas recuperen su carga positiva, lo que aumenta la interacción entre las histonas y el ADN, dificultando el acceso de la maquinaria de transcripción a los genes.⁶⁰

Por lo tanto, la desacetilación epigenética por HDACs tiene un impacto directo en la regulación de la expresión génica, al suprimir la actividad de los genes y favorecer un estado de silenciamiento. Esta regulación es esencial en una variedad de procesos biológicos, como el desarrollo, la diferenciación celular y la respuesta a señales ambientales.⁶¹

3.3.2.2 SIRT6

Las sirtuinas son una clase especial de proteínas borradoras que participan en la eliminación de grupos aceto de las histonas y otras proteínas, lo que modifica la estructura de la cromatina y la función de las proteínas asociadas. En la Figura 7 se muestra el mecanismo de desacetilación epigenética por SIRT6, el cual también se describe en detalle a continuación.

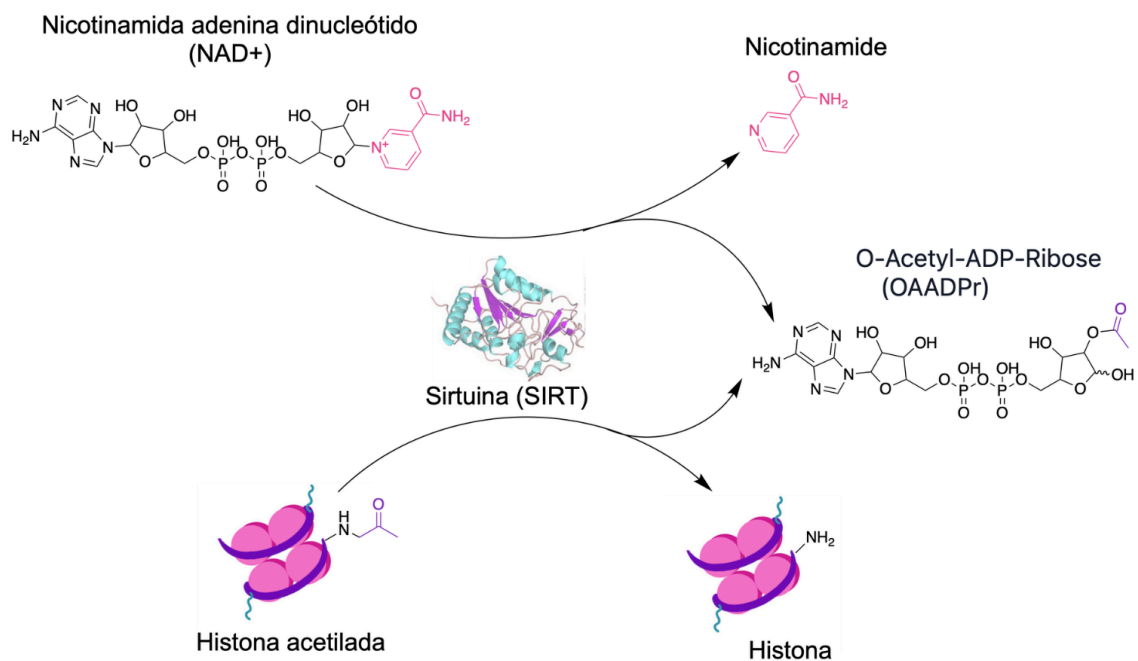


Figura 7. Mecanismo de desacetilación de histonas mediado por SIRT6.⁶²⁻⁶³

En primer lugar, las SIRT6 reconocen y se unen a las histonas acetiladas o a residuos de lisina acetilada en histonas específicas. Este reconocimiento se basa en la interacción entre el dominio de unión a la nicotinamida adenina

dinucleótido (NAD⁺) de las SIRT6 y los grupos acetilo en las histonas. Esta interacción permite que las SIRT6 seleccionen sustratos específicos para la desacetilación. Una vez que las SIRT6 se han unido a las histonas acetiladas, inician el proceso de desacetilación utilizando NAD⁺ como cofactor. Las SIRT6 catalizan la eliminación de los grupos acetilo de las lisinas en las histonas, lo que da como resultado histonas no acetiladas. Esta acción tiene un profundo impacto en la regulación de la cromatina y en la expresión génica, ya que como se ha mencionado anteriormente, la desacetilación promueve una cromatina más compacta y dificulta el acceso de la maquinaria de transcripción a los genes.⁶²⁻⁶³

Las sirtuinas desempeñan un papel crítico en una variedad de procesos biológicos, incluyendo la regulación del envejecimiento, la respuesta al estrés y la modulación de la expresión génica.⁶⁴

3.3.2.3 LSD1

LSD1 es una enzima desmetilasa que desempeña un papel fundamental en la eliminación de grupos metilo de residuos de lisina en histonas, lo que permite la activación de genes previamente silenciados. La acción de LSD1 en la desmetilación de histonas se centra en residuos de lisina metilados, específicamente en la lisina 4 en la histona H3 (H3K4), que está asociada con la activación de la expresión génica. La enzima LSD1 actúa específicamente en las histonas H3 y H4, contribuyendo a la modulación de la cromatina y la regulación de la expresión génica. Para llevar a cabo su función, LSD1 utiliza flavinas, como el FAD (dinucleótido de flavina-adenina), como cofactores, y su actividad está regulada por la interacción con diversas proteínas asociadas.⁶⁵

El mecanismo de desmetilación por LSD1 involucra una reacción de desaminación oxidativa, en la cual la enzima oxida el grupo metilo del residuo de lisina a una forma intermedia de hidroximetil-lisina. Posteriormente, esta hidroximetil-lisina es hidrolizada para producir lisina no metilada. La reacción de desaminación oxidativa requiere la participación de oxígeno molecular y la presencia de cofactores, como el FAD, que actúa como un agente oxidante. La actividad de LSD1 es altamente regulada y puede ser influenciada por la

interacción con proteínas cofactoras, como el correpresor del factor de transcripción silenciador (CoREST) y HDACs, que participan en la remodelación de la cromatina y la represión de la expresión génica. Además, las modificaciones post-traduccionales de LSD1, como la acetilación y la ubiquitinación, pueden modular su función y su estabilidad en el núcleo celular, lo que agrega un nivel adicional de control en la desmetilación epigenética.^{58,66}

3.3.2.4 KDMs

Uno de los mecanismos fundamentales de desmetilación implica la acción de las KDMs, que específicamente eliminan grupos metilo de las histonas, proteínas que envuelven el ADN en el núcleo de las células y contribuyen a la compactación de la cromatina. Este proceso permite la activación de genes que previamente estaban silenciados debido a la metilación de sus histonas, lo que tiene implicaciones significativas en el desarrollo, la diferenciación celular y la respuesta a estímulos ambientales.

Las KDMs son una familia de enzimas que se dividen en dos clases principales: las KDMs dependientes de hierro y oxígeno (KDM1) y las KDMs dependientes de oxígeno y 2-oxoglutarato (KDM2, KDM3, KDM4, KDM5 y KDM6). Estas enzimas actúan de manera específica en los residuos de lisina metilada en las histonas, catalizando la remoción de los grupos metilo. La actividad de las KDMs está regulada por la disponibilidad de cofactores y la interacción con otras proteínas, lo que permite un control fino de la desmetilación epigenética.⁶⁷

El mecanismo de desmetilación por KDMs comienza con la unión de la enzima a la histona que contiene la lisina metilada. Luego, la KDM cataliza la oxidación de la metilación a través de la acción de su dominio catalítico. En el caso de las KDMs dependientes de oxígeno y 2-oxoglutarato, la reacción implica la oxidación del grupo metilo a través de una serie de pasos que requieren oxígeno y 2-oxoglutarato como cofactores. Esta reacción produce una forma intermediaria de hidroximetil-lisina, que es posteriormente procesada para generar lisina no metilada. En el caso de las KDM1, la desmetilación se realiza mediante una reacción de desaminación oxidativa, donde la lisina metilada se convierte en

una forma desmetilada y una molécula de formaldehído. Este mecanismo se encuentra ejemplificado en la Figura 8.^{58,68}

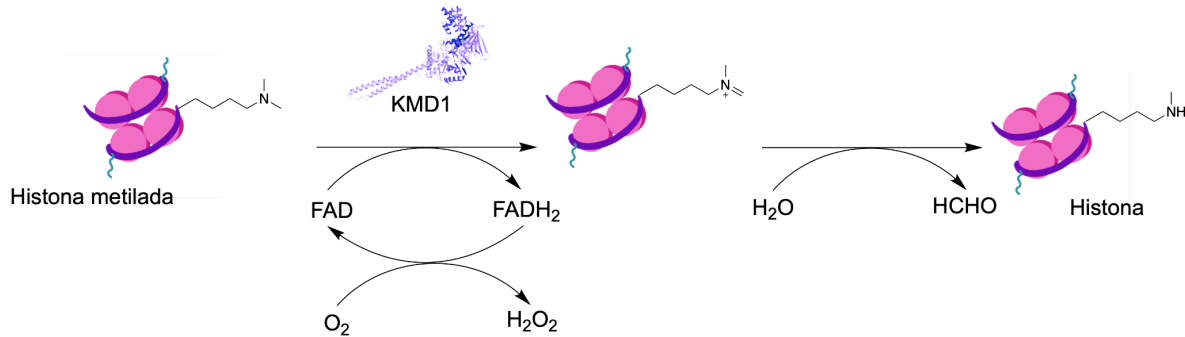


Figura 8. Mecanismo de desmetilación por KDM1 usando FAD como cofactor.^{58,68}

3.3.3 Lectoras

Las proteínas lectoras son macromoléculas que reconocen y se unen a las marcas epigenéticas, como las modificaciones en histonas o el metiloma del ADN, para interpretar la información almacenada en la cromatina y regular la expresión génica. Una marca epigenética es una modificación química en el ADN o en las proteínas asociadas al ADN que pueden influir en la actividad génica sin cambiar la secuencia de ADN. Estas proteínas lectoras son esenciales para la formación de complejos de regulación de la cromatina y desempeñan un papel fundamental en la traducción de señales desde las marcas epigenéticas hasta la maquinaria de transcripción. Entre las proteínas lectoras más importantes, se encuentra la proteína 4 de la familia del dominio y bromodominio extra-terminal (BET/BRD4).⁶⁹⁻⁷⁰

3.3.3.1 BET/BRD4

BET/BRD4 es un elemento clave que permite reconocer y unirse a marcas epigenéticas específicas, en particular a la acetilación de residuos de lisina en histonas. Los bromodominios tienen una alta afinidad por los grupos acetilo, lo que les permite interactuar de manera específica con las histonas acetiladas.

Una vez que BET/BRD4 se une a las histonas acetiladas, su función es actuar como un puente molecular que conecta las marcas epigenéticas con los componentes de la maquinaria de transcripción génica. BET/BRD4 forma complejos con proteínas coactivadoras y componentes de la polimerasa de ARN, facilitando así la activación de la transcripción génica. Esto implica una serie de interacciones proteína-proteína que permiten la formación de un ambiente propicio para la transcripción, incluyendo la remodelación de la cromatina y la liberación de la maquinaria de transcripción de la represión.⁶⁹

La acción de BET/BRD4 no se limita a la activación de la transcripción génica. También está involucrado en la regulación de la elongación de la transcripción, donde interactúa con el complejo de Elongación de la Transcripción Positivo del Factor b (P-TEFb, por sus siglas en inglés) para estimular la liberación de la polimerasa de ARN, permitiendo una mayor transcripción de los genes objetivos.⁷⁰

3.4 Enfermedades asociadas

Las dianas epigenéticas mencionadas y descritas en la sección 3.3.2 se encuentran asociadas a diversas enfermedades que involucran diversos tipos de cáncer y desórdenes metabólicos. La Tabla 1 presenta de manera detallada 46 dianas epigenéticas organizadas en las tres categorías: 8 proteínas escritoras, 37 proteínas borradoras y 1 proteína lectora. La Tabla 1 también resume información de los genes asociados y las enfermedades relacionadas, respectivamente.

Tabla 1. Dianas epigenéticas implicadas en la regulación de genes específicos y su asociación con diversas enfermedades.

Diana epigenética	Clasificación	Gen	Enfermedad asociada	Referencia
DNMT1	Escritora	<i>CEACAM5</i> , <i>DHRS3</i> , <i>RGS16</i>	Cáncer de mama	71
		<i>RARB</i>	Cáncer de mama	
		<i>RARB</i> , <i>CDK2AP1</i> y	Cáncer de cuello uterino	

		<i>CDK6</i>		
		<i>APC</i>	Cáncer de cuello uterino	
		<i>CDKN1A</i>	Cáncer de próstata	
		<i>APC y RARB</i>	Cáncer de próstata	
		No descrito	Cáncer de colon	
		<i>SLIT2 y TIMP-3</i>	Cáncer de pulmón	
		<i>WIF-1</i>	Cáncer de pulmón de células no pequeñas	
		<i>HIF-1α</i>	Leucemia mieloide	
		<i>cdk4</i>	Leucemia mieloide	
		<i>CDKN2A y RARB</i>	Cáncer de estómago	
		<i>RASSF1A</i>	Cáncer de próstata	72
		<i>NR4A1</i>	Diabetes tipo 2	73
		<i>PDX1</i>	Diabetes tipo 2	73
DNMT3a	Escritora	<i>RARB</i>	Cáncer de mama	71
		<i>CEACAM5, DHRS3 y RGS16</i>	Cáncer de mama	
		<i>CDKN2A y RARB</i>	Cáncer de estómago	
		<i>CDK6, HAI-2/PB y NQO1</i>	Carcinoma hepatocelular	
		<i>WIF-1</i>	Cáncer de pulmón de células no pequeñas	
DNMT3b	Escritora	<i>RASSF1A</i>	Cáncer de colon	71
		<i>RASSF1A</i>	Cáncer de pulmón	
		<i>RASSF1A</i>	Cáncer de próstata	
		<i>RASSF1A</i>	Cáncer de próstata	
HDAC1	Borradora	<i>HIF-1α</i>	Cáncer de colon	71
		<i>MITF-M</i>	Melanoma	
		<i>Brca1, Brca2, Ras, ERα, ERβ, PR, HER2, CDKN1A y TP53</i>	Cáncer de mama	
		<i>HIN-1</i>	Cáncer de tiroides	
		<i>CDKN1A</i>	Cáncer de próstata	
		<i>DAPK1</i>	Cáncer de estómago	
		No descrito	Cáncer de estómago	

		No descrito	Cáncer de mama	
		No descrito	Cáncer de esófago	
		No descrito	Cáncer de colon	
		No descrito	Linfoma cutáneo de células T	
		No descrito	Cáncer de próstata	
HDAC2	Borradora	No descrito	Melanoma	71
		No descrito	Cáncer de colon	
		<i>MICA</i>	Cáncer de pulmón de células no pequeñas	
		<i>WWOX y CDKN1B</i>	Cáncer de ovario	
		<i>Brca1, Brca2, Ras, ERα, ERβ, PR y HER2</i>	Cáncer de mama	74
		No descrito	Cáncer de colon	
		No descrito	Cáncer de próstata	
		No descrito	Cáncer de estómago	
		No descrito	Cáncer endometrial	
		No descrito	Linfoma cutáneo de células T	77
		No descrito	Alzheimer	
HDAC3	Borradora	<i>CDKN1A</i>	Leucemia promielocítica	71
		<i>SUV39 y G9a</i>	Cáncer de pulmón	
		<i>CDK6, SALL3 y GATA4</i>	Cáncer de colon y Cáncer de páncreas	
		<i>PHD2 y CDKN1A</i>	Cáncer de cuello uterino	
		<i>ERα y ERβ</i>	Cáncer de mama	74
		No descrito	Cáncer endometrial	
		No descrito	Cáncer de mama	
		No descrito	Cáncer de colon	
		No descrito	Cáncer de próstata	
		No descrito	Aterosclerosis	78
No descrito	Alzheimer	77		
HDAC4	Borradora	<i>CDKN1A</i>	Leucemia promielocítica	71
		<i>SUV39 y G9a</i>	Cáncer de pulmón	
		<i>CDK6, SALL3 y</i>	Cáncer de colon y Cáncer de páncreas	

		<i>GATA4</i>		74
		<i>PHD2 y CDKN1A</i>	Cáncer de cuello uterino	
		<i>ERα y ERβ</i>	Cáncer de mama	
		No descrito	Cáncer endometrial	
		<i>RECK y MMP-2</i>	Cáncer de ovario	
		No descrito	Cáncer de mama	
		No descrito	Bladder carcinoma	
HDAC5	Borradora	No descrito	Cáncer de pulmón	74
		No descrito	Melanoma	
		No descrito	Leucemia mieloide aguda	
HDAC6	Borradora	<i>Brca1, Brca2, Ras, ERα, ERβ, PR y HER2</i>	Cáncer de mama	71
		No descrito	Cáncer de mama	74
		No descrito	Cáncer de boca	
		No descrito	Linfoma cutáneo de células T	
HDAC7	Borradora	No descrito	Adenocarcinomas de páncreas	74
		No descrito	Cáncer colorrectal	75
		<i>PDX1</i>	Diabetes tipo 2	
HDAC8	Borradora	<i>CDKN1A</i>	Leucemia promielocítica	71
		<i>SUV39 y G9a</i>	Cáncer de pulmón	
		<i>CDK6, SALL3 y GATA4</i>	Cáncer de colon y Cáncer de páncreas	
		<i>PHD2 y CDKN1A</i>	Cáncer de cuello uterino	
		<i>ERα, ERβ</i>	Cáncer de mama	
		No descrito	Carcinoma oral de células escamosas	
		No descrito	Cáncer de pulmón de células no pequeñas	
		No descrito	Cáncer de colon	
		No descrito	Cáncer de mama	
		No descrito	Neuroblastoma	
HDAC9	Borradora	No descrito	Cáncer de pulmón	74
		No descrito	Meduloblastomas	
		No descrito	Astrocitomas	

HDAC10	Borradora	No descrito	Cáncer de pulmón	74	
		No descrito	Carcinoma hepatocelular		
HDAC11	Borradora	No descrito	Linfoma de Hodgkin	74	
SIRT1	Borradora	No descrito	Cáncer de ovario	71	
		No descrito	Cáncer de colon		
		No descrito	Cáncer de cuello uterino		
		<i>CDKN1A, pRB y cdk4</i>	Cáncer endometrial	74	
		<i>TP53</i>	Cáncer de mama		
		<i>TP53</i>	Leucemia mieloide aguda		
		<i>TP53</i>	Cáncer de colon		
		<i>TP53</i>	Cáncer de vejiga		
		<i>TP53</i>	Cáncer de próstata	76	
		<i>TP53</i>	Cáncer de ovario		
		No descrito	Aterosclerosis	76	
<i>EP300</i>	Alzheimer	77			
SIRT2	Borradora	No descrito	Glioma	74	
SIRT3	Borradora	No descrito	Cáncer de mama	74	
SIRT4	Borradora	No descrito	Cáncer de mama	74	
SIRT5	Borradora	No descrito	Cáncer de mama	74	
		No descrito	Pancreatic cancer		
SIRT6	Borradora	<i>ER</i>	Cáncer de mama	71	
		<i>Myc ZNF278 (coactivador de Myc), NM1, HOXB6 y MKRN3</i>	Leucemia mieloide		
		<i>PI3K, AKT1, ERK1 y MAPK1</i>	Carcinoma de células escamosas del esófago		
		<i>TMPRSS2: ERG y TP53</i>	Cáncer de próstata		
		<i>CDKN1A</i>	Cáncer de páncreas		
		<i>CDKN1A, CDKN1B y CDKN1C</i>	Cáncer de colon		
		<i>ATP2A3</i>	Carcinoma hepatocelular		
		No descrito	Cáncer de colon		74
		No descrito	Cáncer de mama		
		SIRT7	Borradora		No descrito
HAT/Ep300	Escritora	<i>EP300, IκBα y RELA</i>	Células tumorales cervicales	71	
		<i>EP300, IκBα y RELA</i>	Leucemia mieloide		

		<i>EP300, IκBα y RELA</i>	Linfoma de células T	
		<i>EP300, IκBα y RELA</i>	Cáncer de próstata	
HAT2B/Ep300	Escritora	<i>EP300 y KAT2B</i>	Cáncer de cuello uterino	
		No descrito	Cáncer de mama	
		No descrito	Carcinoma hepatocelular	
		<i>EP300 y CBP</i>	Cáncer de esófago	
HAT3B/p300	Escritora	<i>EP300 y TP53</i>	Cáncer de hígado	
PRMT1	Escritora	<i>AMI-1</i>	Carcinoma hepatocelular	71
EZH2	Escritora	<i>EZH2, EIF4E e IMPDH1</i>	Cáncer de mama	71
		<i>EZH2, EIF4E e IMPDH1</i>	Cáncer de cerebro	
		<i>EZH2, EIF4E e IMPDH1</i>	Cáncer de cuello uterino	
		<i>EZH2, EIF4E e IMPDH1</i>	Cáncer de colon	
		<i>EZH2, EIF4E e IMPDH1</i>	Cáncer de próstata	
		<i>UNC13B</i>	Neuropatía diabética	79
		No descrito	Aterosclerosis	76
LSD1	Borradora	No descrito	Cáncer de vejiga	71
		No descrito	Cáncer de colon	
		No descrito	Leucemia promielocítica	
		<i>NTRK2</i>	Leucemia promielocítica	
		No descrito	Cáncer de próstata	
		No descrito	Glioblastoma	
BET/BDR4	Lectora	No descrito	Leucemia MLL	71
KDM1A	Borradora	<i>TP53 y E2F1</i>	Leucemia, Cáncer de próstata, Cáncer de mama y Neuroblastoma	80
KDM1B		No descrito	Cáncer de mama	
KDM2A		<i>PCD4 y E2F1</i>	Leucemia, Cáncer de estómago y Cáncer de mama	
KDM2B		<i>KRAS</i>	Leucemia, Cáncer de páncreas, Cáncer de ovario, Cáncer de estómago y Glioma	
KDM3A		No descrito	Cáncer de mama,	

			Cáncer de ovario, sarcoma de Ewing y Cáncer de próstata
KDM3B		No descrito	Leucemia
KDM4A		No descrito	Cáncer endometrial, Cáncer de pulmón, Cáncer de próstata, y Cáncer de mama
KDM4B		<i>PTEN</i>	Cáncer de mama y cáncer de colon
KDM4C		No descrito	Cáncer de próstata
KDM4D		No descrito	Cáncer de próstata
KDM5A		<i>pRB</i>	Leucemia, Cáncer de mama, Cáncer de ovario y Melanoma
KDM5B		<i>Brca1 y CAV1</i>	Cáncer de mama, Cáncer de próstata y Melanoma
KDM5C		No descrito	Cáncer de cuello uterino
KDM5D		No descrito	Cáncer de próstata
KDM6A		<i>MYC y RUNX3</i>	Cáncer de vejiga, Cáncer de cuello uterino, Cáncer de mama, Mieloma múltiple, Cáncer de pulmón, y Cáncer de páncreas
KDM6B		<i>FOXO1</i>	Cáncer de pulmón, Cáncer de colon, Cáncer de páncreas, Cáncer de próstata, Mieloma múltiple, Leucemia mieloide aguda, Melanoma, y Cáncer de ovario
KDM7A		No descrito	Melanoma, Cáncer de cuello uterino y Cáncer de próstata
KDM8		<i>CCND1</i>	Cáncer de mama y Cáncer de próstata

3.5 Nutri epigenética

La nutri epigenética es una extensión de la epigenética. La cual se centra en cómo la dieta y los nutrientes influyen en la regulación de la expresión génica y, por lo tanto, en la salud. A través de mecanismos epigenéticos, los nutrientes pueden alterar la actividad de los genes sin modificar su secuencia de ADN, lo que tiene un profundo impacto en el fenotipo y la susceptibilidad a enfermedades.⁸¹

En este contexto, se han identificado varios compuestos dietéticos que pueden desempeñar un papel crucial en la nutri epigenética. Algunos ejemplos de estos compuestos son el ácido fólico, los polifenoles, los ácidos grasos omega-3 y los metabolitos producidos por la microbiota intestinal. Estos componentes pueden influir en la metilación del ADN, las modificaciones de las histonas y otros mecanismos epigenéticos, lo que a su vez afecta la regulación de la actividad génica y la respuesta ante diversas enfermedades.⁸²⁻⁸³

La nutri epigenética se puede clasificar en tres categorías: la directa, la indirecta y la transgeneracional. La nutri epigenética directa se refiere a los cambios epigenéticos que ocurren a lo largo de la vida de un individuo como resultado de su dieta y nutrición, estos cambios pueden influir en su salud y bienestar a corto y largo plazo. La nutri epigenética indirecta se relaciona con las modificaciones epigenéticas que ocurren en el útero como resultado de eventos durante la gestación, estos cambios pueden tener un impacto significativo en la salud a lo largo de la vida de un individuo. La nutri epigenética transgeneracional abarca los cambios epigenéticos que afectan a generaciones anteriores y que de alguna manera se transmiten a través de modificaciones en los gametos o el entorno intrauterino, esto puede influir en la salud de las generaciones futuras.⁸⁴

Un ejemplo de nutri epigenética directa es un estudio *in vivo* donde se observó una correlación entre la ingesta de folato en la dieta y la expresión del gen supresor de tumores p16 debido a que el folato inhibe la metilación del ADN en el promotor del gen p16, lo que a su vez modula su expresión. Este fenómeno ilustra cómo los nutrientes que consumimos pueden influir directamente en la actividad génica y los patrones epigenéticos que regulan la salud celular.⁸⁵

En resumen, la nutri epigenética es un campo de investigación en constante evolución que examina cómo la dieta y la nutrición pueden influir en la epigenética y, en última instancia, en la salud y la susceptibilidad a enfermedades, tanto en individuos como en generaciones futuras. Además de los estudios *in vivo* e *in vitro*, la quimioinformática puede permitir explorar y comprender en mayor profundidad estas interacciones.

3.6 Quimioinformática

La quimioinformática es una disciplina que surge de la combinación de datos químicos y recursos informáticos, y se utiliza para gestionar, representar y analizar de manera sistemática la información relacionada con la química. Esta disciplina facilita el manejo y análisis eficiente de grandes volúmenes de datos. Mediante enfoques quimioinformáticos, es factible organizar y gestionar información, visualizar el espacio químico, realizar minería de datos y establecer relaciones matemáticas entre la estructura y la actividad. Un ejemplo clásico de esto son las relaciones cuantitativas estructura-actividad (QSAR, por sus siglas en inglés).⁸⁶⁻⁸⁷

La quimioinformática desempeña un papel fundamental en el campo de la investigación farmacéutica y el diseño de medicamentos, ya que permite predecir cómo la estructura de una molécula química se relaciona con su actividad biológica. Esto es esencial para optimizar la eficacia y la seguridad de los compuestos farmacéuticos. Además, el uso de representaciones moleculares, como la nomenclatura SMILES (*Simplified Molecular Input-Line Entry System*) e InChI, son esenciales para la comunicación, el intercambio de información y la búsqueda eficiente de datos químicos en bases de datos.⁸⁸⁻⁸⁹

La quimioinformática ha revolucionado la forma en que se abordan los desafíos en el campo de la farmacología, la biología, las ciencias biomédicas, el descubrimiento de fármacos, la química analítica, la química orgánica, la investigación agroquímica, la ciencia de alimentos, la ciencia regulatoria, la ciencia de materiales y el control de procesos; permitiendo un enfoque más estructurado y basado en datos para la toma de decisiones en la investigación y el desarrollo de compuestos químicos.⁹⁰⁻⁹¹

3.6.1 Bases de datos moleculares

Las bases de datos moleculares juegan un papel fundamental en la quimioinformática y en la investigación química en general. Estas plataformas ofrecen un acceso organizado y eficiente a una amplia gama de información relacionada con moléculas químicas, compuestos orgánicos, productos naturales y

sus propiedades asociadas. Al proporcionar una vía centralizada para el almacenamiento y recuperación de datos químicos, las bases de datos moleculares permiten acelerar la investigación, descubrimiento y desarrollo de nuevos compuestos químicos y fármacos.⁹²

Uno de los ejemplos más notables de bases de datos moleculares de acceso libre es ZINC, una base de datos pública de compuestos químicos disponibles comercialmente, con versiones descargables en 2D y 3D, y un sitio web para búsquedas rápidas; su última versión disponible cuenta con más de 2 mil millones de compuestos.⁹³ Otro ejemplo es *COCONUT* (COLlection of Open Natural ProdUcTs), que se enfoca en productos naturales y compuestos orgánicos; *COCONUT* proporciona información detallada sobre la estructura química de miles de productos naturales, así como datos relacionados con su origen, propiedades biológicas y bibliografía asociada. La base de datos *COCONUT* fue desarrollada y actualizada por el grupo Steinbeck de la Universidad Friedrich-Schiller en Jena, Alemania.⁹⁴

Otras bases de datos enfocadas en productos naturales son *LANaPDB* (*Latin American Natural Product Database*) y *FooDB*. *LANaPDB* es una base de datos de libre acceso que contiene 12,959 compuestos recopilados a partir de nueve bases de datos de seis países latinoamericanos, proporciona información detallada, como las estructuras químicas en formato SMILES y propiedades tipo fármaco; los compuestos incluidos provienen de diversas fuentes, como plantas, animales terrestres y marinos, hongos y bacterias.⁹⁵

FooDB, una base de datos que contiene información sobre compuestos químicos presentes en alimentos, es mantenida por el Instituto de Investigación de Informática Médica de la Universidad de Alberta y ha tenido aplicaciones principalmente en las áreas de formulación de alimentos y evaluación de la composición nutricional de dietas.⁹⁶

3.6.2 Núcleos estructurales base (*molecular scaffolds*)

En quimioinformática, los núcleos estructurales base, también conocidos como estructuras base, andamiajes moleculares o, en inglés, *molecular scaffolds*; son elementos fundamentales de las moléculas químicas que sirven como

esqueletos o marcos estructurales sobre los cuales se construyen compuestos químicos más complejos.⁹⁷ En quimioinformática, la relevancia de los núcleos estructurales base se manifiesta en múltiples aplicaciones. Uno de estos casos es para analizar la diversidad de una biblioteca química. Este análisis permite identificar los componentes esenciales de una colección de compuestos químicos. Además, los núcleos estructurales base también son una herramienta valiosa para el diseño de nuevos compuestos farmacológicos. Comenzando con un núcleo estructural base, se pueden modificar sus grupos funcionales y sustituyentes para optimizar las propiedades de un compuesto, como su actividad biológica o solubilidad. Así mismo, también se utilizan en algoritmos de agrupamiento y análisis de datos para identificar similitudes estructurales entre compuestos químicos, lo cual es esencial en la minería de datos químicos y construcción de modelos QSAR para predecir propiedades biológicas. Los núcleos estructurales base también son esenciales en el proceso de descubrimiento de nuevos fármacos, ya que ayudan a identificar estructuras centrales que son responsables de la actividad biológica y, por lo tanto, guían el diseño de compuestos más efectivos y selectivos.⁹⁸

3.6.3 Espacio químico y su representación visual

El concepto del espacio químico desempeña un papel central en el campo de la quimioinformática y es la base conceptual de esta disciplina.⁹⁹ Este concepto constituye una representación abstracta y multidimensional que abarca todas las moléculas concebibles, donde cada molécula se representa como un punto en dicho espacio. La finalidad de estas representaciones radica en la capacidad de comparar, clasificar y analizar moléculas en función de sus características químicas, lo cual resulta esencial en la investigación farmacéutica y la química computacional.¹⁰⁰

En el espacio químico, cada dimensión corresponde a un descriptor molecular. Estos descriptores moleculares pueden abarcar tanto propiedades físicas como el peso molecular, la solubilidad y la polaridad; como propiedades químicas, tales como la presencia de grupos funcionales y enlaces químicos. Cada molécula se traduce en un vector propio dentro del espacio químico, en el

que las coordenadas de este vector reflejan los valores de las características químicas en las dimensiones pertinentes. Por lo tanto, cada molécula se convierte en un punto en este espacio multidimensional. La distancia entre los puntos en el espacio químico se puede calcular utilizando diversas métricas de distancia, lo que permite determinar la similitud química entre las moléculas. Moléculas cercanas en el espacio químico son químicamente similares, mientras que aquellas más alejadas exhiben diferencias más marcadas. Para representar visualmente este espacio multidimensional, se suelen aplicar técnicas de reducción de dimensiones; como el Análisis de Componentes Principales (PCA, por siglās en inglés),¹⁰¹ el Encaje Estocástico de Vecinos t-Distribuido (t-SNE, por siglās del inglés *T-Distributed Stochastic Neighbor Embedding*),¹⁰² los mapas autoorganizados (SOM, por sus siglas en inglés)¹⁰³ y los gráficos de mapas de árbol (TMAP, por sus siglas en inglés).¹⁰⁴ Estas técnicas simplifican las dimensiones del espacio químico a un número manejable, lo que facilita la visualización en un plano o en tres o dos dimensiones.¹⁰⁵

El empleo del espacio químico en quimioinformática es destacado por diversas razones. Primordialmente, permite la identificación de nuevas moléculas con potenciales propiedades o características de interés al facilitar la búsqueda de moléculas con similitudes a otras estructuras químicas ya conocidas que se tienen de referencia, lo que puede conducir al descubrimiento de nuevos compuestos que tengan la actividad de interés. Además, posibilita la clasificación de compuestos químicos en función de sus propiedades, lo que resulta valioso en diversos campos como la toxicología, el diseño de fármacos, la química de alimentos y la química de materiales. Asimismo, contribuye a la optimización de propiedades de las moléculas, como la solubilidad, biodisponibilidad y actividad biológica, y permite la selección de una variedad de compuestos químicos para su síntesis y posterior evaluación experimental.¹⁰⁶⁻¹⁰⁷

3.6.4 Relaciones cuantitativas estructura-actividad

Los métodos QSAR se basan en la premisa de que la actividad biológica de una molécula está intrínsecamente relacionada con su estructura química y sus propiedades físico-químicas.¹⁰⁸ El proceso típico de análisis QSAR involucra varios

pasos fundamentales. Comenzando con la selección y recopilación de datos con actividad biológica, que representan la respuesta de un compuesto en particular en términos de, por ejemplo, su concentración inhibitoria 50 (CI_{50}). Posteriormente, las estructuras químicas de los compuestos se codifican en forma numérica mediante descriptores moleculares, como el coeficiente de partición octanol-agua ($\log P$), energías de enlace, entre otros. Estos descriptores se utilizan para construir modelos matemáticos que relacionan las características moleculares con la actividad biológica.¹⁰⁹

Para evaluar la calidad y capacidad de predicción de un modelo matemático, es esencial realizar una validación utilizando conjuntos de datos de prueba. Diversas técnicas estadísticas y de aprendizaje automático se emplean en el proceso, como análisis de regresión lineal, PCAs para reducir la dimensionalidad de los datos; así como redes neuronales, máquinas de soporte vectorial y bosques aleatorios, para abordar problemas de mayor complejidad.¹¹⁰ Ejemplos de las visualizaciones que se pueden obtener a partir de estos análisis se muestran en la Figura 9.

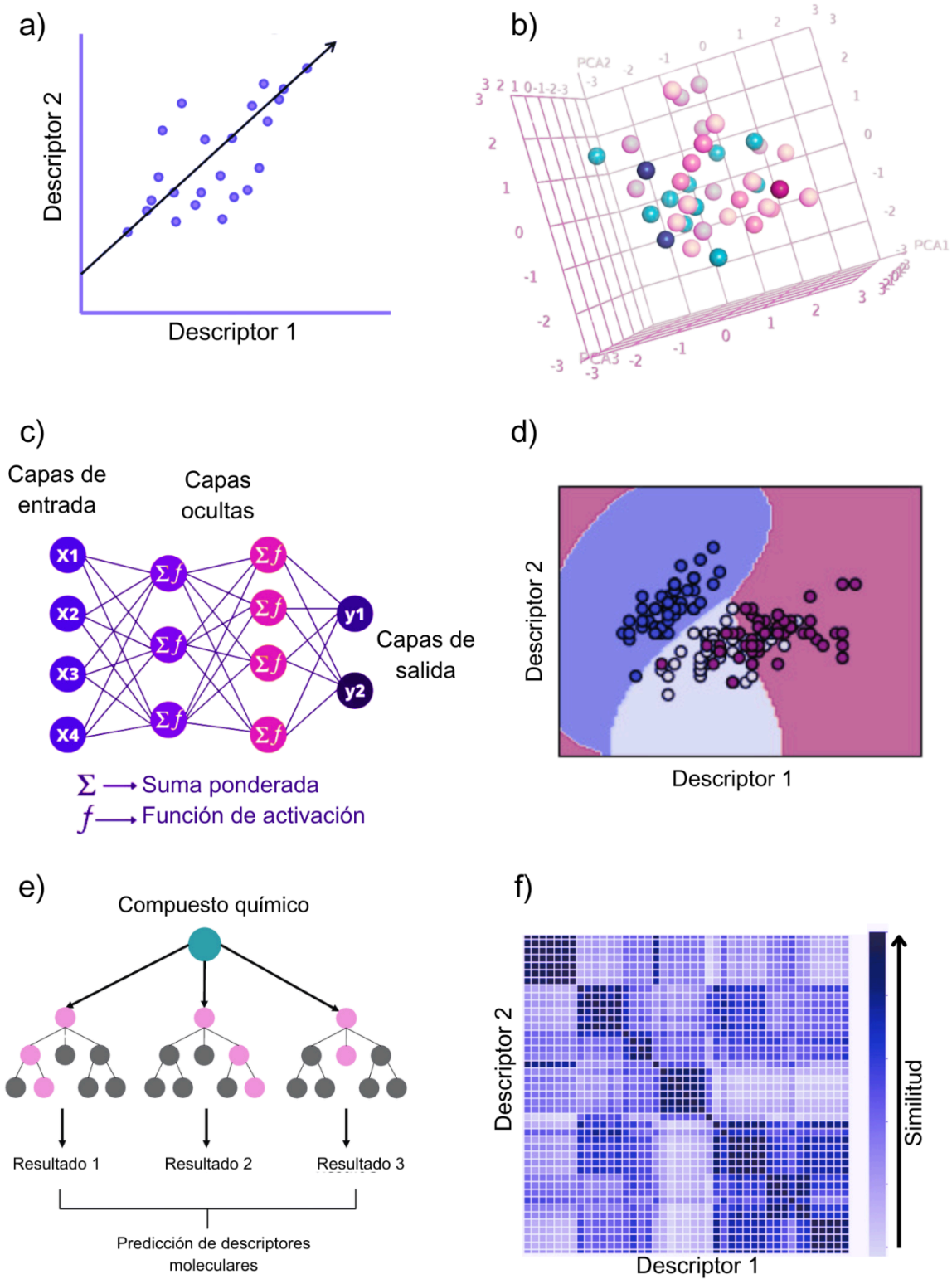


Figura 9. Representación visual de técnicas estadísticas y de aprendizaje automático que se emplean en quimioinformática para analizar relaciones estructura-propiedad (incluyendo actividad biológica); a) análisis de regresión

lineal, b) análisis de componentes principales (PCA), c) redes neuronales artificiales, d) máquinas de soporte vectorial (SVM), e) bosques aleatorios, f) mapa de calor en un análisis de similitud.

Los panoramas de actividad (del inglés, *activity landscapes*) también conocidos en la literatura en Español como paisajes de actividad, son una herramienta que ayuda a comprender y visualizar la relación entre las estructuras químicas de los compuestos y su actividad biológica en un espacio multidimensional. Esta representación visual es especialmente útil en el contexto de los métodos QSAR, ya que proporciona una vista integral de cómo las diferencias en las estructuras moleculares se correlacionan con las variaciones en la actividad biológica.¹¹¹

En un panorama de actividad, cada molécula se representa como un punto en un espacio multidimensional, donde cada dimensión corresponde a un descriptor molecular. La actividad biológica se representa en una escala de color, con colores más cálidos (por ejemplo, rojo) indicando una mayor actividad y colores más fríos (por ejemplo, azul) indicando una menor actividad. Al observar estos panoramas de actividad, es posible identificar patrones, cúmulos o regiones de alta o baja actividad de estructuras químicas que se correlacionan con actividades específicas. La representación de un panorama de actividad puede tomar diversas formas, como mapas de contorno, mapas de calor, gráficos de dispersiones multidimensionales, entre otros. Estas representaciones permiten identificar áreas de interés, diseñar compuestos con características específicas y optimizar la actividad biológica de una molécula.¹¹²⁻¹¹³

3.6.4.1 Panoramas de actividad

En quimioinformática, los panoramas o paisajes de actividad constituyen un enfoque esencial para la identificación y evaluación de compuestos químicos con propiedades biológicas similares o relacionadas. La base de este concepto se encuentra en la noción de similitud, un elemento clave que permite comparar moléculas y entender su relación en términos de estructura química y actividad

biológica. La similitud en quimioinformática se mide a través de diversas métricas, entre las más comunes se encuentran el coeficiente de Tanimoto, el coeficiente de Dice, el coeficiente de Jaccard y la similitud de Tversky. Estas métricas ofrecen diferentes perspectivas para evaluar la similitud entre moléculas.¹¹⁴

La representación de moléculas es un paso fundamental en la quimioinformática y contribuye significativamente a la comparación de compuestos. Algunas de las representaciones más utilizadas son las huellas digitales moleculares o *Fingerprints* y los patrones/perfiles de actividad biológica. Las huellas digitales moleculares son representaciones binarias que codifican características estructurales de las moléculas como bits 0 o 1, generadas por algoritmos como el ECFP (*Extended Connectivity Fingerprint*) o MACCS keys (del inglés, *Molecular ACCess System keys*). Por otro lado, los patrones o perfiles de actividad biológica describen la actividad de las moléculas en relación con sus dianas biológicas específicas y se basan en datos provenientes de ensayos biológicos de alto rendimiento o bases de datos de actividades biológicas previamente registradas.¹¹⁵⁻¹¹⁶

Para evaluar la similitud entre moléculas, se pueden utilizar diferentes coeficientes según el contexto y los objetivos del análisis. El coeficiente de Tanimoto-Jaccard (Ecuación 1)¹¹⁷ se emplea para comparar conjuntos de moléculas, midiendo el número de elementos compartidos entre los conjuntos en relación con el número total de elementos. El coeficiente de Dice (Ecuación 2) es similar al coeficiente de Tanimoto, pero se calcula considerando el doble del número de elementos compartidos en relación con la suma de los tamaños de los conjuntos. La similitud de Tversky (Ecuación 3) permite ajustar la importancia relativa de los elementos compartidos y no compartidos en los conjuntos moleculares, lo que puede ser útil en casos específicos.¹¹⁸

$$T(A, B) = \frac{|A \cap B|}{|A \cup B|} \dots \dots \dots \text{Ecuación 1}$$

$$D(A, B) = \frac{2 * |A \cap B|}{|A| + |B|} \dots \dots \dots \text{Ecuación 2}$$

$$Tv(A, B) = \frac{|A \cap B|}{\alpha * |A \cap B| + \beta * |A \setminus B| + \gamma * |B \setminus A|} \dots \dots \dots \text{Ecuación 3}$$

Donde: $T(A, B)$, es el coeficiente de Tanimoto entre los conjuntos A y B. $D(A, B)$, es el coeficiente de Dice entre los conjuntos A y B. $Tv(A, B)$, es la similitud de Tversky entre los conjuntos A y B. $|A \cap B|$ es el número de elementos compartidos entre los conjuntos. $|A \cup B|$ es el número total de elementos en los dos conjuntos combinados. $|A|$ es el tamaño del conjunto A. $|B|$ es el tamaño del conjunto B. $|A \setminus B|$ es el número de elementos que están en A pero no en B. $|B \setminus A|$ es el número de elementos que están en B pero no en A. α, β y γ son parámetros que permiten ajustar la importancia relativa de los elementos compartidos y no compartidos en la métrica.

En resumen, los panoramas de actividad en quimioinformática se basan en la similitud y la comparación de moléculas, utilizando diversas métricas y representaciones.^{117,119} Este enfoque es esencial ya que permite identificar compuestos químicos con propiedades biológicas afines, abriendo el camino a avances significativos en la ciencia y la medicina.

3.6.4.2 Acantilados de actividad (*activity cliffs*)

Los acantilados de actividad o *activity cliffs* en inglés, son un concepto importante en el campo de la quimioinformática, que se utiliza para describir situaciones en las que pequeños cambios en la estructura molecular de dos compuestos similares resultan en diferencias significativas en su actividad biológica o propiedades farmacológicas. Estos acantilados de actividad se refieren a una característica donde dos moléculas que son estructuralmente muy similares pueden tener perfiles de actividad drásticamente diferentes, lo que a menudo se representa como un acantilado en un gráfico c (SAR).¹²⁰⁻¹²¹

No todos los acantilados de actividad son necesariamente representativos de diferencias reales en la actividad biológica que presentan. Este fenómeno puede, en algunos casos, ser atribuido a lo que se conoce como artefactos derivados de la representación molecular utilizada en el análisis. En otras

palabras, un artefacto es considerado como una discrepancia en la actividad que compuestos similares pueden tener con un blanco de acción o el efecto terapéutico producido.¹²²

Los acantilados de actividad son un fenómeno intrigante y desafiante, ya que pequeños cambios en una molécula, como la adición o eliminación de un átomo o grupo funcional, pueden dar lugar a cambios significativos en su capacidad para interactuar con un blanco de acción o lograr un efecto terapéutico. En la quimioinformática, se utilizan técnicas de modelado molecular y análisis de datos para estudiar las relaciones entre la estructura química de compuestos y sus efectos biológicos o propiedades farmacológicas. Y a pesar de que se ha avanzado considerablemente en la identificación y caracterización de acantilados de actividad mediante diversos enfoques computacionales, la dependencia de la representación molecular plantea la posibilidad de que algunos compuestos detectados como acantilados de actividad sean más artefactos que reflejos de diferencias en la actividad sustanciales.¹²³

4. Planteamiento del problema

La epigenética ha revelado que los cambios en la expresión génica pueden ser mediados por factores ambientales y, en particular, por compuestos químicos presentes en los alimentos y productos naturales. La interacción entre productos químicos y dianas epigenéticas ha despertado el interés en la comunidad científica, ya que puede tener implicaciones significativas en la salud humana, incluyendo el desarrollo de enfermedades crónicas como el cáncer, las enfermedades cardiovasculares y trastornos metabólicos.

A pesar del avance de la investigación en este campo, hay desafíos importantes. Uno de ellos es la falta de una recopilación y análisis integral de cómo la estructura química de los compuestos presentes en alimentos y productos naturales influye en su actividad sobre dianas epigenéticas específicas. Además, la diversidad química de estos compuestos y su relación con su actividad epigenética aún no se ha explorado en forma sistemática. Por lo que la necesidad de avanzar en la comprensión de estas áreas aún no exploradas es evidente, ya que una comprensión más profunda de la relación entre la estructura química y la actividad epigenética de los compuestos podría conducir al desarrollo de estrategias más efectivas para la prevención y el tratamiento de enfermedades relacionadas con la epigenética.

Por lo tanto, en el presente estudio se analiza el avance en la investigación sobre productos químicos y componentes de alimentos que interactúan con dianas epigenéticas. La carencia de una base de datos integral dificulta la identificación y el análisis sistemático de los compuestos. Esta falta de una base de datos que integre información sobre la estructura química de productos químicos dietéticos y productos naturales con actividad epigenética reportada en la literatura limita nuestra comprensión. Este estudio se centrará en analizar la literatura reciente y crear una base de datos molecular que facilite obtener información disponible sobre los productos químicos presentes en alimentos y productos naturales que interactúan con dianas epigenéticas. La base de datos permitirá buscar, identificar y analizar patrones y relaciones entre la estructura química de estos compuestos y su actividad epigenética reportada en la literatura científica (revisada por pares). Se espera que la aportación de este trabajo facilite investigaciones futuras para

contribuir al desarrollo de estrategias más efectivas para la prevención y el tratamiento de enfermedades relacionadas con la epigenética.

5. Hipótesis

La estructura química de los compuestos presentes en alimentos y productos naturales está estrechamente relacionada con su capacidad para influir en dianas epigenéticas. Esta relación puede ser investigada y caracterizada mediante el uso de herramientas quimioinformáticas y técnicas de visualización del espacio químico, lo que permitirá una comprensión más profunda de los mecanismos subyacentes a la actividad epigenética de estos compuestos.

6. Objetivos

6.1 General

Analizar el avance de la investigación sobre productos químicos y componentes de alimentos que interactúan con dianas epigenéticas.

6.2 Particulares

➤ Construir una base de datos (*Epi Food Chemical Database*) de compuestos que integre información sobre la estructura química de moléculas dietéticas y productos naturales con actividad epigenética reportada en la literatura del 2017-2023.

➤ Evaluar la diversidad en el espacio químico de los compuestos en la *Epi Food Chemical Database* utilizando métodos cuantitativos y técnicas de visualización.

➤ Explorar el concepto de multiverso químico, es decir, el espacio químico generado con múltiples representaciones de estructuras químicas, para comprender cómo varía la diversidad química de los compuestos.

➤ Explorar las relaciones entre las estructuras químicas y el perfil de actividad epigenética utilizando el concepto de panoramas de actividad (*activity landscapes*).

7. Metodología

La Figura 10 resume las principales etapas de la metodología para alcanzar los objetivos. Las sub-secciones siguientes detallan cada una de las etapas.

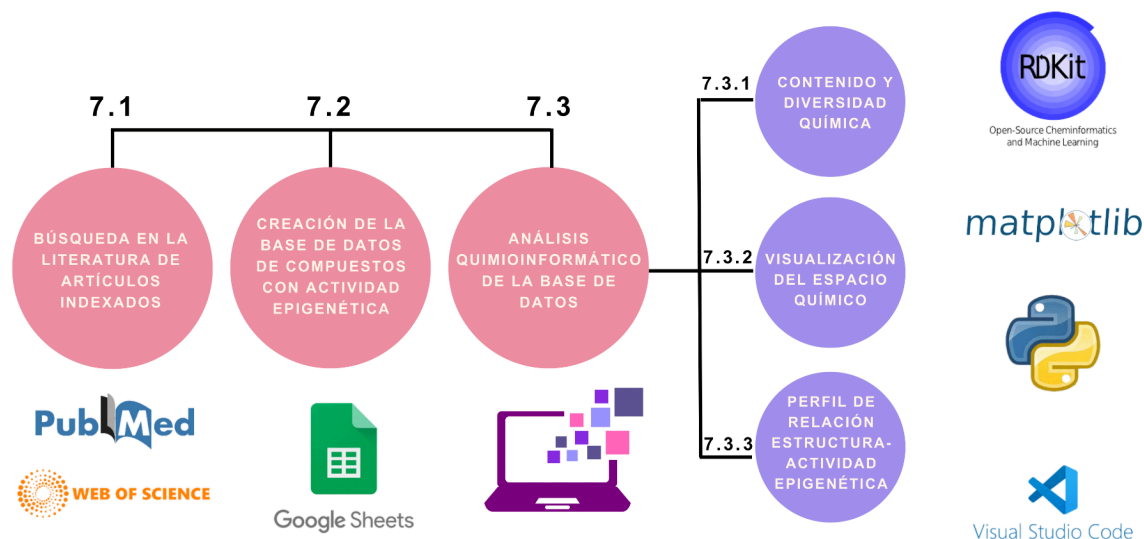


Figura 10. Diagrama visual de las etapas principales de la metodología empleada en el presente trabajo.

7.1 Búsqueda en la literatura de artículos indizados

Para la primera versión de la base de datos se decidió realizar un metaanálisis de la literatura científica que abarcó el período de enero de 2017 a marzo de 2023, centrándose en investigaciones publicadas en revistas indizadas y con números de identificación de objetos digitales (DOI). Esta revisión se enfocó en el estudio de sustancias químicas presentes en alimentos que tienen interacciones con dianas epigenéticas y que podrían tener aplicaciones terapéuticas o ser relevantes para la prevención de enfermedades. Este rango temporal se seleccionó debido al creciente interés y avances en la investigación en epigenética y nutrición a partir de la última década, reflejados en un aumento significativo en el número de estudios publicados.

La búsqueda bibliográfica se realizó en *PubMed*¹²⁴ y *Web of Science Core Collection*,¹²⁵ utilizando una de las siguientes palabras clave: ("epigenetics" Y "food chemical(s)") O ("epigenetics" Y "natural products") O ("epigenetics" Y "therapeutic

application") O ("*epigenetics*" Y "*disease*") O ("*epigenetics*" Y "*drug discovery*") O ("*epigenetics*" Y "*drug development*") O "*epigenetic targets*" O "*epigenetic therapy*" O "*epigenetic mechanisms*" O "*epigenetic regulation*" O "*epigenetic modifiers*" O "*epidrugs*" O "*nutritional epigenetics*" O "*nutrigenetics*").

Como parte del análisis, se identificaron los compuestos dietéticos y productos naturales mencionados en los resúmenes de los artículos seleccionados.

7.2 Creación de la base de datos de compuestos químicos presentes en alimentos y productos naturales con actividad epigenética

A partir de la búsqueda y el análisis de la literatura descrita en la Sección 6.1, se creó una base de datos compuesta que ha sido denominada en el presente documento como *Epi Food Chemical Database*. La versión actual se encuentra en el repositorio de GitHub:

https://DIFACQUIM.github.io/Epi_food_Chemical_Database/EpiFoodChemicalDatabase.html, disponible en formatos html y csv.

Epi Food Chemical Database contiene la siguiente información:

- El nombre de cada uno de estos compuestos.
- Su estructura química en notación SMILES.
- La principal fuente natural en dónde se encuentran presentes.
- Las respectivas notaciones lineales InChI e INChIKey.
- El enlace correspondiente a las bases de datos *FoodDB* y *COCONUT* si es que se encuentran disponibles (utilizando los identificadores correspondientes en esas bases de datos públicas).
- Las referencias (usando DOIs) de los artículos indizados asociados.
- El perfil de actividad que indica con cuáles dianas epigenéticas el compuesto respectivo ha reportado actividad en la literatura.

Para facilitar análisis posteriores y para identificar tendencias en los datos de manera eficaz usando diversos métodos quimiocinformáticos, el perfil de actividad se representó mediante un vector de "1"s y "0"s para indicar si el

compuesto ha reportado actividad o no contra las dianas epigenéticas específicas, respectivamente.

7.3 Análisis quimioinformático de la base de datos *Epi Food Chemical Database*

El análisis del contenido y la diversidad de las estructuras químicas de los 187 compuestos presentes en la base de datos de *Epi Food Chemical Database* se analizaron mediante tres principales enfoques: a) la evaluación de los núcleos estructurales base y la diversidad química a partir del uso de huellas digitales moleculares y la generación de andamios moleculares; b) la distribución del espacio químico; y c) la exploración de relaciones descriptivas entre la estructura y la actividad de los compuestos, basándose en el concepto de panoramas de actividad. En las secciones 6.3.1 a 6.3.3 se detalla cada uno de estos tres tipos de análisis.

7.3.1 Contenido y diversidad química

El análisis de contenido de núcleos estructurales base se llevó a cabo a partir de la definición de Bemis y Murcko, que considera un núcleo base como los anillos de una molécula y los conectores de los mismos.¹²⁶ Este análisis se realizó utilizando un código propio desarrollado en Python con el módulo *MurckoScaffold* de la biblioteca de *RDKit*.

A su vez, las estructuras químicas de la base de datos *Epi Food Chemical Database* se analizaron utilizando protocolos bien establecidos y ampliamente utilizados para caracterizar y evaluar la diversidad química. Es decir, se obtuvieron los núcleos estructurales base más frecuentes en la distribución; así como, la curva de la recuperación del sistema cíclico (CSR, por sus siglas en inglés).

7.3.2 Visualización del espacio químico

El espacio químico de la base de datos químicos de *Epi Food Chemical Database* se visualizó en una representación gráfica t-SNE. Como referencia, se incluyó la base de datos *FooDB*, la cual contiene 52.856 compuestos. El t-SNE se

realizó a partir de los 209 descriptores del módulo *MoleculeDescriptors* de *RDKit*. Algunos de los descriptores incluidos en este módulo son el peso molecular, el coeficiente octanol/agua (logP), el número de átomos donadores de hidrógeno (HBD), el número de átomos aceptores de hidrógeno (HBA), el área de superficie polar topológica (TPSA), el número de heterociclos en la molécula, el número de anillos aromáticos, el número de heteroátomos y el número de enlaces rotables. La Tabla S3 en el material suplementario muestra detalladamente los 209 descriptores utilizados para la generación del t-SNE. El análisis correspondiente se encuentra detallado en la sección de resultados y discusión respectiva (sección 8.3.2).

7.3.3 Perfil de relación estructura-actividad epigenética

Se llevó a cabo el cálculo de dos principales comparaciones pareadas para los 187 compuestos de la *Epi Food Chemical Database*. La primera basada en sus respectivas huellas digitales moleculares para encontrar la similitud estructural y la segunda a partir de los vectores generados del perfil de actividad contra las dianas epigenéticas (sección 7.2) para encontrar la similitud en el perfil de actividad.

Las huellas digitales moleculares son una representación matemática que captura las características estructurales de una molécula, facilitando así su análisis mediante herramientas quimioinformáticas. Para poder analizar la diversidad estructural se emplearon cuatro huellas digitales moleculares diferentes: MACCS keys (166 bits); ECFP de radio 2 (ECFP4) y de radio 3 (ECFP6); y los Fingerprints de *RDKit* (RDKit). Todos los análisis de similitud se calcularon mediante el índice de Tanimoto-Jaccard (Ecuación 1).

ECFP se basa en una variante del algoritmo de Morgan y opera en tres fases secuenciales.¹²⁷

- 1) En la primera etapa, conocida como asignación inicial, se otorga a cada átomo de la molécula un identificador único, considerando una serie de características como el número de átomos vecinos diferentes de hidrógeno, el número de enlaces, el número atómico, la masa atómica, la presencia en anillos, entre otros.

- 2) En la etapa de actualización iterativa, estos identificadores se recalculan teniendo en cuenta los identificadores de los átomos vecinos, utilizando para ello una matriz que refleje el nivel de iteración, el número de identificador del átomo, el orden de enlace y el número de identificador del átomo vecino, seguido de la aplicación de una función hash para generar un nuevo identificador. La huella digital resultante es una matriz que contiene la concatenación de estos identificadores a lo largo de las iteraciones. El proceso se repite hasta que se consideran todos los átomos vecinos dentro del radio definido.
- 3) En la etapa de eliminación de identificadores duplicados, se reducen las apariciones múltiples de identificadores repetidos a identificadores únicos. El resultado es una una matriz para cada identificador de cada átomo para cada nivel de iteración. Finalmente se crea un vector de 2048 *bits* a partir de las matrices generadas utilizando una función *hash*.

MACCS Keys es un tipo de huella digital conformada por un conjunto de preguntas a las que se les asignan respuestas binarias (0 o 1), generando así una representación única de la molécula. Este sistema se compone de un diccionario de preguntas, con dos versiones disponibles: una de 960 bits y otra de 166 bits. Sin embargo, en la actualidad, solo la versión de 166 bits está disponible públicamente en diversas herramientas y programas quimioinformáticos de código abierto.¹²⁸

Las huellas digitales de RDKit se generan identificando la topología de la molécula. Este proceso se puede dividir en cuatro etapas.¹²⁹

- 1) La molécula se somete a un preprocesamiento, que incluye la eliminación de hidrógenos y la normalización de la estructura para garantizar una representación consistente y comparable.
- 2) Se identifican varias características estructurales de la molécula, como fragmentos, subestructuras y patrones de conectividad.
- 3) Se genera un vector binario a partir de un patrón de conectividad generado en la etapa anterior, el cual representa la presencia o ausencia de una característica estructural específica en la molécula a

partir de caminos topológicos. RDKit utiliza un conjunto predeterminado de parámetros para la generación de los caminos topológicos: tamaño mínimo de camino: 1 enlace, tamaño máximo de camino: 7 enlaces, tamaño mínimo de fingerprint: 64 bits.

- 4) Una vez que se han identificado todos los caminos, se eliminan los vectores duplicados. El resultado es una matriz generada a partir de los vectores únicos. Finalmente se crea un vector de 2048 *bits* a partir de las matrices generadas utilizando una función *hash*.

La Figura 11 muestra un prototipo de un mapa de similitud de estructura-actividad (SAS, por sus siglas en inglés) donde la similitud del perfil de actividad epigenética se representa en el eje de las ordenadas, mientras que la similitud estructural basada en huellas digitales moleculares se traza en el eje de las abscisas. Un mapa SAS se puede dividir aproximadamente en cuatro regiones como se describe en la Figura 11; en la Región I hay pares de compuestos con perfiles de actividad muy similares, pero similitud estructural muy diferente. En la Región II se encuentran pares de compuestos con alta similitud estructural y alta similitud en sus perfiles de actividad. La Región III identifica pares de compuestos con baja similitud estructural y perfiles de actividad muy diferentes. En la Región IV, hay pares de compuestos con alta similitud estructural, pero perfiles de actividad epigenética muy diferentes.¹³⁰⁻¹³¹

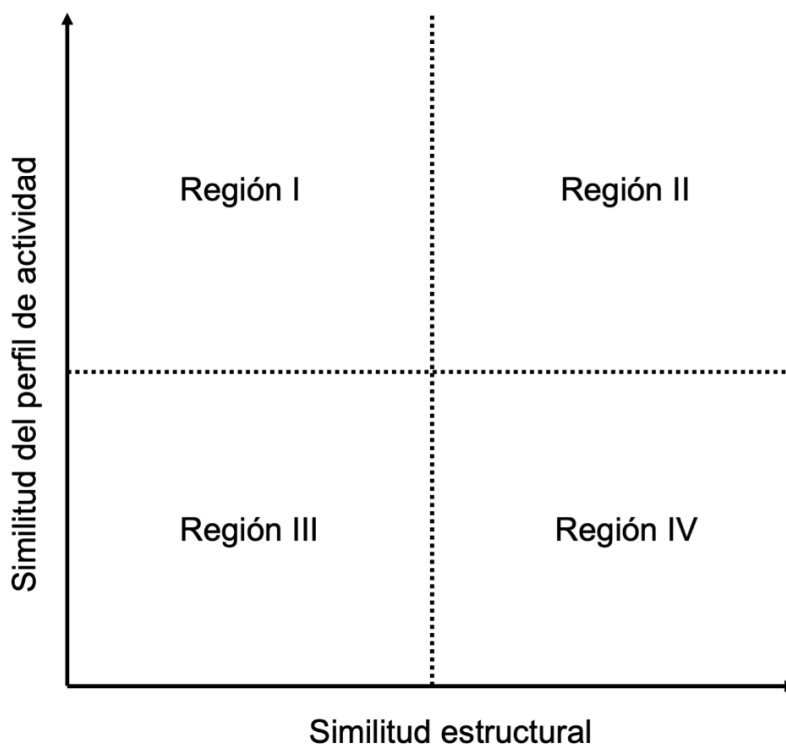


Figura 11. Gráfico prototipo de un mapa de similitud estructura-actividad (SAS). Los pares de compuestos en las regiones I y III tienen una similitud estructural baja, mientras que los de las regiones II y IV tienen una similitud estructural alta. Los pares de compuestos en las regiones I y II tienen una gran similitud en sus perfiles de actividad epigenética, aunque los compuestos químicos en las regiones III y IV tienen perfiles de actividad epigenética muy diferentes.

8. Resultados y discusión

8.1 Análisis de la literatura

De acuerdo a las palabras clave descritas en la sección 7.1, se encontraron 7,430 artículos indizados en *PubMed* y 5,960 en *Web of Science*; de los cuales 4,484 se encontraron en común en las dos bases de datos; y 2,946 eran únicos en *PubMed* y 1,476 en *Web of Science*. En total se obtuvieron 8,906 referencias únicas en la literatura entre ambas bases de datos.

8.2 *Epi Food Chemical Database*

De los 8.906 artículos únicos revisados de ambas bases de datos (*PubMed* y *Web of Science*), se seleccionaron solamente aquellos que hacían referencia explícita a compuestos dietéticos y productos naturales en sus resúmenes, resultando en la selección de 436 artículos utilizados para construir y curar la *Epi Food Chemical Database*. Es importante destacar que ningún compuesto dietético o producto natural fue excluido durante este proceso de selección.

La versión actual de la *Epi Food Chemical Database* contiene 187 compuestos únicos, de los cuales 121 compuestos han informado actividad específica contra al menos uno de las dianas epigenéticas y 66 compuestos han informado actividad general contra al menos una familia de dianas epigenéticas. En el contexto de este estudio, se estableció el criterio de considerar una actividad como específica cuando se mencionaba de manera explícita la acción de un compuesto contra una proteína específica. No obstante, es importante destacar que hay reportes que describen la actividad de un compuesto frente a una familia de proteínas, debido a que aún no se ha identificado con precisión la proteína específica contra la cual los compuestos tienen actividad dentro de esa familia. En estos casos, se catalogó la actividad como general.

Epi Food Chemical Database contiene diez columnas con información general; y cuarenta y nueve columnas que codifican el perfil de actividad epigenética de los compuestos en las cuarenta y seis dianas epigenéticas. El perfil de actividad epigenética está codificado como vectores de bits de “0”s y “1”s, lo

que indica la ausencia o presencia de actividad informada para cada uno de las 46 dianas epigenéticas.

Las dianas epigenéticas están ordenadas y organizadas en tres categorías principales: escritoras, borradoras y lectoras. Bajo esta clasificación, se identificaron; 8 escritoras (DNMT1, DNMT3a, DNMT3b, HAT/Ep300, HAT2B/Ep300, HAT3B/p300, EZH2, PRMT1); 37 borradoras (HDAC1, HDAC2, HDAC3, HDAC4, HDAC5, HDAC6, HDAC7, HDAC8, HDAC9, HDAC10, HDAC11, SIRT1, SIRT2, SIRT3, SIRT4, SIRT5, SIRT6, SIRT7, LSD1, KDM1A, KDM1B, KDM2A, KDM2B, KDM3A, KDM3B, KDM4A, KDM4B, KDM4C, KDM4D, KDM5A, KDM5B, KDM5C, KDM5D, KDM6A, KDM6B, KDM7A, KDM8) y 1 lectora (BET/BRD4).

Las quince dianas epigenéticas más frecuentes con actividad reportada de los compuestos en la *Epi Food Chemical Database* se muestran en la Figura 12. Se observa que la diana epigenética más frecuente es DNMT1 (63 compuestos), seguida de DNMT3B (35 compuestos), DNMT3A (34 compuestos con actividad reportada), HDAC6 (31 compuestos), y HDAC1 (28 compuestos).

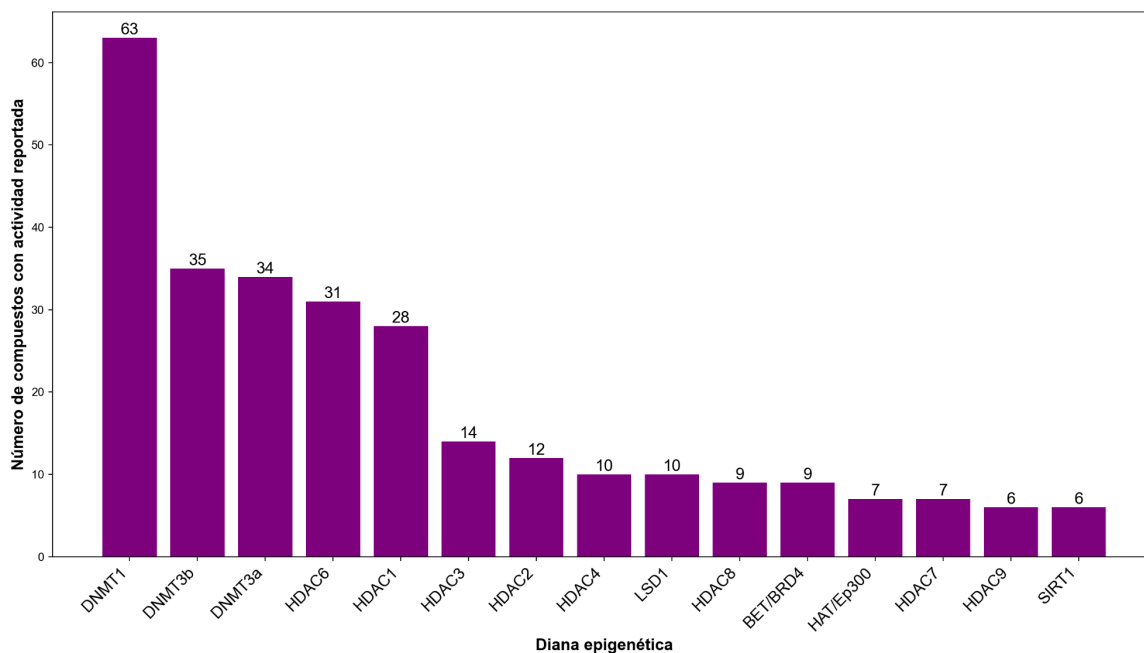


Figura 12. Histograma que muestra las 15 dianas epigenéticas más frecuentes con las que los compuestos de la *Epi Food Chemical Database* tiene actividad reportada.

Epi Food Chemical Database tiene 58 compuestos con actividad específica reportada para una única diana epigenética. DNMT1 y HDAC6 son las dianas epigenéticas con más compuestos con actividad asociada (18 compuestos cada diana). Seguidas por LSD1 con ocho compuestos; BET/BRD4 con cuatro compuestos; y DNMT3a, DNMT3b, HAT/Ep300, KDM4a, cada una con actividad frente a dos compuestos. Así mismo, tres dianas epigenéticas muestran actividad específica reportada ante un solo compuesto: HDAC1 con isotiocianato de fenetilo, SIRT1 con pterostilbeno y SIRT5 con glutamato.

De manera destacada, se identificaron cinco compuestos que exhiben actividad frente a varias dianas epigenéticas simultáneamente: la biotina asociada a 27, berberina a 15, alfa-cetoglutarato a 13, tricostatina A a 12 y buteína a 11. Los 10 compuestos presentes en la *Epi Food Chemical Database* con actividad reportada ante el mayor número de dianas epigenéticas se presentan en la Figura 13, detallando la estructura química y entre paréntesis el número de dianas epigenéticas correspondientes asociadas respectivamente. Las dianas específicas

contra las cuales tiene actividad reportada cada compuesto están disponibles en las versiones de la *Epi Food Chemical Database* (sección de anexos). Además, es importante destacar que el mecanismo de acción de cada compuesto es único para cada diana, determinado a su vez por la función y mecanismo de acción epigenético asociado a cada proteína correspondiente.

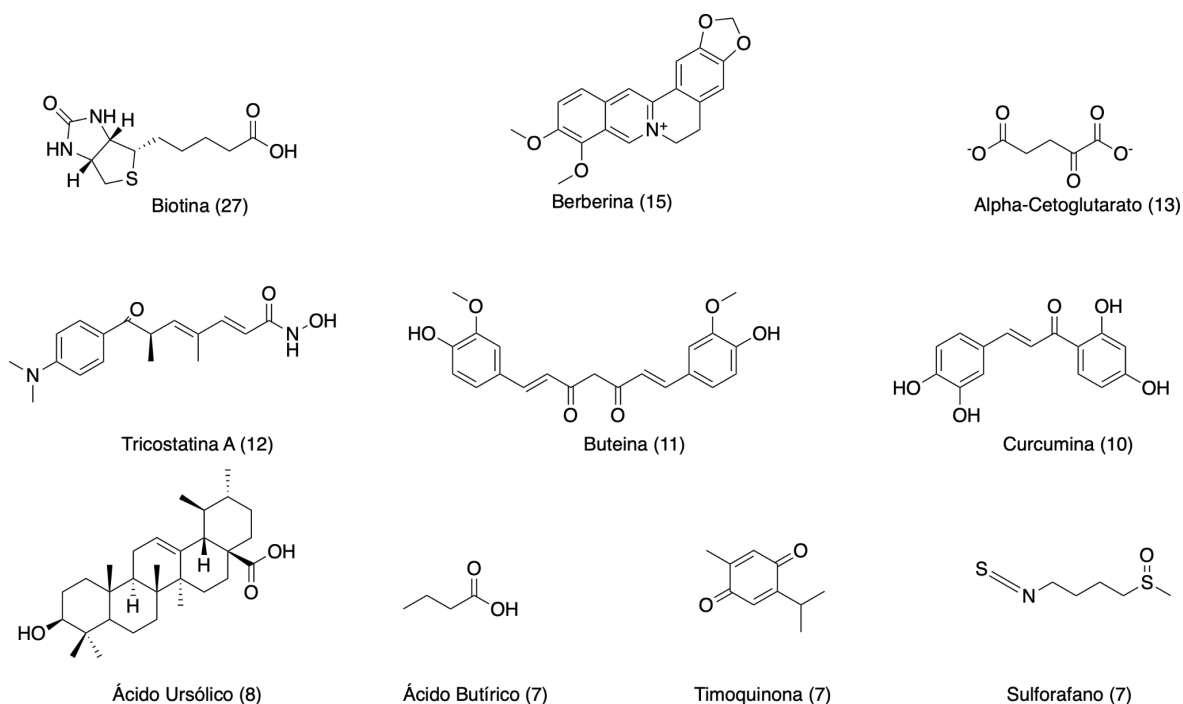


Figura 13. Diez compuestos principales en la *Epi Food Chemical Database* con actividad asociada al mayor número de dianas epigenéticas.

Las principales fuentes alimentarias donde se encuentran presentes los 187 compuestos químicos de la *Epi Food Chemical Database* son la carne de res, las legumbres, los cereales integrales, las bayas, las aves de corral, las bellotas, las acerolas y las nueces. La lista completa del total de las principales fuentes alimentarias respectivas para cada uno de los 187 compuestos identificados se encuentra en la *Epi Food Chemical Database*.

8.3 Análisis quimioinformático

En esta sección, se presentan los resultados del análisis quimioinformático realizado sobre la base de datos *Epi Food Chemical Database*. El objetivo principal fue evaluar el contenido y la diversidad química de los 187 compuestos incluidos en la base de datos mediante diferentes estrategias detalladas en la sección 7.3.

8.3.1 Contenido y diversidad química

Para evaluar y analizar los resultados del análisis quimioinformático del contenido y la diversidad química de la *Epi Food Chemical Database* se siguieron dos estrategias principales: 1) la generación y el análisis de los núcleos base, conforme a la definición de Bemis y Murcko, y 2) el análisis de la diversidad de núcleos base utilizando la curva CSR.

8.3.1.1 Análisis de núcleos base (*molecular scaffolds*)

En total se obtuvieron 90 núcleos base para los 187 compuestos. La Figura 14 muestra los diez núcleos base más frecuentes, junto con la frecuencia y la proporción porcentual, estos diez núcleos base representan el 35,54% de la distribución total de la *Epi Food Chemical Database*. Los andamios más frecuentes fueron el benceno (10,37%), seguido de la flavona (5,93%) y el flavilio (2,96%). Otros andamios frecuentes fueron el indol (2,96%), la piridina (2,22%), el hexano (2,22%) y la isoflavona (1,48%).

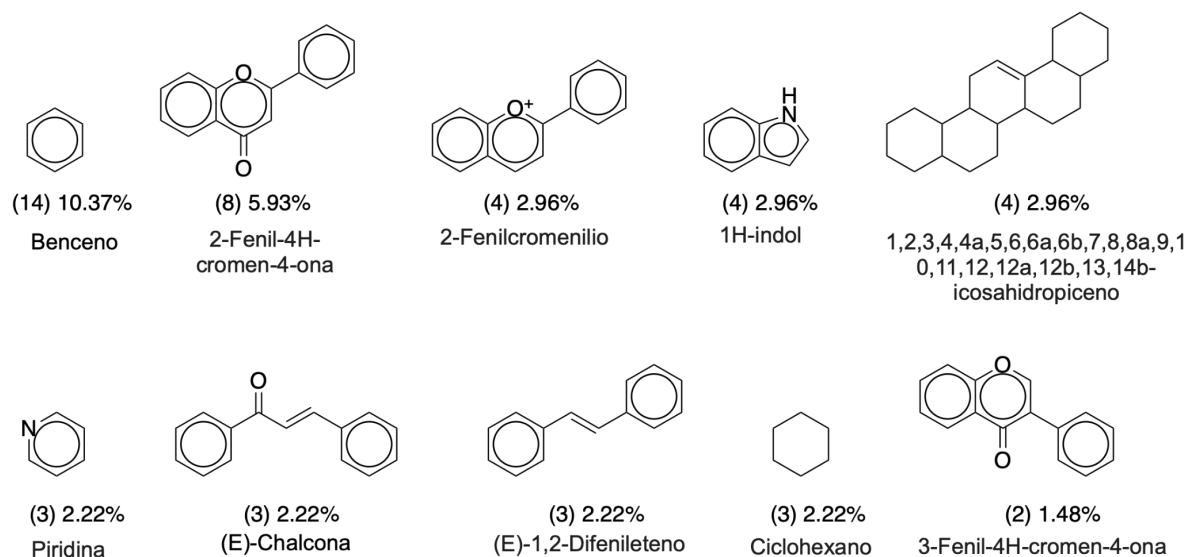


Figura 14. Se muestra la frecuencia (entre paréntesis) y el porcentaje de la distribución de los diez núcleos base más comunes en la *Epi Food Chemical Database*.

8.3.1.2 Curva de recuperación cíclica del sistema

La Figura 15 exhibe la curva CSR de los núcleos base presentes en la *Epi Food Chemical Database*, la cual evalúa la diversidad de los núcleos base. Esta curva representa la proporción de moléculas en un conjunto o base de datos que pertenecen a una fracción específica de núcleos base.

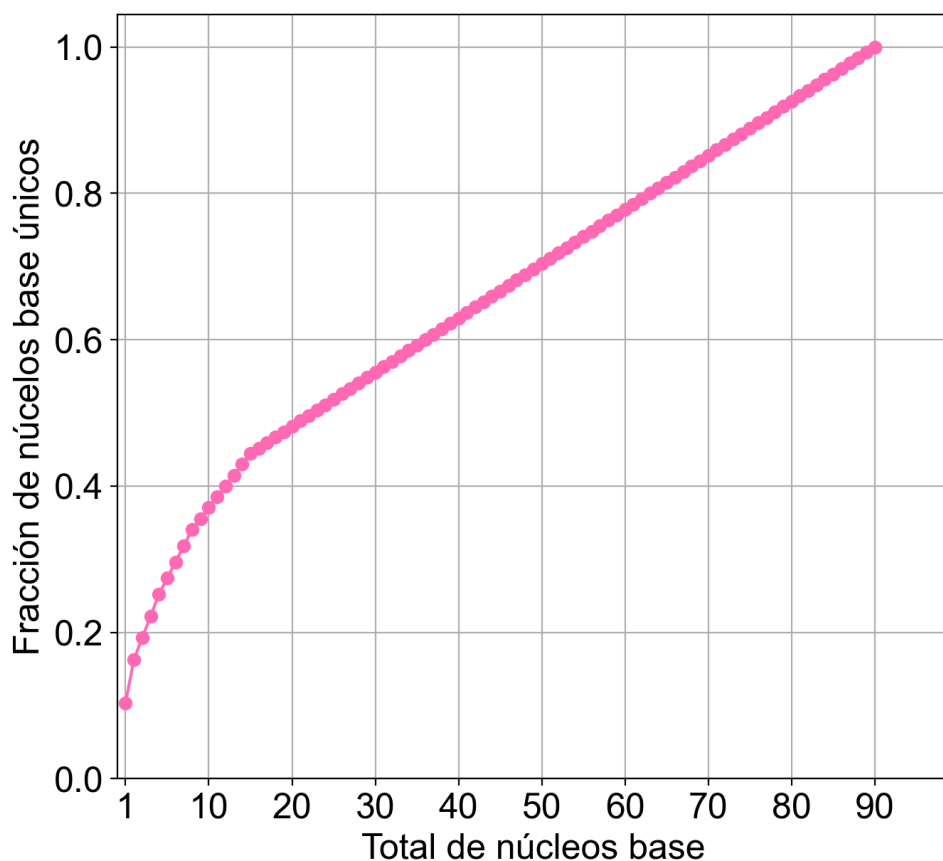


Figura 15. Curva CSR para los núcleos base de la *Epi Food Chemical Database*.

En un conjunto de datos altamente diverso, cada molécula de la biblioteca corresponde a un andamio distinto, resultando en una diagonal con un área bajo la curva (ABC) de 0,5. Con la disminución de la diversidad de los núcleos base, la curva se desvía de la orientación diagonal. En contraste, el caso de un conjunto de datos con la menor diversidad posible se observaría donde todos los compuestos comparten la misma estructura química; en tal caso, la curva CSR se presentaría como una línea vertical, acompañada por un ABC de 1,0.¹³⁴ La forma de la curva CSR en la Figura 15 indica una alta diversidad estructural en la base de datos de sustancias químicas de *Epi Food Chemical Database*, con un ABC de 0,75.

8.3.2 Visualización de espacio químico

El t-SNE es un método para la reducción de dimensionalidad, destacando por su capacidad para visualizar conjuntos de datos complejos de alta dimensión en espacios más manejables. Su funcionamiento se basa en el cálculo de una matriz de similitud entre las muestras en el espacio original, empleando medidas como la distancia euclidiana. En este estudio, se utilizó la distancia euclidiana para calcular la matriz de similitud entre las muestras del conjunto de datos original, considerando las 209 dimensiones correspondientes a los descriptores calculados para cada estructura química (ver sección 7.3.2). Este enfoque no se aplica directamente a las estructuras químicas, sino a los valores numéricos asociados a los descriptores moleculares de cada estructura química en la base de datos.

A partir de estas similitudes, el algoritmo genera distribuciones de probabilidad conjunta tanto en el espacio original como en el de baja dimensión. El proceso de optimización implica la minimización de la divergencia de Kullback-Leibler entre estas distribuciones, ajustando la posición de los puntos en el espacio de menor dimensión. Este enfoque asegura la preservación de la estructura local, manteniendo la proximidad entre puntos que eran vecinos cercanos en el espacio de alta dimensión. Además, la capacidad del t-SNE para identificar patrones no lineales en los datos es esencial para revelar estructuras complejas que podrían pasar desapercibidas con métodos más convencionales. Esta metodología asegura una evaluación eficiente de las similitudes entre las muestras, lo cual es indispensable para la aplicación subsiguiente del algoritmo t-SNE en la visualización de las relaciones entre las muestras en un espacio de menor dimensión. Esto facilita la interpretación de la estructura de los datos y proporciona una representación visual más comprensible.¹³⁵

La visualización mediante el método t-SNE del espacio químico de la base de datos de *Epi Food Chemical Database*, usando como referencia a la base de datos *FooDB*, se muestra en la Figura 16. En este caso el t-SNE utilizó un enfoque probabilístico para colocar muestras de un espacio de 209 dimensiones representadas por los descriptores moleculares calculados (Tabla S4) en uno de dos dimensiones, preservando la identidad de los los 53,043 compuestos químicos

más cercanos; 187 compuestos de *Epi Food Chemical Database* y 52,856 de *FooDB*.

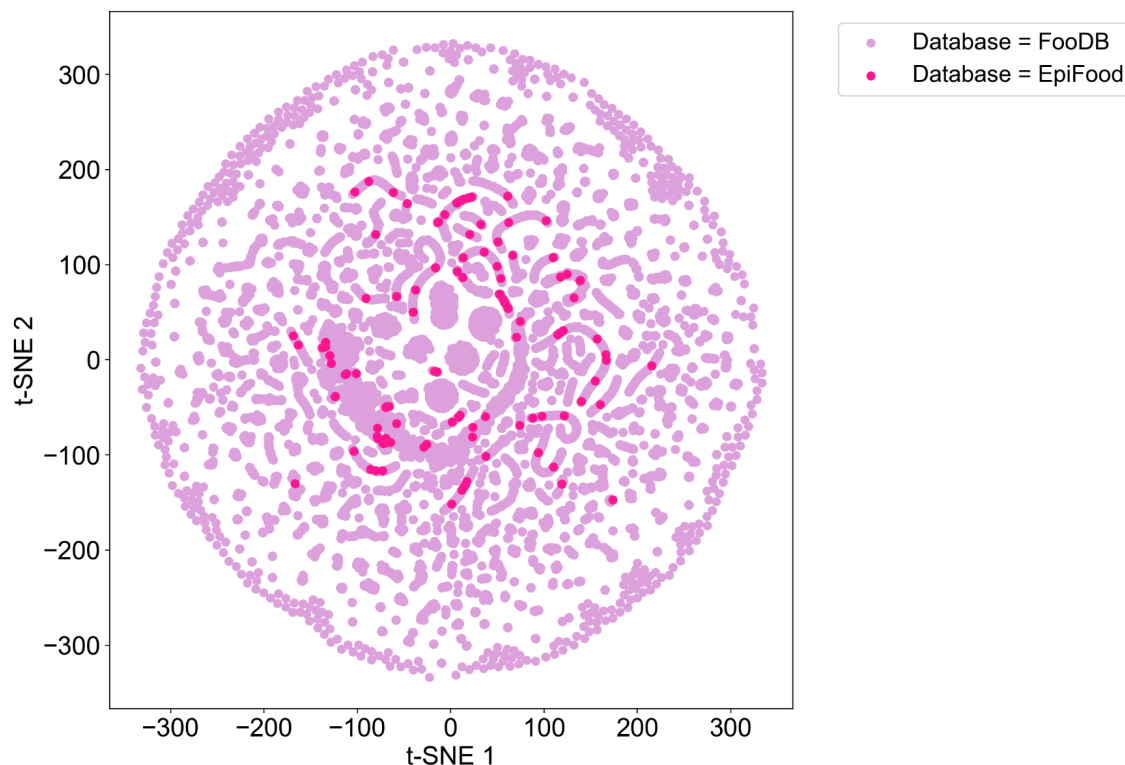


Figura 16. Representación visual del espacio químico mediante t-SNE de *Epi Food Chemical Database* (rosa intenso) en comparación con el espacio químico de *FooDB* (lila).

La Figura 16 sugiere que los puntos, que en este caso representan a los compuestos químicos presentes en *Epi Food Chemical Database* y *FooDB*, comparten espacio químico al ocupar regiones cercanas en el espacio de dos dimensiones. La proximidad de estos indica similitudes en las propiedades químicas calculadas subyacentes de los compuestos representados por esos puntos, sugiriendo que hay categorías de compuestos químicos o perfiles moleculares comunes entre *Epi Food Chemical Database* y *FooDB*. Las agrupaciones de puntos denotan compuestos con propiedades similares, mientras que aquellos posicionados en el borde de la gráfica son moléculas que exhiben menor similitud en relación con otras dentro del espacio de alta dimensionalidad.

Se eligió *FooDB*⁹⁵ como referencia para este análisis debido a que es la base de alimentos disponible más grande hasta la fecha, lo que le confiere una amplitud y representatividad considerable en términos de la diversidad de compuestos químicos presentes en los alimentos. Es importante destacar que *Epi Food Chemical Database* fue desarrollada de manera independiente a *FooDB*. Por lo tanto, es relevante considerar que hay compuestos en la *Epi Food Chemical Database* que no se encuentran en *FooDB*. Estos compuestos pueden ser identificados con la etiqueta "Not in database" en la columna correspondiente llamada "*FooDB ID*" dentro de la base de datos *Epi Food Chemical Database*, la cual se encuentra disponible en el repositorio de GitHub (consultar sección de anexos).

8.3.3 Perfil epigenético

En total, para cada huella digital molecular se calcularon 17.590 comparaciones pareadas (incluidas las auto comparaciones) y 17.430 comparaciones pareadas (excluyendo auto comparaciones). La similitud del perfil de actividad epigenética y la similitud estructural para cada huella digital molecular se representó en un diagrama de dispersión, también conocidos como mapas SAS.

En las subsecciones siguientes se presentan los mapas SAS de los pares de compuestos de la *Epi Food Chemical Database* a partir de los valores obtenidos de la similitud estructural calculada con el coeficiente de Tanimoto y las huellas digitales ECFP4 (Figura 17), ECFP6 (Figura 18), MACCS Keys (Figura 19) y RDKit (Figura 20). En todos los mapas se calculó (eje de las ordenadas) el valor de similitud obtenido del perfil de actividad epigenética según está descrito en la sección 7.3.). Los puntos en rosa representan a los pares de compuestos químicos en la región II de los mapas SAS (según la Figura 11), que corresponden a compuestos muy similares tanto en estructura como en perfil de actividad. Estos pares de compuestos también se conocen como pares de compuestos que tienen relación SAR continua u homogénea. Por el contrario, los puntos de datos verdes en los mapas SAS representan pares de compuestos en la región IV (de acuerdo con la Figura 11), los cuales corresponden a pares de compuestos con perfiles de

actividad similar, pero estructuras químicas muy diferentes, también conocidos como pares de compuestos con acantilados de actividad. Los puntos en morado son aquellos que corresponden a pares de compuestos con baja similitud estructural, correspondientes a las regiones I y III de la Figura 11.

Para seleccionar los pares de compuestos con actividad SAR se seleccionaron aquellos que tenían un valor de similitud estructural mayor al valor promedio más dos desviaciones estándar de la distribución de la comparación pareada de la *Epi Food Chemical Database* y un valor de similitud en el perfil de actividad mayor a 0.1. Mientras que para los pares de compuestos acantilados de actividad se seleccionaron aquellos que tenían un valor de similitud estructural mayor al promedio más dos desviaciones estándar de la distribución de la comparación pareada de la *Epi Food Chemical Database* y un valor de similitud en el perfil de actividad menor o igual a 0.1.

En las secciones 8.3.3.1 a 8.3.3.4 se muestran los mapas SAS y los pares de compuestos SAR y acantilados de actividad para cada huella digital molecular. Así mismo los vínculos para visualizar los mapas SAR interactivos se encuentran disponibles en la sección de anexos.

8.3.3.1 ECFP4

La Figura 17 muestra el mapa SAS para los pares de compuestos en la *Epi Food Chemical Database* usando la huella digital molecular *Extended Connectivity Fingerprint* radio 2 para el cálculo de la similitud estructural.

La Tabla 2 presenta los veintidós pares de compuestos, marcados con puntos rosas en la Figura 17, que exhiben una relación simultánea entre su estructura y su actividad, identificados como compuestos con un SAR continuo. Por otro lado, la Tabla 3 muestra los ciento veintinueve pares de compuestos, representados por puntos verdes en la misma figura que tienen relación entre su estructura, pero no entre su perfil de actividad, también conocidos como compuestos acantilados de actividad. Los valores de similitud fueron calculados utilizando el coeficiente de Tanimoto, así como la huella digital molecular ECFP4, para el cálculo de la similitud estructural (SE); y los vectores del perfil de actividad

respectivos para cada compuesto para el cálculo de la similitud del perfil de actividad (SPA).

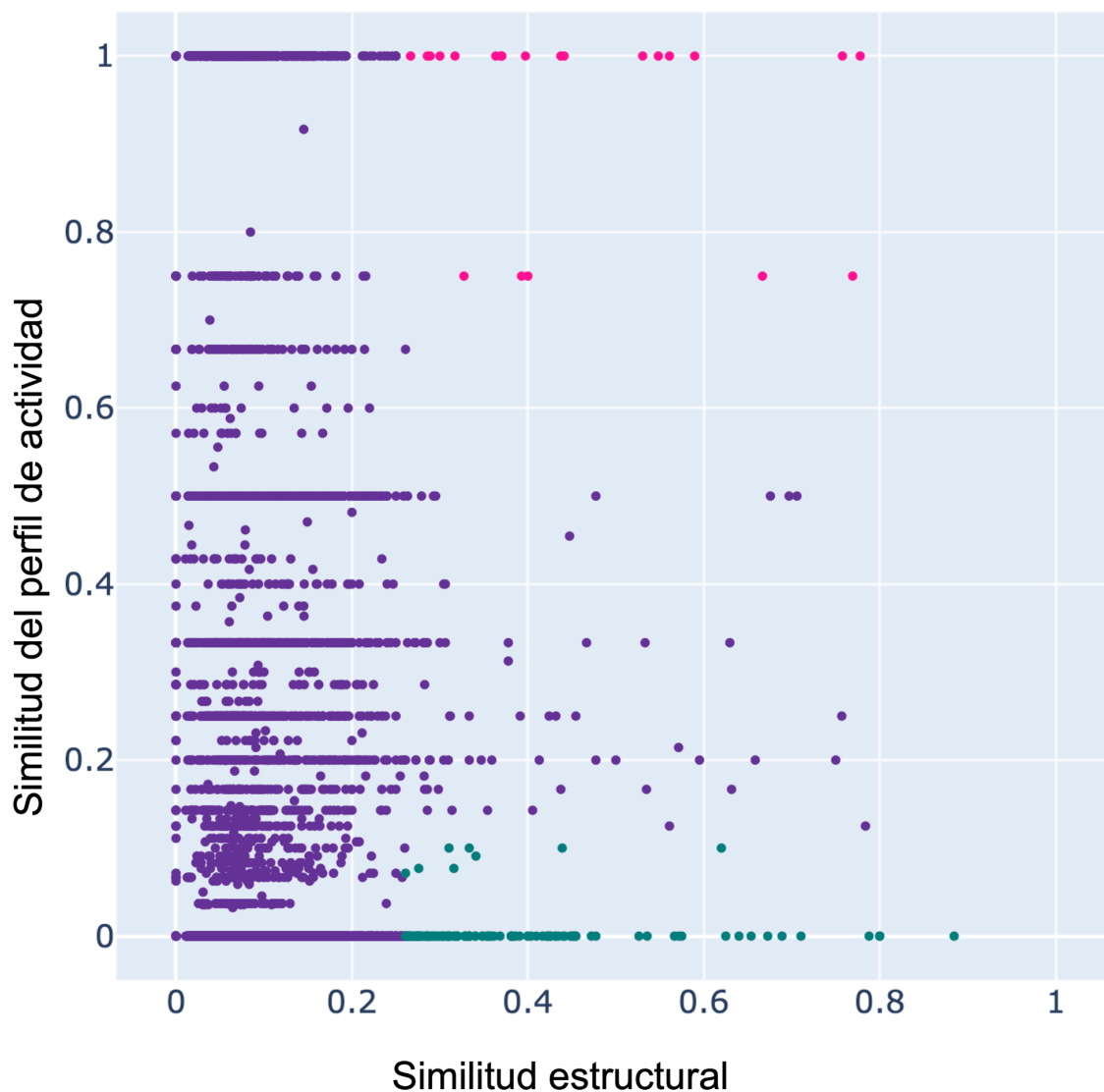


Figura 17. Mapa de similitud estructura-actividad (SAS) de los compuestos en *Epi food Chemical Database* usando la huella digital molecular ECFP de radio 2. En rosa están los pares compuestos con relación SAR y en verde están los pares compuestos con acantilados de actividad.

Tabla 2. Valores de similitud estructural (SE) y similitud del perfil de actividad (SPA) de los pares de compuestos SAR en la *Epi Food Chemical Database* con *Extended Connectivity Fingerprint* radio 2.

Par de compuestos	SE	SPA
Apigenina vs Luteolina	0.78	1.00
Ácido docosahexaenoico vs Ácido eicosapentaenoico	0.77	0.75
Cianidina vs Pelargonidina	0.76	1.00
Ácido eicosapentaenoico vs Ácido linoleico	0.67	0.75
Malvidina vs Pelargonidina	0.59	1.00
Cianidina vs Malvidina	0.56	1.00
Ácido docosahexaenoico vs Ácido linoleico	0.55	1.00
Floretina vs Florizina	0.53	1.00
Citrulina vs Teanina	0.44	1.00
Harmalin vs Harmina	0.44	1.00
Butirato vs Ácido docosahexaenoico	0.40	0.75
Colecalciferol vs Colesterol	0.40	1.00
Butirato vs Ácido linoleico	0.39	0.75
Butirato vs Ácido eicosapentaenoico	0.37	1.00
Luteína vs Retinol	0.37	1.00
Ácido linoleico vs Prostaglandina E2	0.36	1.00
Ácido eicosapentaenoico vs Prostaglandina E2	0.33	0.75
Isoquercitrina vs Wogonósido	0.32	1.00
Eugenol vs Tetrahidrocurcumina	0.30	1.00
Betacaroteno vs Licopeno	0.29	1.00
Betaína vs Colina	0.29	1.00
Ácido clorogénico vs Tetrahidrocurcumina	0.27	1.00

Tabla 3. Valores de similitud estructural (SE) y similitud del perfil de actividad (SPA) de los pares de compuestos acantilados de actividad en la *Epi Food Chemical Database* con *Extended Connectivity Fingerprint* radio 2.

Par de compuestos	SE	SPA
Ácido linoleico vs Ácido oleico	0.88	0.00
Apigenina vs Crisina	0.80	0.00

Betacaroteno vs Retinol	0.80	0.00
Buteína vs Isoliquiritigenina	0.79	0.00
Crisina vs Luteolina	0.71	0.00
Amentoflavona vs Apigenina	0.69	0.00
Berberina vs Epiberberina	0.67	0.00
Niacina vs Nicotinamida	0.65	0.00
Piceatannol vs Resveratrol	0.64	0.00
Amentoflavona vs luteolina	0.63	0.00
Curcumina vs Tetrahidrocurcumina	0.62	0.10
Apigenina vs Kaempferol	0.58	0.00
Amentoflavona vs Crisina	0.57	0.00
Ácido eicosapentaenoico vs Ácido oleico	0.57	0.00
Glutamato vs Teanina	0.54	0.00
Isoquercitrina vs Quercetina	0.53	0.00
Kaempferol vs Luteolina	0.48	0.00
Buteína vs Piceatannol	0.47	0.00
Crisina vs Kaempferol	0.45	0.00
Luteolina vs Miricetina	0.45	0.00
Epicatequina vs Galato de epigallocatequina	0.45	0.00
Ácido docosahexaenoico vs Ácido oleico	0.45	0.00
Grifolina vs Licopeno	0.45	0.00
Citrulina vs Glutamato	0.45	0.00
Apigenina vs Miricetina	0.44	0.00
Butirato vs Ácido oleico	0.44	0.00
Curcumina vs Eugenol	0.44	0.10
Crisina vs Quercetina	0.43	0.00
Crisina vs Miricetina	0.43	0.00
Betacaroteno vs Luteína	0.43	0.00
Amentoflavona vs Kaempferol	0.43	0.00
Fisetina vs Luteolina	0.43	0.00
Piceatannol vs Pterostilbeno	0.42	0.00
Isoquercitrina vs Florizina	0.42	0.00
Alfa-cetoglutarato vs Butirato	0.42	0.00
Isoquercitrina vs Kaempferol	0.42	0.00

Isoquercitrina vs Luteolina	0.41	0.00
Capsaicina vs Eugenol	0.40	0.00
Amentoflavona vs Genisteína	0.40	0.00
Isoquercitrina vs Miricetina	0.40	0.00
Tangeretina vs Wogonósido	0.39	0.00
Amentoflavona vs Quercetina	0.39	0.00
Ácido oleico vs Prostaglandina E2	0.38	0.00
Capsaicina vs Curcumina	0.38	0.00
Crisina vs Wogonósido	0.38	0.00
Fisetina vs Isoquercitrina	0.38	0.00
Ácido geranilgeranoico vs Licopeno	0.37	0.00
Alfa-viniferina vs Epicatequina	0.36	0.00
Isoliquiritigenina vs Pterostilbeno	0.36	0.00
Buteína vs Ácido clorogénico	0.36	0.00
Apigenina vs Daidzeína	0.36	0.00
Apigenina vs Fisetina	0.35	0.00
Oleaceína vs Oleuropeína	0.35	0.00
Genisteína vs Isoquercitrina	0.35	0.00
Butirato vs Glutamato	0.35	0.00
Curcumina vs Isoliquiritigenina	0.34	0.09
Cianidina vs Luteolina	0.34	0.00
Ácido clorogénico vs Piceatannol	0.34	0.00
Alfa-cetoglutarato vs Glutamato	0.33	0.00
Amentoflavona vs Miricetina	0.33	0.00
Apigenina vs Isoquercitrina	0.33	0.00
Astaxantina vs Betacaroteno	0.33	0.00
Curcumina vs Pterostilbeno	0.33	0.10
Oleuropeína vs Florizina	0.33	0.00
Isoquercitrina vs Oleuropeína	0.33	0.00
Epicatequina vs Silibinina	0.33	0.00
Astaxantina vs Luteína	0.33	0.00
Ácido cafeico vs fumarato	0.32	0.00
Crisina vs Tangeretina	0.32	0.00
Colesterol vs Guggulsterona	0.32	0.00

Amentoflavona vs Tangeretina	0.32	0.00
Buteína vs Resveratrol	0.32	0.08
Isoliquiritigenina vs Piceatannol	0.32	0.00
Florizina vs Sacarosa	0.31	0.00
Apigenina vs Pelargonidina	0.31	0.00
Ácido clorogénico vs Curcumina	0.31	0.10
Ácido gálico vs Niacina	0.31	0.00
Amarogentina vs Isoquercitrina	0.31	0.00
Crisina vs Isoquercitrina	0.31	0.00
Isoquercitrina vs Sacarosa	0.31	0.00
Ácido geranilgeranoico vs Grifolin	0.30	0.00
Apigenina vs Wogonósido	0.30	0.00
Ácido cafeico vs Resveratrol	0.30	0.00
D-antraquinonol vs Ácido geranilgeranoico	0.30	0.00
Amentoflavona vs Wogonósido	0.30	0.00
Indol-3-carbinol vs Melatonina	0.30	0.00
Ácido anacárdico vs Ácido linoleico	0.30	0.00
Crisina vs Fisetina	0.29	0.00
Amentoflavona vs Daidzeína	0.29	0.00
Eugenol vs Papaverina	0.29	0.00
Amentoflavona vs Isoquercitrina	0.29	0.00
Ácido docosahexaenoico vs Ácido geranilgeranoico	0.29	0.00
Betanina vs Isoquercitrina	0.29	0.00
Alfa-cetogluturato vs Ácido docosahexaenoico	0.29	0.00
Ácido anacárdico vs Ácido oleico	0.29	0.00
Daidzeína vs Luteolina	0.29	0.00
Harmina vs Melatonina	0.29	0.00
Capsaicina vs Tetrahydrocurcumina	0.28	0.00
Epicatequina vs Luteolina	0.28	0.00
Licopeno vs Retinol	0.28	0.00
Luteolina vs Piceatannol	0.28	0.00
Floretina vs Resveratrol	0.28	0.00
Astaxantina vs Licopeno	0.28	0.00
Alfa-cetogluturato vs Ácido oleico	0.28	0.08

Astaxantina vs Retinol	0.28	0.00
Epiberberina vs Peperomina E	0.28	0.00
Luteolina vs Wogonósido	0.28	0.00
Ácido eicosapentaenoico vs Ácido geranilgeranoico	0.28	0.00
Buteína vs Fisetina	0.27	0.00
Colesterol vs Solamargina	0.27	0.00
Buteína vs Pterostilbeno	0.27	0.00
Florizina vs Wogonósido	0.27	0.00
Colesterol vs Withaferin A	0.27	0.00
Pelargonidina vs Resveratrol	0.27	0.00
Alfa-viniferina vs Galato de epigallocatequina	0.27	0.00
Alfa-cetoglutarato vs Ácido eicosapentaenoico	0.27	0.00
Butirato vs Teanina	0.27	0.00
Epicatequina vs Piceatannol	0.27	0.00
Buteína vs Tetrahidrocurcumina	0.27	0.00
Crisina vs Daidzeína	0.27	0.00
Ácido cafeico vs Oleaceína	0.26	0.00
D-antroquinonol vs Licopeno	0.26	0.00
Piceatannol vs Tetrahidrocurcumina	0.26	0.00
Apigenina vs Ácido elágico	0.26	0.00
Eugenol vs Pterostilbeno	0.26	0.00
Ácido geranilgeranoico vs Ácido linoleico	0.26	0.00
Eugenol vs Silibinina	0.26	0.00
Alfa-cetoglutarato vs Betaína	0.26	0.00
Buteína vs Floretina	0.26	0.07

8.3.3.2 ECFP6

La Figura 18 muestra el mapa SAS para los pares de compuestos en la *Epi Food Chemical Database* usando la huella digital molecular ECFP6 para el cálculo de la similitud estructural.

La Tabla 4 presenta los veinticuatro pares de compuestos, marcados con puntos rosas en la Figura 18, que exhiben una relación simultánea entre su

estructura y su actividad, identificados como compuestos con un SAR continuo. Por otro lado, la Tabla 5 muestra los ciento seis pares de compuestos, representados por puntos verdes en la misma figura que tienen relación entre su estructura, pero no entre su perfil de actividad, también conocidos como compuestos acantilados de actividad. Los valores de similitud fueron calculados utilizando el coeficiente de Tanimoto, así como la huella digital molecular MACCS Keys, para el cálculo de la SE; y los vectores del perfil de actividad respectivos para cada compuesto para el cálculo de la SPA.

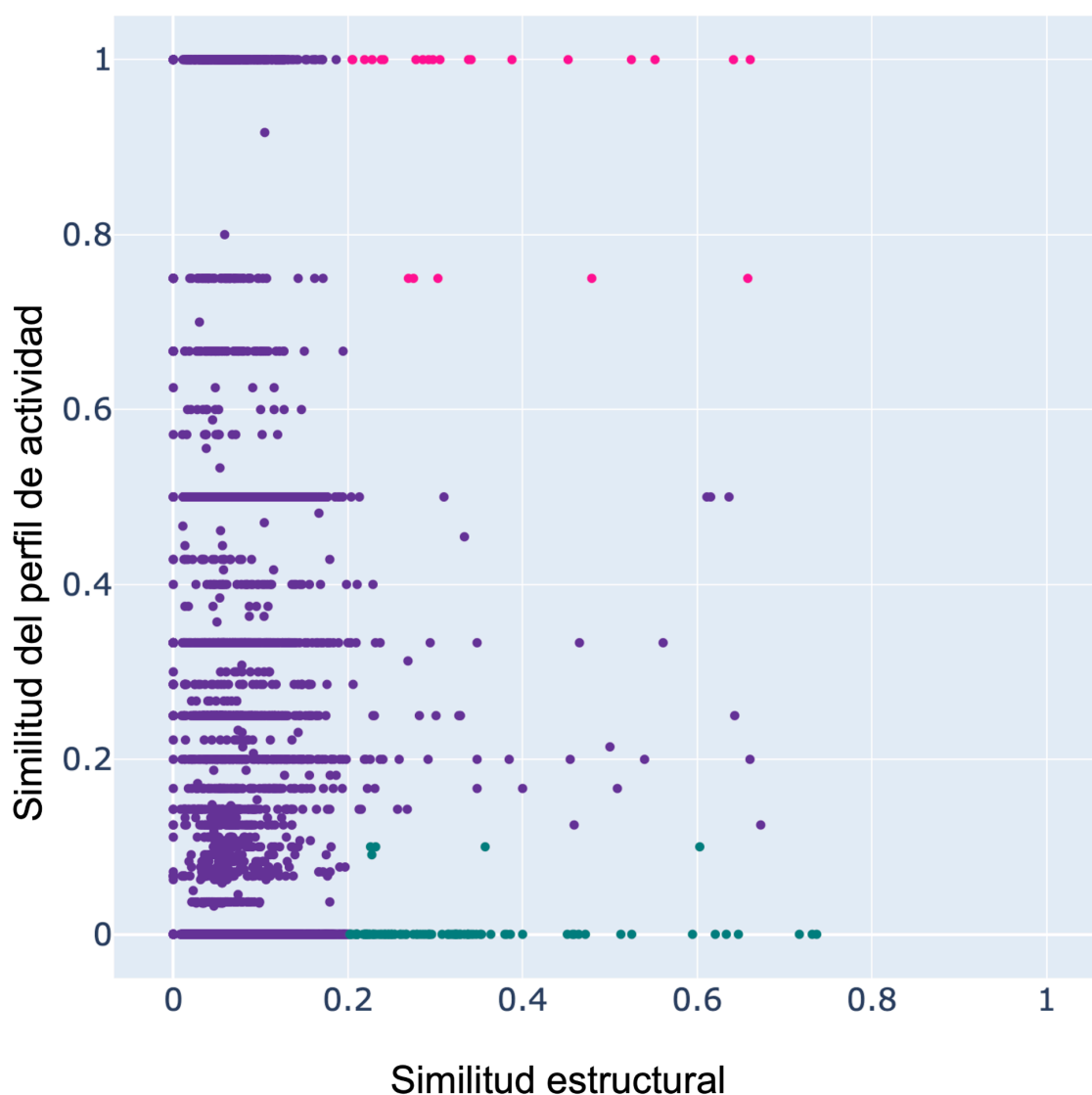


Figura 18. Mapa SAS de los compuestos en *Epi food Chemical Database* usando la huella digital molecular *Extended Connectivity Fingerprint* radio 3. En rosa están

los pares compuestos con relación SAR y en verde están los pares compuestos con acantilados de actividad.

Tabla 4. Valores de similitud estructural (SE) y similitud del perfil de actividad (SPA) de los pares de compuestos SAR en la *Epi Food Chemical Database* con *Extended Connectivity Fingerprint* radio 3.

Par de compuestos	SE	SPA
Apigenina vs Luteolina	0.66	1.00
Ácido docosahexaenoico vs Ácido eicosapentaenoico	0.66	0.75
Cianidina vs Pelargonidina	0.64	1.00
Malvidina vs Pelargonidina	0.55	1.00
Cianidina vs Malvidina	0.52	1.00
Ácido eicosapentaenoico (EPA) vs Ácido linoleico	0.48	0.75
Floretina vs Florizina	0.45	1.00
Ácido docosahexaenoico (DHA) vs Ácido linoleico	0.39	1.00
Citrulina vs Teanina	0.34	1.00
Harmalin vs Harmina	0.34	1.00
Luteína vs Retinol	0.31	1.00
Butirato vs Ácido docosahexaenoico	0.30	0.75
Colecalciferol vs Colesterol	0.30	1.00
Betacaroteno vs Licopeno	0.29	1.00
Betaína vs Colina	0.29	1.00
Butirato vs Ácido eicosapentaenoico	0.28	1.00
Butirato vs Ácido linoleico	0.28	0.75
Ácido eicosapentaenoico vs Prostaglandina E2	0.27	0.75
Eugenol vs Tetrahidrocurcumina	0.24	1.00
Ácido linoleico vs Prostaglandina E2	0.24	1.00
Isoquercitrina vs Wogonósido	0.23	1.00
Ácido clorogénico vs Tetrahidrocurcumina	0.22	1.00
Oleuropeína vs Sacarosa	0.21	1.00
D-antroquinonol vs Grifolin	0.20	1.00

Tabla 5. Valores de similitud estructural (SE) y similitud del perfil de actividad (SPA) de los pares de compuestos acantilados de actividad en la *Epi Food Chemical Database* con *Extended Connectivity Fingerprint* radio 3.

Par de compuestos	SE	SPA
Betacaroteno vs Retinol	0.74	0.00
Ácido linoleico vs Ácido oleico	0.73	0.00
Apigenina vs Crisina	0.72	0.00
Buteína vs Isoliquiritigenina	0.65	0.00
Niacina vs Nicotinamida	0.63	0.00
Crisina vs Luteolina	0.62	0.00
Curcumina vs Tetrahidrocurcumina	0.60	0.10
Amentoflavona vs Apigenina	0.59	0.00
Piceatannol vs Resveratrol	0.53	0.00
Amentoflavona vs Luteolina	0.51	0.00
Berberina vs Epiberberina	0.47	0.00
Isoquercitrina vs Quercetina	0.46	0.00
Apigenina vs Kaempferol	0.46	0.00
Glutamato vs Teanina	0.46	0.00
Amentoflavona vs Crisina	0.45	0.00
Betacaroteno vs Luteína	0.40	0.00
Grifolina vs Licopeno	0.39	0.00
Buteína vs Piceatannol	0.38	0.00
Epicatequina vs Galato de epigallocatequina	0.38	0.00
Alfa-cetoglutarato vs Butirato	0.36	0.00
Curcumina vs Eugenol	0.36	0.10
Kaempferol vs Luteolina	0.35	0.00
Citrulina vs Glutamato	0.35	0.00
Ácido eicosapentaenoico vs Ácido oleico	0.35	0.00
Crisina vs Kaempferol	0.34	0.00
Astaxantina vs Luteína	0.34	0.00
Capsaicina vs Eugenol	0.34	0.00
Isoquercitrina vs Kaempferol	0.34	0.00
Isoquercitrina vs Miricetina	0.34	0.00
Luteolina vs Miricetina	0.33	0.00

Astaxantina vs Betacaroteno	0.33	0.00
Apigenina vs Miricetina	0.33	0.00
Crisina vs Wogonósido	0.33	0.00
Isoquercitrina vs Florizina	0.32	0.00
Crisina vs Quercetina	0.32	0.00
Crisina vs Miricetina	0.32	0.00
Isoquercitrina vs Luteolina	0.32	0.00
Oleaceína vs Oleuropeína	0.32	0.00
Fisetina vs Isoquercitrina	0.32	0.00
Buteína vs Ácido clorogénico	0.32	0.00
Amentoflavona vs Kaempferol	0.31	0.00
Butirato vs Ácido oleico	0.31	0.00
Butirato vs Glutamato	0.31	0.00
Ácido clorogénico vs Piceatannol	0.30	0.00
Fisetina vs Luteolina	0.29	0.00
Isoliquiritigenina vs Pterostilbeno	0.29	0.00
Tangeretina vs Wogonósido	0.29	0.00
Piceatannol vs Pterostilbeno	0.29	0.00
Ácido docosahexaenoico vs Ácido oleico	0.29	0.00
Amentoflavona vs Genisteína	0.28	0.00
Astaxantina vs Licopeno	0.28	0.00
Capsaicina vs Curcumina	0.28	0.00
Isoquercitrina vs Oleuropeína	0.27	0.00
Alfa-cetoglutarato vs Glutamato	0.27	0.00
Ácido geranilgeranoico vs Licopeno	0.27	0.00
Genisteína vs Isoquercitrina	0.26	0.00
Amentoflavona vs Quercetina	0.26	0.00
D-antroquinonol vs Ácido geranilgeranoico	0.26	0.00
Apigenina vs Daidzeína	0.25	0.00
Ácido oleico vs Prostaglandina E2	0.25	0.00
Apigenina vs Isoquercitrina	0.25	0.00
Ácido gálico vs Niacina	0.25	0.00
Oleuropeína vs Florizina	0.25	0.00
Florizina vs Sacarosa	0.25	0.00

Ácido geranilgeranoico vs Grifolin	0.25	0.00
Ácido anacárdico vs Ácido oleico	0.25	0.00
Alfa-viniferina vs Epicatequina	0.24	0.00
Ácido cafeico vs Fumarato	0.24	0.00
Isoquercitrina vs Sacarosa	0.24	0.00
Cianidina vs Luteolina	0.24	0.00
Epicatequina vs Silibinina	0.24	0.00
Betanina vs Isoquercitrina	0.24	0.00
Crisina vs Isoquercitrina	0.23	0.00
Ácido clorogénico vs Surcumina	0.23	0.10
Amentoflavona vs Miricetina	0.23	0.00
Alfa-cetogluturato vs Betaína	0.23	0.00
Astaxantina vs Retinol	0.23	0.00
Butirato vs Teanina	0.23	0.00
Colesterol vs Guggulsterona	0.23	0.00
Curcumina vs Isoliquiritigenina	0.23	0.09
Curcumina vs Pterostilbeno	0.23	0.10
Apigenina vs Pelargonidina	0.23	0.00
Luteína vs Licopeno	0.22	0.00
Indol-3-carbinol vs Melatonina	0.22	0.00
Amentoflavona vs Tangeretina	0.22	0.00
Apigenina vs Wogonósido	0.22	0.00
Amentoflavona vs Wogonósido	0.22	0.00
Apigenina vs Fisetina	0.22	0.00
Crisina vs Tangeretina	0.22	0.00
Colesterol vs Solamargina	0.22	0.00
Licopeno vs Retinol	0.22	0.00
Capsaicina vs Tetrahidrocurcumina	0.22	0.00
Betaína vs Butirato	0.22	0.00
Amarogentina vs Isoquercitrina	0.21	0.00
Butirato vs Fumarato	0.21	0.00
Butirato vs Lactato	0.21	0.00
Ácido clorogénico vs Ácido salvianólico B	0.21	0.00
Epicatequina vs Luteolina	0.21	0.00

Ácido anacárdico vs Ácido linoleico	0.21	0.00
Ácido cafeico vs Oleaceína	0.21	0.00
Alfa-cetoglutarato vs Ácido docosahexaenoico	0.21	0.00
Amentoflavona vs Isoquercitrina	0.20	0.00
Betanina vs Oleuropeína	0.20	0.00
D-antroquinonol vs Licopeno	0.20	0.00
Alfa-viniferina vs Galato de epigallocatequina	0.20	0.00

8.3.3.3 MACCS Keys

La Figura 19 muestra el mapa SAS para los pares de compuestos en la *Epi Food Chemical Database* usando la huella digital molecular MACCS Keys para el cálculo de la similitud estructural.

La Tabla 6 presenta los quince pares de compuestos, marcados con puntos rosas en la Figura 19, que exhiben una relación simultánea entre su estructura y su actividad, identificados como compuestos con un SAR continuo. Por otro lado, la Tabla 7 muestra los ciento veintiocho pares de compuestos, representados por puntos verdes en la misma figura que tienen relación entre su estructura, pero no entre su perfil de actividad, también conocidos como compuestos acantilados de actividad. Los valores de similitud fueron calculados utilizando el coeficiente de Tanimoto, así como la huella digital molecular MACCS Keys, para el cálculo de la SE; y los vectores del perfil de actividad respectivos para cada compuesto para el cálculo de la SPA.

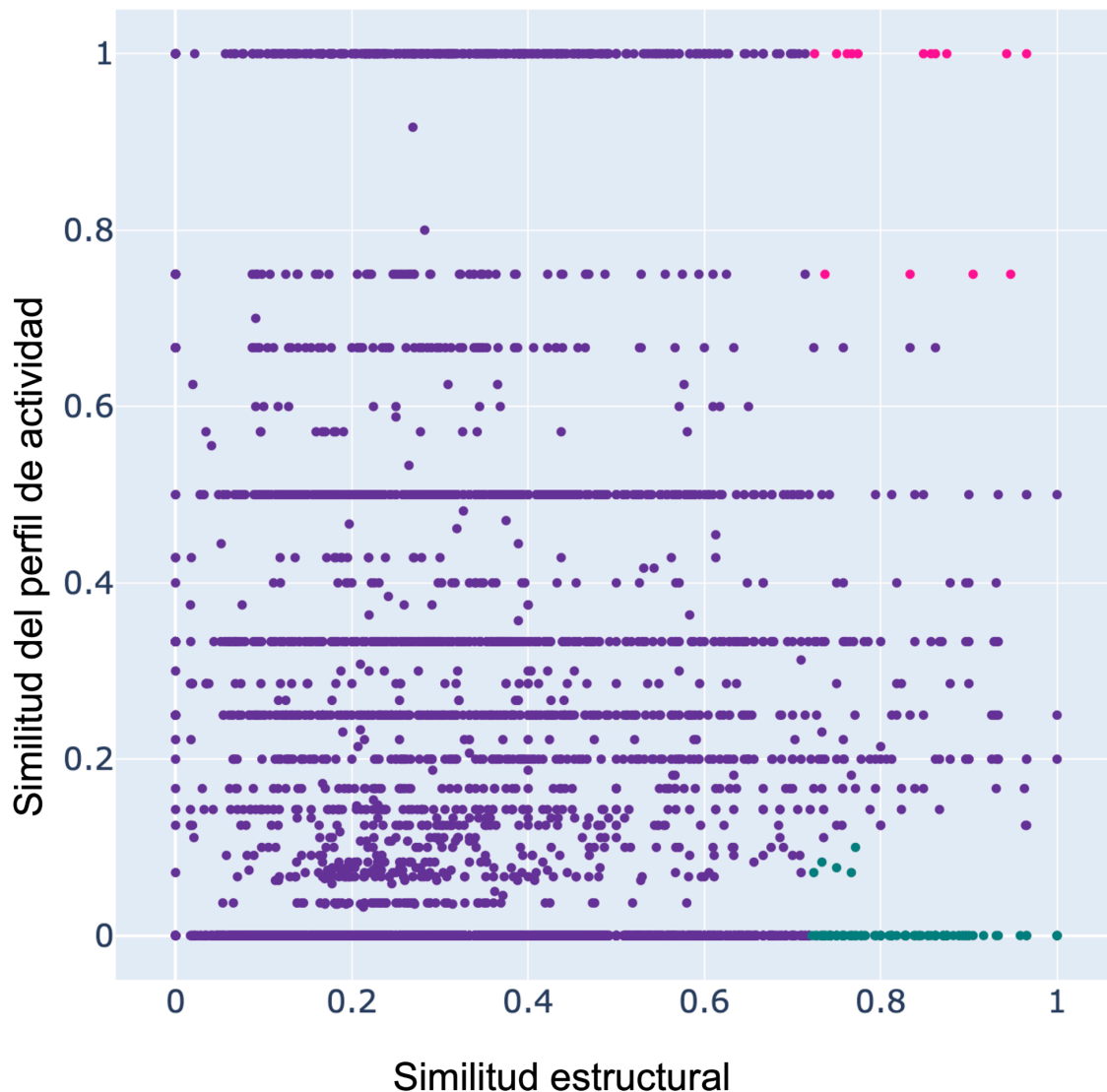


Figura 19. Mapa SAS de los compuestos en *Epi food Chemical Database* usando la huella digital molecular *Molecular ACCess System keys*. En rosa están los pares compuestos con relación SAR y en verde están los pares compuestos con acantilados de actividad.

Tabla 6. Valores de similitud estructural (SE) y similitud del perfil de actividad (SPA) de los pares de compuestos SAR en la *Epi Food Chemical Database* con *Molecular ACCess System keys*.

Par de compuestos	SE	SPA
Cianidina vs Pelargonidina	0.97	1.00
Ácido docosahexaenoico vs Ácido eicosapentaenoico	0.95	0.75

Colecalciferol vs Colesterol	0.94	1.00
Ácido eicosapentaenoico vs Ácido linoleico	0.90	0.75
Malvidina vs Pelargonidina	0.88	1.00
Apigenina vs Luteolina	0.86	1.00
Ácido docosahexaenoico vs Ácido linoleico	0.86	1.00
Cianidina vs Malvidina	0.85	1.00
Daidzeína vs Delfinidina	0.83	0.75
Alfa-viniferina vs Amentoflavona	0.77	1.00
Isoquercitrina vs Ácido salvianólico B	0.77	1.00
Isoquercitrina vs Wogonósido	0.76	1.00
Crisina vs Epicatequina	0.75	1.00
Butirato vs Ácido docosahexaenoico	0.74	0.75
Citrulina vs Teanina	0.73	1.00

Tabla 7. Valores de similitud estructural (SE) y similitud del perfil de actividad (SPA) de los pares de compuestos acantilados de actividad en la *Epi Food Chemical Database* con *Molecular ACCess System keys*.

Par de compuestos	SE	SPA
Apigenina vs Crisina	1.00	0.00
Berberina vs Epiberberina	1.00	0.00
Ácido linoleico vs Ácido oleico	1.00	0.00
Luteolina vs Miricetina	1.00	0.00
Fisetina vs Luteolina	0.97	0.00
Kaempferol vs Luteolina	0.97	0.00
Buteína vs Piceatannol	0.96	0.00
Cianidina vs Luteolina	0.93	0.00
Amentoflavona vs Genisteína	0.93	0.00
Buteína vs Isoliquiritigenina	0.92	0.00
Ácido eicosapentaenoico vs Ácido oleico	0.90	0.00
Amentoflavona vs Kaempferol	0.90	0.00
Luteolina vs Pelargonidina	0.90	0.00
Amentoflavona vs Daidzeína	0.90	0.00
Daidzeína vs Luteolina	0.90	0.00
Apigenina vs Kaempferol	0.89	0.00

Crisina vs Kaempferol	0.89	0.00
Apigenina vs Daidzeína	0.89	0.00
Crisina vs Daidzeína	0.89	0.00
Epicatequina vs Galato de epigallocatequina	0.88	0.00
Cianidina vs Epicatequina	0.88	0.00
Delfinidina vs Epicatequina	0.88	0.00
Epicatequina vs Luteolina	0.88	0.00
Epicatequina vs Miricetina	0.88	0.00
Epicatequina vs Quercetina	0.88	0.00
Isoliquiritigenina vs Piceatannol	0.88	0.00
Amentoflavona vs Ácido elágico	0.87	0.00
Amentoflavona vs Luteolina	0.87	0.00
Amentoflavona vs Miricetina	0.87	0.00
Amentoflavona vs Quercetina	0.87	0.00
Amentoflavona vs Apigenina	0.86	0.00
Amentoflavona vs Crisina	0.86	0.00
Apigenina vs Miricetina	0.86	0.00
Crisina vs Luteolina	0.86	0.00
Crisina vs Miricetina	0.86	0.00
Crisina vs Quercetina	0.86	0.00
Ácido docosahexaenoico vs Ácido oleico	0.86	0.00
Isoquercitrina vs Florizina	0.85	0.00
Isoquercitrina vs Silibinina	0.85	0.00
Galato de epigallocatequina vs Isoquercitrina	0.85	0.00
Epicatequina vs Fisetina	0.84	0.00
Epicatequina vs Kaempferol	0.84	0.00
Epicatequina vs Pelargonidina	0.84	0.00
Buteína vs Ácido Gálico	0.84	0.00
Amentoflavona vs Pelargonidina	0.84	0.00
Galato de epigallocatequina vs Wogonósido	0.84	0.00
Amentoflavona vs Wogonósido	0.83	0.00
Ácido elágico vs Wogonósido	0.83	0.00
Luteolina vs Wogonósido	0.83	0.00
Miricetina vs Wogonósido	0.83	0.00

Quercetina vs Wogonósido	0.83	0.00
Apigenina vs Fisetina	0.83	0.00
Apigenina vs Pelargonidina	0.83	0.00
Crisina vs Fisetina	0.83	0.00
Crisina vs Pelargonidina	0.83	0.00
Triacetato de naringenina vs Tangeretina	0.82	0.00
Amarogentina vs Isoquercitrina	0.81	0.00
Amentoflavona vs Cianidina	0.81	0.00
Amentoflavona vs Delfinidina	0.81	0.00
Ácido elágico (EA) vs Luteolina	0.81	0.00
Malvidin vs Wogonósido	0.81	0.00
Betacaroteno vs Betaelemeno	0.81	0.00
Apigenina vs Cianidina	0.80	0.00
Crisina vs Cianidina	0.80	0.00
Crisina vs Delfinidina	0.80	0.00
Fisetina vs Wogonósido	0.80	0.00
Kaempferol vs Wogonósido	0.80	0.00
Silibinina vs Wogonósido	0.80	0.00
Galato de epigallocatequina vs Ácido salvianólico B	0.79	0.00
Luteolina vs Malvidina	0.79	0.00
Alfa-viniferina vs Daidzeína	0.79	0.00
Piceatannol vs Resveratrol	0.78	0.00
Daidzeína vs Epicatequina	0.78	0.00
Amarogentin vs Oleuropeína	0.78	0.00
Cianidina vs Wogonósido	0.78	0.00
Delfinidina vs Wogonósido	0.78	0.00
Ácido gálico vs Inositol	0.77	0.00
Amentoflavona vs Galato de epigallocatequina	0.77	0.00
Curcumina vs Tetrahidrocurcumina	0.77	0.10
Genisteína vs Wogonósido	0.77	0.00
Malvidina vs Silibinina	0.77	0.00
Alfa-viniferina vs Genisteína	0.77	0.00
Amentoflavona vs Emodina	0.77	0.00
Buteína vs Miricetina	0.77	0.00

Buteína vs Quercetina	0.77	0.07
Amentoflavona vs Epicatequina	0.76	0.00
Alfa-viniferina vs Apigenina	0.76	0.00
Alfa-viniferina vs Crisina	0.76	0.00
Crisina vs galato de Epigallocatequina	0.76	0.00
Fisetina vs Tangeretina	0.76	0.00
Colesterol vs Guggulsterona	0.76	0.00
Kaempferol vs Silibinina	0.76	0.00
Tangeretina vs Wogonósido	0.76	0.00
Alfa-mangostin vs Shikonin	0.76	0.00
Apigenina vs Emodina	0.75	0.00
Apigenina vs Epicatequina	0.75	0.00
Buteína vs Resveratrol	0.75	0.08
Crisina vs Emodina	0.75	0.00
Epicatequina vs Isoquercitrina	0.75	0.00
Epicatequina vs Malvidina	0.75	0.00
Oleaceína vs Oleuropeína	0.75	0.00
Pelargonidina vs Wogonósido	0.75	0.00
Isoquercitrina vs Luteolina	0.74	0.00
Isoquercitrina vs Miricetina	0.74	0.00
Isoquercitrina vs Quercetina	0.74	0.00
Amentoflavona vs Malvidina	0.74	0.00
Daidzein vs Wogonósido	0.74	0.00
Alfa-viniferina vs Kaempferol	0.74	0.00
Alfa-viniferina vs Pelargonidina	0.74	0.00
Apigenina vs Ácido elágico	0.74	0.00
Crisina vs Ácido elágico	0.74	0.00
Inositol vs Piceatannol	0.74	0.00
Amentoflavona vs Silibinina	0.74	0.00
Epicatequina vs Wogonósido	0.74	0.00
Miricetina vs Silibinina	0.74	0.00
Miricetina vs Tangeretina	0.74	0.00
Oleuropeína vs Florizina	0.73	0.00
Buteína vs Fisetina	0.73	0.00

Buteína vs Ácido Lacaico	0.73	0.08
Luteolina vs Piceatannol	0.73	0.00
Miricetina vs Piceatannol	0.73	0.00
Alfa-mangostin vs Malvidin	0.73	0.00
Amentoflavona vs Ácido lacaico	0.73	0.00
Apigenina vs Malvidina	0.73	0.00
Crisina vs Malvidina	0.73	0.00
Alfa-mangostina vs tangeretina	0.73	0.00
Buteína vs Floretina	0.72	0.07
Papaverina vs Tangeretina	0.72	0.00

8.3.3.4 RDKit

La Figura 20 muestra el mapa SAS para los pares de compuestos en la *Epi Food Chemical Database* usando la huella digital molecular RDKit para el cálculo de la similitud estructural.

La Tabla 8 presenta los veintinueve pares de compuestos, marcados con puntos rosas en la Figura 20, que exhiben una relación simultánea entre su estructura y su actividad, identificados como compuestos con un SAR continuo. Por otro lado, la Tabla 9 muestra los ciento treinta y siete pares de compuestos, representados por puntos verdes en la misma figura que tienen relación entre su estructura, pero no entre su perfil de actividad, también conocidos como compuestos acantilados de actividad. Los valores de similitud fueron calculados utilizando el coeficiente de Tanimoto, así como la huella digital molecular RDKit, para el cálculo de la SE; y los vectores del perfil de actividad respectivos para cada compuesto para el cálculo de la SPA.

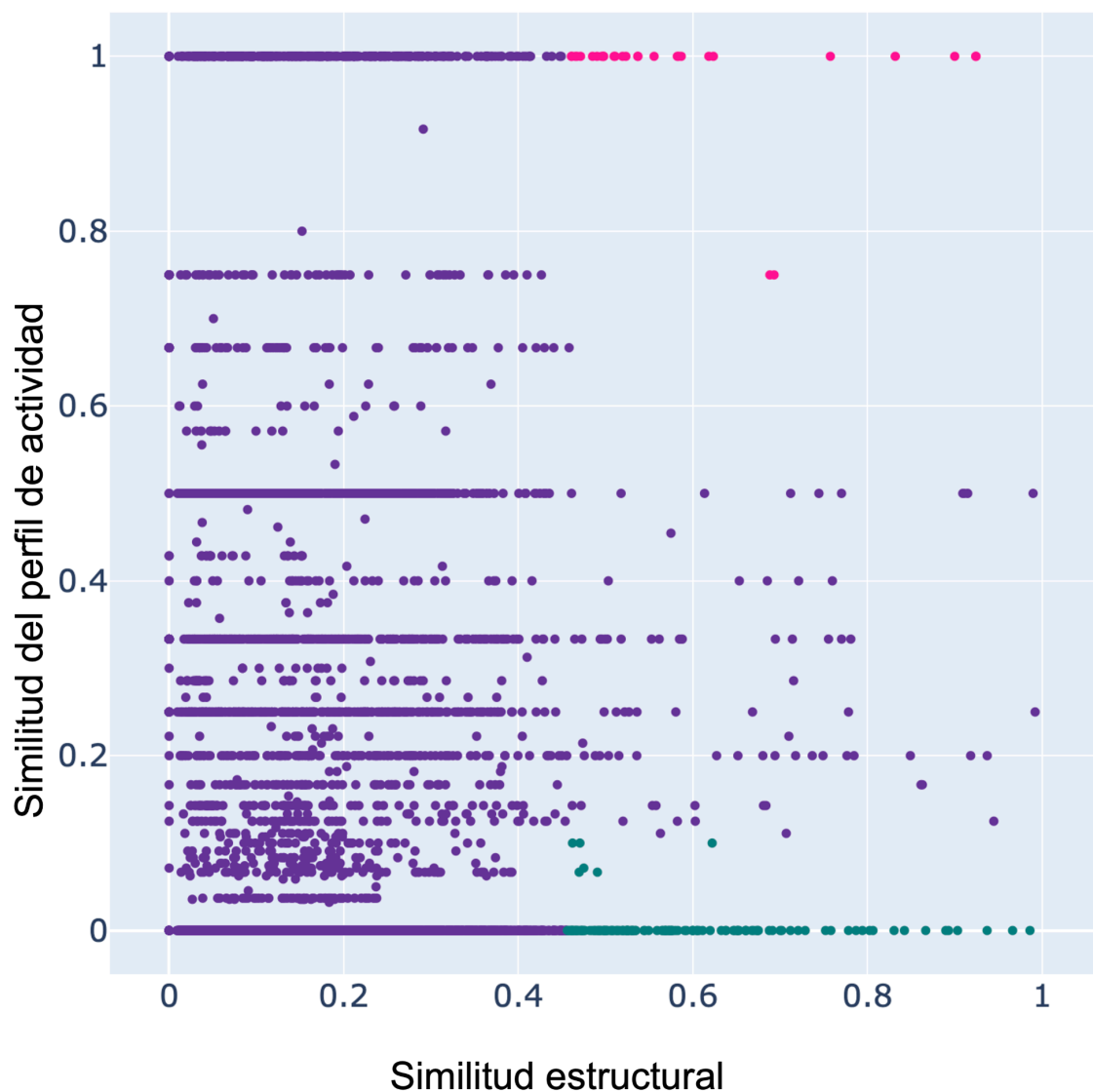


Figura 20. Mapa SAS de los compuestos en *Epi food Chemical Database* usando la huella digital molecular RDKit. En rosa están los pares compuestos con relación SAR y en verde están los pares compuestos con acantilados de actividad.

Tabla 8. Valores de similitud estructural (SE) y similitud del perfil de actividad (SPA) de los pares de compuestos SAR en la *Epi Food Chemical Database* con RDKit.

Par de compuestos	SE	SPA
Cianidina vs Pelargonidina	0.92	1.00
Apigenina vs Luteolina	0.92	1.00
Cianidina vs Malvidina	0.90	1.00

Malvidina vs Pelargonidina	0.83	1.00
Isoquercitrina vs Wogonósido	0.76	1.00
Ácido eicosapentaenoico vs Ácido linoleico	0.69	0.75
Ácido docosahexaenoico vs Ácido eicosapentaenoico	0.69	0.75
Colecalciferol vs Colesterol	0.62	1.00
Evodiamina vs Reserpina	0.62	1.00
Citrulina vs Teanina	0.59	1.00
Colecalciferol vs Luteína	0.58	1.00
Luteína vs Retinol	0.58	1.00
Betaína vs Colina	0.56	1.00
Amarogentin vs Betanin	0.54	1.00
Ácido docosahexaenoico vs Ácido linoleico	0.54	1.00
Alfa-mangostin vs Wogonoside	0.52	1.00
Ácido salvianólico B vs Wogonósido	0.52	1.00
Colesterol vs Luteína	0.51	1.00
Isoquercitrina vs Ácido salvianólico B	0.51	1.00
Reserpina vs Tanshinona IIA	0.50	1.00
Evodiamina vs Tanshinona IIA	0.50	1.00
Alfa-mangostin vs Isoquercitrina	0.50	1.00
D-antroquinonol vs Vitamina E	0.49	1.00
Fructosa vs Inositol	0.49	1.00
Amarogentin vs Solamargina	0.47	1.00
Epiberberina vs Ácido Salvianólico B	0.47	1.00
Alfa-mangostin vs Ácido salvianólico B	0.47	1.00
Epiberberina vs Isoquercitrina	0.46	1.00
Epiberberina vs Wogonósido	0.46	1.00

Tabla 9. Valores de similitud estructural (SE) y similitud del perfil de actividad (SPA) de los pares de compuestos acantilados de actividad en la *Epi Food Chemical Database* con RDKit.

Par de compuestos	SE	SPA
Berberina vs Epiberberina	0.99	0.00
Apigenina vs Crisina	0.97	0.00
Betacaroteno vs Retinol	0.94	0.00

Piceatannol vs Resveratrol	0.90	0.00
Crisina vs Luteolina	0.89	0.00
Buteína vs Isoliquiritigenina	0.89	0.00
Ácido linoleico vs Ácido oleico	0.87	0.00
Luteolina vs Miricetina	0.84	0.00
Apigenina vs Kaempferol	0.83	0.00
Kaempferol vs Luteolina	0.81	0.00
Crisina vs Kaempferol	0.80	0.00
Fisetina vs Luteolina	0.79	0.00
Epicatequina vs Galato de epigallocatequina	0.79	0.00
Apigenina vs Miricetina	0.78	0.00
Crisina vs Quercetina	0.76	0.00
Crisina vs Miricetina	0.75	0.00
Apigenina vs Fisetina	0.73	0.00
Piceatannol vs Pterostilbeno	0.72	0.00
Miricetina vs Tangeretina	0.71	0.00
Crisina vs Fisetina	0.70	0.00
Cianidina vs Luteolina	0.70	0.00
Amentoflavona vs Luteolina	0.69	0.00
Tangeretina vs Wogonósido	0.69	0.00
Isoquercitrina vs Quercetina	0.68	0.00
Amentoflavona vs Apigenina	0.67	0.00
Isoquercitrina vs Miricetina	0.67	0.00
Epicatequina vs Silibinina	0.67	0.00
Apigenina vs Pelargonidina	0.67	0.00
Amentoflavona vs Miricetina	0.66	0.00
Crisina vs Tangeretina	0.66	0.00
Amentoflavona vs Quercetina	0.66	0.00
Luteolina vs Malvidina	0.65	0.00
Amentoflavona vs Kaempferol	0.65	0.00
Amentoflavona vs Crisina	0.65	0.00
Luteolina vs Pelargonidina	0.65	0.00
Triacetato de naringenina vs Silibinina	0.65	0.00
Fisetina vs Tangeretina	0.65	0.00

Crisina vs Pelargonidina	0.64	0.00
Isoquercitrina vs Kaempferol	0.64	0.00
Apigenina vs Cianidina	0.63	0.00
Curcumina vs Tetrahidrocurcumina	0.62	0.10
Fisetina vs Isoquercitrina	0.62	0.00
Amarogentin vs Oleuropeína	0.61	0.00
Isoquercitrina vs Tangeretina	0.61	0.00
Crisina vs Cianidina	0.60	0.00
Betanina vs Wogonósido	0.60	0.00
Epicatequina vs Triacetato de naringenina	0.60	0.00
Crisina vs Delfinidina	0.60	0.00
Amentoflavona vs Tangeretina	0.60	0.00
Apigenina vs Daidzeína	0.59	0.00
Apigenina vs Malvidina	0.59	0.00
Ácido salvianólico B vs Silibinina	0.58	0.00
Crisina vs Daidzeína	0.58	0.00
Betanina vs Isoquercitrina	0.58	0.00
Reserpina vs Ácido salvianólico B	0.58	0.00
Amentoflavona vs Genisteína	0.58	0.00
Ácido eicosapentaenoico vs Ácido oleico	0.57	0.00
Isoquercitrina vs Luteolina	0.57	0.00
Betacaroteno vs Luteína	0.57	0.00
Amentoflavona vs Isoquercitrina	0.57	0.00
Galato de epigallocatequina vs Triacetato de naringenina	0.57	0.00
Quercetina vs Wogonósido	0.57	0.00
Daidzeína vs Luteolina	0.57	0.00
Miricetina vs Wogonósido	0.57	0.00
Crisina vs Malvidina	0.57	0.00
Betanina vs Ácido Salvianólico B	0.56	0.00
Kaempferol vs Wogonósido	0.56	0.00
Amentoflavona vs Wogonósido	0.55	0.00
Alfa-viniferina vs Ácido salvianólico B	0.55	0.00
Isoquercitrina vs Malvidina	0.54	0.00
Reserpina vs Wogonósido	0.53	0.00

Evodiamina vs Ácido Salvianólico B	0.53	0.00
Amentoflavona vs Delfinidina	0.53	0.00
Amentoflavona vs Cianidina	0.53	0.00
Galato de epigalocatequina vs Ácido salvianólico B	0.53	0.00
Apigenina vs Isoquercitrina	0.53	0.00
Isoquercitrina vs Reserpina	0.53	0.00
Amentoflavona vs Malvidina	0.53	0.00
Colesterol vs Guggulsterona	0.53	0.00
Amarogentina vs Isoquercitrina	0.52	0.00
Alfa-mangostina vs Tangeretina	0.52	0.00
Luteolina vs Wogonósido	0.52	0.00
Cianidina vs Isoquercitrina	0.52	0.00
Delfinidina vs Isoquercitrina	0.52	0.00
Amarogentin vs Wogonósido	0.52	0.00
Fisetina vs Wogonósido	0.52	0.00
Triacetato de naringenina vs Florizina	0.52	0.00
Amentoflavona vs Pelargonidina	0.52	0.00
Ácido cafeico vs Resveratrol	0.51	0.00
Apigenina vs Wogonósido	0.51	0.00
Crisina vs Isoquercitrina	0.51	0.00
Crisina vs Wogonósido	0.51	0.00
Florizina vs Wogonósido	0.51	0.00
Evodiamina vs Isoquercitrina	0.51	0.00
Aristoyagonina vs Evodiamina	0.51	0.00
Epiberberina vs Reserpina	0.50	0.00
Alfa-viniferina vs Silibinina	0.50	0.00
Aristoyagonina vs Reserpina	0.50	0.00
Reserpina vs Silibinina	0.50	0.00
Ácido salvianólico B vs Tanshinona IIA	0.50	0.00
Evodiamina vs Wogonósido	0.50	0.00
Betanina vs Silibinina	0.50	0.00
Amarogentina vs Ácido Salvianólico B	0.49	0.00
Amentoflavona vs Daidzeína	0.49	0.00
Aristoyagonina vs Ácido salvianólico B	0.49	0.00

Glutamato vs Teanina	0.49	0.00
Berberina vs Betanina	0.49	0.07
Aristoyagonina vs Betanina	0.49	0.00
Betanina vs Epiberberina	0.49	0.00
Harmalina vs Melatonina	0.49	0.00
Isoquercitrina vs Pelargonidina	0.48	0.00
Isoquercitrina vs Florizina	0.48	0.00
Malvidin vs Wogonósido	0.48	0.00
Alfa-mangostina vs Quercetina	0.48	0.00
Buteína vs Floretina	0.48	0.07
Alfa-mangostin vs Miricetina	0.47	0.00
Betanina vs Oleuropeína	0.47	0.00
Tanshinona IIA vs Wogonósido	0.47	0.00
Alfa-viniferina vs Galato de epigallocatequina	0.47	0.00
Curcumina vs Pterostilbeno	0.47	0.10
Amarogentin vs Berberina	0.47	0.07
Evodiamina vs Silibinina	0.47	0.00
Niacina vs Nicotinamida	0.47	0.00
Alfa-mangostin vs Kaempferol	0.47	0.00
Amarogentin vs Epiberberina	0.47	0.00
Berberina vs Acido salvianólico B	0.47	0.00
Oleuropeína vs Wogonósido	0.46	0.00
Berberina vs Wogonósido	0.46	0.00
Isoquercitrina frente a Tanshinona IIA	0.46	0.00
Curcumina vs Eugenol	0.46	0.10
Alfa-mangostin vs Amentoflavona	0.46	0.00
Berberina vs Isoquercitrina	0.46	0.00
Isoquercitrina vs Oleuropeína	0.46	0.00
Alfa-mangostin vs Luteolina	0.46	0.00
Epicatequina vs Ácido Salvianólico B	0.46	0.00
Oleuropeína vs Solamargina	0.46	0.00
Amarogentina vs Silibinina	0.46	0.00

8.3.3.5 Pares de compuestos consenso con relación estructura-actividad epigenética

Fueron identificados un total de diez pares de compuestos SAR que exhiben una relación simultánea entre su estructura y actividad simultáneamente para las cuatro huellas digitales moleculares (ECFP4, ECFP6, MACCS Keys y RDKit).

Los primeros tres pares de compuestos consenso con relación SAR continua son el ácido docosahexaenoico vs el ácido eicosapentaenoico; el ácido docosahexaenoico vs el ácido linoleico; y el ácido eicosapentaenoico vs el ácido linoleico. La Figura 21 muestra las estructuras y la relación de estos pares de compuestos.

El ácido docosahexaenoico y el ácido linoleico tienen actividad reportada vs DNMT1, DNMT3a, DNMT3b; mientras que el ácido eicosapentaenoico tiene actividad reportada vs DNMT1, DNMT3a (ver Figura 2), DNMT3b y HDAC1. A su vez, las principales fuentes alimentarias donde se encuentran presentes el ácido docosahexaenoico y el ácido eicosapentaenoico son los pescados grasos como el salmón, la caballa y las sardinas; en tanto que el ácido linoleico se encuentra esencialmente en aceites vegetales como el aceite de maíz, de soya y de girasol.¹³⁶⁻¹³⁸ Cabe destacar que, debido a la relación estructural entre estos tres compuestos es altamente probable que el ácido docosahexaenoico y el ácido linoleico también tengan actividad vs la diana epigenética HDAC1, lo que demostraría que en este caso se cumple una relación SAR. Debido a esto se sugiere llevar a cabo estudios experimentales para estudiar si esta predicción es correcta.

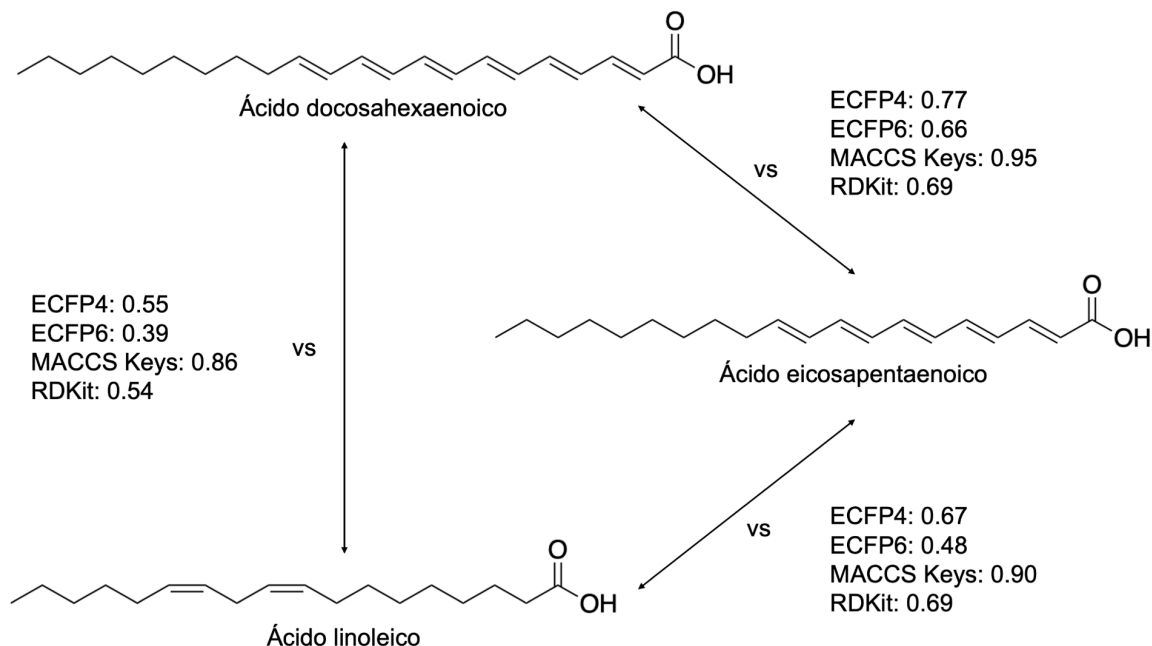


Figura 21. Estructuras químicas de los pares con relación SAR entre el ácido eicosapentaenoico, ácido docosahexaenoico y ácido linoleico; así como sus valores de similitud calculados obtenidos usando las huellas digitales moleculares: ECFP4 (*Extended Connectivity Fingerprint* radio 2), ECFP6 (*Extended Connectivity Fingerprint* radio 3), MACCS Keys (*Molecular ACCess System*) y RDKit (*RDKit Fingerprint*).

Otro par de compuestos que presentan relación SAR son la apigenina con la luteolina, ambos con actividad reportada vs HDAC1 y HDAC3. Estos se encuentran primordialmente en hierbas como el perejil, la manzanilla y el tomillo; y verduras como el apio, la cebolla y los pimientos.¹³⁹⁻¹⁴¹ La Figura 22 muestra las estructuras químicas de este par de compuestos.

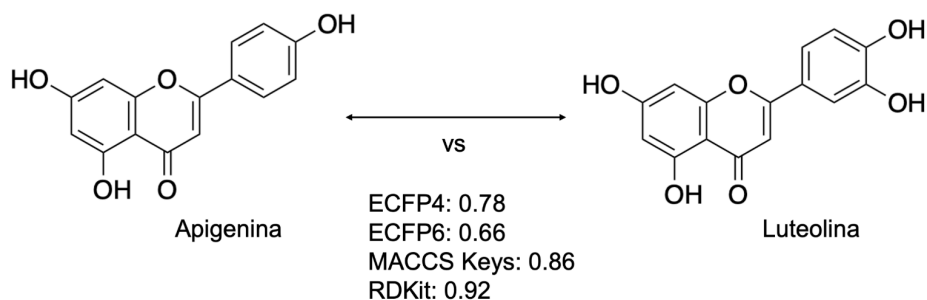


Figura 22. Estructuras químicas del par con relación SAR entre la apigenina y la luteolina; así como sus valores de similitud calculados obtenidos usando las huellas digitales moleculares: ECFP4 (*Extended Connectivity Fingerprint* radio 2), ECFP6 (*Extended Connectivity Fingerprint* radio 3), MACCS Keys (*Molecular ACCess System*) y RDKit (*RDKit Fingerprint*).

El siguiente par de compuestos con relación SAR identificados son el colecalciferol, también conocido como vitamina D3 y el colesterol. El colecalciferol se puede encontrar de diversos alimentos como pescados grasos (salmón, truchas, sardinas, atún, etc), la yema de huevo y algunos productos fortificados;¹⁴¹ por otra parte, algunas de las fuentes alimentarias del colesterol son alimentos de origen animal como el huevo, las vísceras y los productos lácteos ricos en grasas.¹⁴²⁻¹⁴³ Este par de compuestos tienen actividad reportada vs HDAC6. Las estructuras químicas del colecalciferol y el colesterol se muestran en la Figura 23.

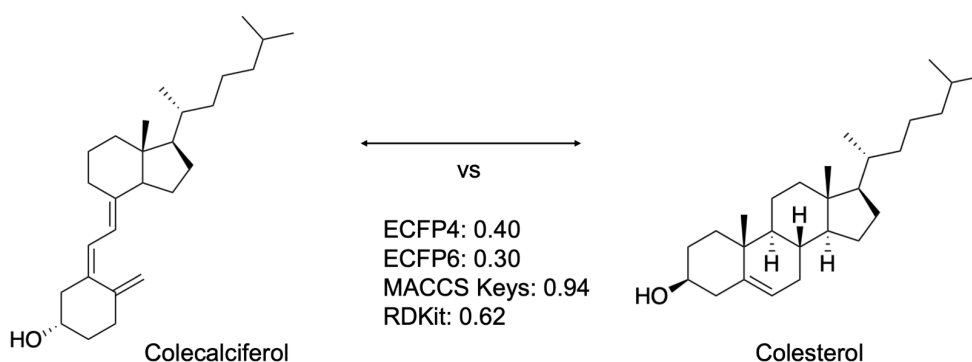


Figura 23. Estructuras químicas del par con relación SAR entre el colecalciferol y el colesterol; así como sus valores de similitud calculados obtenidos usando las huellas digitales moleculares: ECFP4 (*Extended Connectivity Fingerprint* radio 2),

ECFP6 (*Extended Connectivity Fingerprint* radio 3), MACCS Keys (*Molecular ACCess System*) y RDKit (*RDKit Fingerprint*).

A su vez, otro par de compuestos que tienen actividad reportada vs HDAC6 son los aminoácidos citrulina y teanina, mostrados en la Figura 24. La citrulina se puede encontrar en alimentos como legumbres y algunas verduras y frutas, principalmente en la sandía.¹⁴⁴⁻¹⁴⁵ Por otra parte, la teanina se encuentra en hojas de té verde como el matcha.¹⁴⁶⁻¹⁴⁷

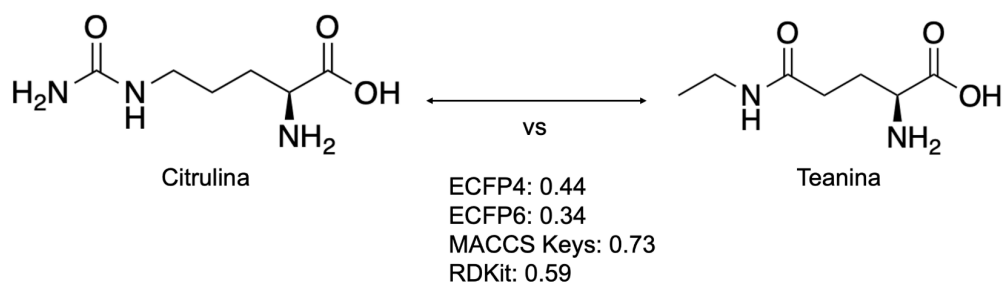


Figura 24. Estructuras químicas del par con relación SAR entre la citrulina y la teanina; así como sus valores de similitud calculados obtenidos usando las huellas digitales moleculares: ECFP4 (*Extended Connectivity Fingerprint* radio 2), ECFP6 (*Extended Connectivity Fingerprint* radio 3), MACCS Keys (*Molecular ACCess System*) y RDKit (*RDKit Fingerprint*).

Otra relación SAR que se identificó fue entre la isoquercitrina, un flavonoide polifenólico presente principalmente en frutas como las bayas, manzanas y uvas;¹⁴⁸ y el wogonósido, un flavonoide glicosilado proveniente de la *Scutellaria baicalensis*, una hierba asiática con propiedades medicinales.¹⁴⁹ Estos dos compuestos tienen actividad reportada vs LSD1.

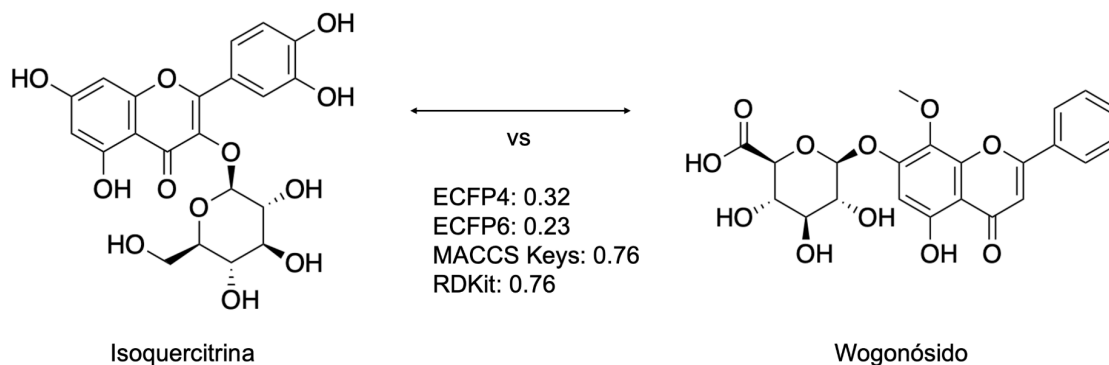


Figura 25. Estructuras químicas del par con relación SAR entre la isoquercitrina y el wogonósido; así como sus valores de similitud calculados obtenidos usando las huellas digitales moleculares: ECFP4 (*Extended Connectivity Fingerprint* radio 2), ECFP6 (*Extended Connectivity Fingerprint* radio 3), MACCS Keys (*Molecular ACCess System*) y RDKit (*RDKit Fingerprint*).

Los últimos tres pares de compuestos con actividad SAR se dan entre la cianidina, la pelargonidina y la malvidina. Estos tres compuestos se encuentran principalmente en frutas como las frambuesas, cerezas, aronias negras, moras, arándanos y algunas otras bayas. También se encuentran en el frijol y la col rizada.¹⁵⁰⁻¹⁵³ Estos compuestos tienen actividad epigenética reportada contra dos dianas epigenéticas: DNMT1 y DNMT3b.

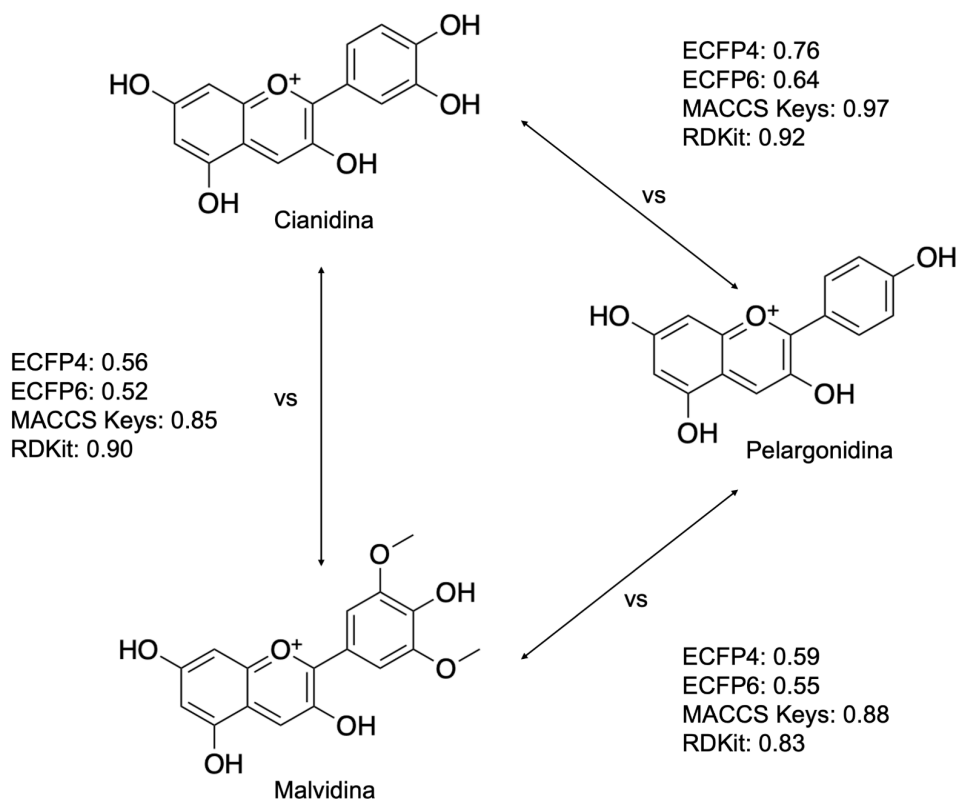


Figura 26. Estructuras químicas de los pares con relación SAR entre la cianidina, pelargonidina y malvidina; así como sus valores de similitud calculados obtenidos usando las huellas digitales moleculares: ECFP4 (*Extended Connectivity Fingerprint* radio 2), ECFP6 (*Extended Connectivity Fingerprint* radio 3), MAACS Keys (*Molecular ACCess System*) y RDKit (*RDKit Fingerprint*).

8.3.3.6 Pares de compuestos consenso potenciales acantilados de actividad (*activity cliffs*)

Se identificaron a un total de 36 pares de compuestos consenso para las cuatro huellas digitales moleculares (ECFP4, ECFP6, MACCS Keys y RDKit) en la región de los compuestos conocidos como acantilados de actividad: aquellos que presentan una relación simultánea entre su estructura, pero no en su actividad biológica. Sin embargo, es importante tomar en cuenta que tanto los cálculos de similitud en el perfil de actividad, así como los mapas SAS, se realizaron utilizando la información reportada en la literatura (sección 7.1). Por lo que un valor de cero

en el perfil de actividad no necesariamente significa que el compuesto no tiene actividad contra la respectiva diana epigenética. Existe la posibilidad de que la información no haya sido recopilada de acuerdo con la metodología empleada en esa investigación o que al momento no se hayan llevado a cabo ensayos experimentales de los compuestos contra esas dianas. Por lo tanto, estos pares de compuestos correspondientes a la región IV de la Figura 11 en este proyecto se denominan como “acantilados de actividad potenciales”. La Tabla 10 muestra los 35 pares de “acantilados de actividad potenciales”, así como el nombre de las dianas epigenéticas contra las cuales tienen actividad reportada, respectivamente.

Tabla 10. Pares de compuestos consenso potenciales acantilados de actividad (estructura similar y actividad biológica diferente), así como las dianas epigenéticas con las cuales tienen actividad asociada.

ID	Compuesto	Dianas epigenéticas asociadas	Compuesto	Dianas epigenéticas asociadas
1	Amarogentina	DNMT1	Isoquercitrina	LSD1
2	Amentoflavona	BET/BRD4	Apigenina	HDAC1 y HDAC3
3	Amentoflavona		Crisina	HDAC6
4	Amentoflavona		Genisteína	DNMT1, HDAC1, HDAC5 y HDAC6
5	Amentoflavona		Kaempferol	DNMT1, DNMT3b, HDAC4, HDAC7 y HDAC8
6	Amentoflavona		Luteolina	HDAC1 y HDAC3
7	Amentoflavona		Miricetina	DNMT1
8	Amentoflavona		Quercetina	DNMT1, DNMT3a, HDAC1 y SIRT1
9	Amentoflavona		Wogonósido	LSD1
10	Apigenina		HDAC1 y HDAC3	Crisina
11	Apigenina	Daidzeína		DNMT1, DNMT3a y DNMT3b
12	Apigenina	Fisetina		DNMT1 y

				BET/BRD4
13	Apigenina		Kaempferol	DNMT1, DNMT3b, HDAC4, HDAC7 y HDAC8
14	Apigenina		Miricetina	DNMT1
15	Apigenina		Pelargonidina	DNMT1 y DNMT3b
16	Berberina	DNMT1, DNMT3a, DNMT3b, HDAC1, HDAC2, HDAC3, HDAC4, HDAC5, HDAC6, HDAC7, HDAC8, HDAC9, HDAC10, SIRT3 y KDM6A	Epiberberina	LSD1
17	Buteína	HDAC1, HDAC2, HDAC3, HDAC4, HDAC5, HDAC6, HDAC7, HDAC8, HDAC9, HDAC10 y HDAC11	Isoliquiritigenina	DNMT1 y BET/BRD4
18	Colesterol	HDAC6	Gugulsterona	DNMT1, DNMT3a, DNMT3b y HDAC1
19	Crisina		Kaempferol	DNMT1, DNMT3b, HDAC4, HDAC7 y HDAC8
20	Crisina	HDAC6	Luteolina	HDAC1 y HDAC3
21	Crisina		Miricetina	DNMT1
22	Crisina		Quercetina	DNMT1, DNMT3a, HDAC1 y SIRT1
23	Curcumina	DNMT1, DNMT3a, DNMT3b, HAT/Ep300, HDAC1, HDAC3, HDAC3, HDAC4, HDAC8 y SIRT1	Tetrahidrocurcumina	DNMT1
24	Cianidina	DNMT1 y DNMT3b	Luteolina	HDAC1 y HDAC3
25	Ácido eicosapentaenoico	DNMT1, DNMT3a, DNMT3b y HDAC1	Ácido oleico	KDM4A

26	Epicatequina	HDAC6	Galato de epicatequina	DNMT1, HDAC1, HDAC2, HDAC3, y HDAC8
27	Fisetina	DNMT1 y BET/BRD4	Luteolina	HDAC1 y HDAC3
28	Isoquercitrina	LSD1	Miricetina	DNMT1
29	Isoquercitrina		Florizina	DNMT1, DNMT3a, DNMT3b y HDAC1
30	Isoquercitrina		Quercetina	DNMT1, DNMT3a, HDAC1 y SIRT1
31	Isoquercitrina		Luteolina	HDAC1 y HDAC3
32	Kaempferol	DNMT1, DNMT3b, HDAC4, HDAC7 y HDAC8	Ácido oleico	KDM4A
33	Ácido linoleico	DNMT1, DNMT3a y DNMT3b	Miricetina	DNMT1
34	Luteolina	HDAC1 y HDAC3	Resveratrol	HDAC3, SIRT1 y BET/BRD4
35	Piceatanol	DNMT3a y DNMT3b	Wogonósido	LSD1
36	Tangeretina	DNMT3b, HDAC1, HDAC2, HDAC4, HDAC5 y HDAC9		

Estos potenciales acantilados de actividad pueden servir para proponer predicciones de los panoramas de actividad de los compuestos con el objetivo de identificar futuras posibles áreas de interés, permitiendo priorizar la evaluación experimental de estos enfocada en dianas específicas.¹⁵⁴ Algunos ejemplos de estas propuestas se describen a continuación.

Como se mencionó en la sección 8.3.3.5, la apigenina y luteolina tienen relación SAR con actividad reportada vs HDAC1 y HDAC3. Y algunos compuestos potenciales acantilados de actividad tanto con la apigenina y luteolina simultáneamente son la crisina, el kaempferol, la fisetina y la miricetina. La crisina tiene actividad reportada vs HDAC6; el kaempferol vs DNMT1, DNMT3b, HDAC4, HDAC7 y HDAC8; la fisetina vs DNMT1 y BET/BRD4; y la miricetina vs DNMT1.

Por lo que se sugiere explorar la posible actividad de la apigenina y la luteolina vs DNMT1, DNMT3b, HDAC4, HDAC7 y HDAC8, así como la de la crisina, el kaempferol, la fisetina y la miricetina vs HDAC1 y HDAC3.

Otros potenciales acantilados de actividad con interés sobresaliente son entre el ácido oleico y los ácidos eicosapentanoico y linoleico. El ácido oleico tiene actividad reportada vs KDM4A, mientras que el ácido eicosapentanoico y linoleico son un par de compuestos con relación SAR con actividad reportada vs DNMT1, DNMT3a, DNMT3b. Por lo tanto, el análisis del ácido oleico vs DNMT1, DNMT3a y DNMT3b, así como el de los ácidos eicosapentanoico y linoleico vs KDM4A, también son casos de relevancia especial.

Estos ejemplos particulares descritos presentan oportunidades sustanciales para llevar a cabo investigaciones adicionales, destacando la relevancia de explorar la actividad de compuestos específicos frente a dianas epigenéticas particulares. La identificación y análisis de estos potenciales acantilados de actividad abren la oportunidad de explorar nuevas vías para la comprensión y la predicción de la actividad epigenética mediante la optimización de estrategias quimioinformáticas en el diseño de compuestos con fines terapéuticos.

Es importante destacar que los mecanismos de acción de los 187 compuestos dietéticos y productos naturales son altamente específicos, estando determinados por la función y el mecanismo epigenético de cada una de las 46 dianas respectivas; así como el hecho de que varios de estos compuestos muestran actividad contra múltiples dianas epigenéticas simultáneamente. Los mecanismos específicos asociados a cada compuesto y diana con actividad epigenética pueden explorarse con mayor detalle en las versiones de la *Epi Food Chemical Database* a través de los respectivos DOIs asociados a los artículos que detallan las investigaciones realizadas entre el 2017 y 2023 de estos, disponibles en la sección de anexos.

9. Conclusiones

➤ Se construyó y curó la base de datos *Epi Food Chemical Database*, que contiene 187 compuestos químicos presentes en alimentos y productos naturales. La base de datos contiene información estructural y el perfil de actividad frente a 46 dianas epigenéticas.

➤ A partir análisis quimioinformáticos de la *Epi Food Chemical Database*, se ha observado una alta diversidad estructural entre los compuestos, identificando diez núcleos base que representan una proporción significativa (35.54%) de la base de datos. Además, la visualización del espacio químico mostró que los compuestos en esta base de datos y en FooDB tienden a ocupar regiones cercanas, lo que sugiere similitudes en sus propiedades químicas, así como posibles perfiles de actividad moleculares en común.

➤ Se ha identificado la presencia de dos grupos principales de compuestos: uno exhibe relaciones continuas entre su estructura y su actividad epigenética (relación SAR), mientras que el otro muestra potenciales acantilados de actividad, con estructuras muy similares pero perfiles de actividad epigenética muy diferentes.

➤ La identificación de estos potenciales acantilados de actividad ofrece una oportunidad prometedora para futuras investigaciones, ya que sugiere la posibilidad de que compuestos con estructuras similares presenten actividad epigenética semejante. Esto podría facilitar la selección de candidatos para pruebas experimentales destinadas a evaluar su actividad epigenética y su potencial aplicación en el desarrollo de estrategias terapéuticas.

10. Perspectivas

➤ Establecer colaboraciones estrechas con grupos de investigación experimentales especializados en el ámbito de la nutri epigenómica. El objetivo principal es validar la actividad epigenética de los compuestos identificados en la base de datos mediante experimentos y análisis empíricos. Se buscará dar prioridad a la validación de las relaciones estructura-actividad, centrándose particularmente en casos de potenciales acantilados de actividad.

➤ Implementar un sistema de mantenimiento continuo para asegurar la actualización constante de la *Epi Food Chemical Database* a medida que se desarrollen y publiquen nuevos hallazgos e investigaciones en el campo. Este proceso garantizará que la base de datos refleje de manera precisa el estado más reciente de la investigación en nutri epigenómica, abarcando tanto descubrimientos anteriores a 2017 como aquellos posteriores a marzo de 2023. De esta manera, se asegurará que la base de datos refleje de manera precisa el estado más actualizado de la investigación en el campo, garantizando su exhaustividad y completitud en el contenido.

11. Referencias

1. Dupont, C.; Armant, D.; Brenner, C. Epigenetics: Definition, Mechanisms and Clinical Perspective. *Seminars in Reproductive Medicine* **2009**, *27* (05), 351–357. DOI:10.1055/s-0029-1237423.
2. Lacal, I.; Ventura, R. Epigenetic Inheritance: Concepts, Mechanisms and Perspectives. *Frontiers in Molecular Neuroscience* **2018**, *11*. DOI:10.3389/fnmol.2018.00292.
3. Li, Y. Modern Epigenetics Methods in Biological Research. *Methods* **2021**, *187*, 104–113. DOI:10.1016/j.ymeth.2020.06.022.
4. Peixoto, P.; Cartron, P.-F.; Serandour, A. A.; Hervouet, E. From 1957 to Nowadays: A Brief History of Epigenetics. *International Journal of Molecular Sciences* **2020**, *21* (20), 7571. DOI:10.3390/ijms21207571.
5. Jones, P. A. Functions of DNA Methylation: Islands, Start Sites, Gene Bodies and Beyond. *Nature Reviews Genetics* **2012**, *13* (7), 484–492. DOI:10.1038/nrg3230.
6. Bestor, T. H. The DNA Methyltransferases of Mammals. *Human Molecular Genetics* **2000**, *9* (16), 2395–2402. DOI:10.1093/hmg/9.16.2395.
7. Lyko, F. The DNA Methyltransferase Family: A Versatile Toolkit for Epigenetic Regulation. *Nature Reviews Genetics* **2017**, *19* (2), 81–92. DOI:10.1038/nrg.2017.80.
8. Robertson, K. D. DNA Methylation and Human Disease. *Nature Reviews Genetics* **2005**, *6* (8), 597–610. DOI:10.1038/nrg1655.
9. Suzuki, M. M.; Bird, A. DNA Methylation Landscapes: Provocative Insights from Epigenomics. *Nature Reviews Genetics* **2008**, *9* (6), 465–476. DOI:10.1038/nrg2341.
10. Bird, A. DNA Methylation Patterns and Epigenetic Memory. *Genes & Development* **2002**, *16* (1), 6–21. DOI:10.1101/gad.947102.
11. Luger, K.; Mäder, A. W.; Richmond, R. K.; Sargent, D. F.; Richmond, T. J. Crystal Structure of the Nucleosome Core Particle at 2.8 Å Resolution. *Nature* **1997**, *389* (6648), 251–260. DOI:10.1038/38444.

12. Yusufaly, T. I.; Li, Y.; Singh, G.; Olson, W. K. Arginine-Phosphate Salt Bridges between Histones and DNA: Intermolecular Actuators That Control Nucleosome Architecture. *The Journal of Chemical Physics* **2014**, *141* (16). DOI:10.1063/1.4897978.
13. Rooijers, K.; Markodimitraki, C. M.; Rang, F. J.; de Vries, S. S.; Chialastri, A.; de Luca, K. L.; Mooijman, D.; Dey, S. S.; Kind, J. Simultaneous Quantification of Protein–DNA Contacts and Transcriptomes in Single Cells. *Nature Biotechnology* **2019**, *37* (7), 766–772. DOI:10.1038/s41587-019-0150-y.
14. Strahl, B. D.; Allis, C. D. The Language of Covalent Histone Modifications. *Nature* **2000**, *403* (6765), 41–45. DOI:10.1038/47412.
15. Kouzarides, T. Chromatin Modifications and Their Function. *Cell* **2007**, *128* (4), 693–705. DOI:10.1016/j.cell.2007.02.005.
16. Roth, S. Y.; Denu, J. M.; Allis, C. D. Histone Acetyltransferases. *Annual Review of Biochemistry* **2001**, *70* (1), 81–120. DOI:10.1146/annurev.biochem.70.1.81.
17. Jenuwein, T.; Allis, C. D. Translating the Histone Code. *Science* **2001**, *293* (5532), 1074–1080. DOI:10.1126/science.1063127.
18. Esteller, M. Non-Coding RNAs in Human Disease. *Nature Reviews Genetics* **2011**, *12* (12), 861–874. DOI:10.1038/nrg3074.
19. Rinn, J. L.; Chang, H. Y. Genome Regulation by Long Noncoding RNAs. *Annual Review of Biochemistry* **2012**, *81* (1), 145–166. DOI:10.1146/annurev-biochem-051410-092902.
20. Morris, K. V.; Mattick, J. S. The Rise of Regulatory RNA. *Nature Reviews Genetics* **2014**, *15* (6), 423–437. DOI:10.1038/nrg3722.
21. Marchese, F. P.; Huarte, M. Long Non-Coding RNAs and Chromatin Modifiers. *Epigenetics* **2013**, *9* (1), 21–26. DOI:10.4161/epi.27472.
22. Gupta, R. A.; Shah, N.; Wang, K. C.; Kim, J.; Horlings, H. M.; Wong, D. J.; Tsai, M.-C.; Hung, T.; Argani, P.; Rinn, J. L.; Wang, Y.; Brzoska, P.; Kong, B.; Li, R.; West, R. B.; van de Vijver, M. J.; Sukumar, S.; Chang, H. Y. Long Non-Coding RNA HOTAIR Reprograms Chromatin State to Promote Cancer Metastasis. *Nature* **2010**, *464* (7291), 1071–1076. DOI:10.1038/nature08975.

23. He, R.; Kidder, B. L. H3K4 Demethylase KDM5B Regulates Global Dynamics of Transcription Elongation and Alternative Splicing in Embryonic Stem Cells. *Nucleic Acids Research* **2017**, *45* (11), 6427–6441. DOI:10.1093/nar/gkx251.
24. Statello, L.; Guo, C.-J.; Chen, L.-L.; Huarte, M. Gene Regulation by Long Non-Coding RNAs and Its Biological Functions. *Nature Reviews Molecular Cell Biology* **2020**, *22* (2), 96–118. DOI:10.1038/s41580-020-00315-9.
25. Mello, C. C.; Conte, D. Revealing the World of RNA Interference. *Nature* **2004**, *431* (7006), 338–342. DOI:10.1038/nature02872.
26. Fire, A.; Xu, S.; Montgomery, M. K.; Kostas, S. A.; Driver, S. E.; Mello, C. C. Potent and Specific Genetic Interference by Double-Stranded RNA in *Caenorhabditis Elegans*. *Nature* **1998**, *391* (6669), 806–811. DOI:10.1038/35888.
27. Meister, G.; Landthaler, M.; Patkaniowska, A.; Dorsett, Y.; Teng, G.; Tuschl, T. Human Argonaute2 Mediates RNA Cleavage Targeted by Mirnas and Sirnas. *Molecular Cell* **2004**, *15* (2), 185–197. DOI:10.1016/j.molcel.2004.07.007.
28. Bartel, D. P. MicroRNAs. *Cell* **2004**, *116* (2), 281–297. DOI:10.1016/s0092-8674(04)00045-5.
29. Kim, V. N.; Han, J.; Siomi, M. C. Biogenesis of Small RNAs in Animals. *Nature Reviews Molecular Cell Biology* **2009**, *10* (2), 126–139. DOI:10.1038/nrm2632.
30. Chen, L.-L. The Biogenesis and Emerging Roles of Circular RNAs. *Nature Reviews Molecular Cell Biology* **2016**, *17* (4), 205–211. DOI:10.1038/nrm.2015.32.
31. Vo, J. N.; Cieslik, M.; Zhang, Y.; Shukla, S.; Xiao, L.; Zhang, Y.; Wu, Y.-M.; Dhanasekaran, S. M.; Engelke, C. G.; Cao, X.; Robinson, D. R.; Nesvizhskii, A. I.; Chinnaiyan, A. M. The Landscape of Circular RNA in Cancer. *Cell* **2019**, *176* (4). DOI:10.1016/j.cell.2018.12.021.
32. Kristensen, L. S.; Andersen, M. S.; Stagsted, L. V.; Ebbesen, K. K.; Hansen, T. B.; Kjems, J. The Biogenesis, Biology and Characterization of Circular RNAs. *Nature Reviews Genetics* **2019**, *20* (11), 675–691. DOI:10.1038/s41576-019-0158-7.
33. Marmorstein, R. Structure and Function of Histone Acetyltransferases. *Cellular and Molecular Life Sciences* **2001**, *58* (5), 693–703. DOI:10.1007/pl00000893.

34. Witt, O.; Deubzer, H. E.; Milde, T.; Oehme, I. HDAC Family: What Are the Cancer Relevant Targets? *Cancer Letters* **2009**, *277* (1), 8–21. DOI:10.1016/j.canlet.2008.08.016.
35. Gong, F.; Miller, K. M. Histone Methylation and the DNA Damage Response. *Mutation Research/Reviews in Mutation Research* **2019**, *780*, 37–47. DOI:10.1016/j.mrrev.2017.09.003.
36. Rossetto, D.; Avvakumov, N.; Côté, J. Histone Phosphorylation. *Epigenetics* **2012**, *7* (10), 1098–1108. DOI:10.4161/epi.21975.
37. Boros, I. M. Histone Modification in Drosophila. *Briefings in Functional Genomics* **2012**, *11* (4), 319–331. DOI:10.1093/bfgp/els029.
38. Toma-Fukai, S.; Shimizu, T. Structural Diversity of Ubiquitin E3 Ligase. *Molecules* **2021**, *26* (21), 6682. DOI:10.3390/molecules26216682.
39. Snyder, N. A.; Silva, G. M. Deubiquitinating Enzymes (Dubs): Regulation, Homeostasis, and Oxidative Stress Response. *Journal of Biological Chemistry* **2021**, *297* (3), 101077. DOI:10.1016/j.jbc.2021.101077.
40. Biswas, S.; Rao, C. M. Epigenetic Tools (the Writers, the Readers and the Erasers) and Their Implications in Cancer Therapy. *European Journal of Pharmacology* **2018**, *837*, 8–24. DOI:10.1016/j.ejphar.2018.08.021.
41. Arrowsmith, C. H.; Bountra, C.; Fish, P. V.; Lee, K.; Schapira, M. Epigenetic Protein Families: A New Frontier for Drug Discovery. *Nature Reviews Drug Discovery* **2012**, *11* (5), 384–400. DOI:10.1038/nrd3674.
42. Baylin, S. B.; Jones, P. A. A Decade of Exploring the Cancer Epigenome — Biological and Translational Implications. *Nature Reviews Cancer* **2011**, *11* (10), 726–734. DOI:10.1038/nrc3130.
43. McBride, A. E. 3 Diverse Roles of Protein Arginine Methyltransferases. *Protein Methyltransferases* **2006**, 51–103. DOI:10.1016/s1874-6047(06)80005-3.
44. Tahiliani, M.; Koh, K. P.; Shen, Y.; Pastor, W. A.; Bandukwala, H.; Brudno, Y.; Agarwal, S.; Iyer, L. M.; Liu, D. R.; Aravind, L.; Rao, A. Conversion of 5-Methylcytosine to 5-Hydroxymethylcytosine in Mammalian DNA by MLL Partner TET1. *Science* **2009**, *324* (5929), 930–935. DOI:10.1126/science.1170116.

45. Goll, M. G.; Bestor, T. H. Eukaryotic Cytosine Methyltransferases. *Annual Review of Biochemistry* **2005**, *74* (1), 481–514. DOI:10.1146/annurev.biochem.74.010904.153721.
46. Jones, P. A.; Liang, G. Rethinking How DNA Methylation Patterns Are Maintained. *Nature Reviews Genetics* **2009**, *10* (11), 805–811. DOI:10.1038/nrg2651.
47. Margueron, R.; Reinberg, D. The Polycomb Complex PRC2 and Its Mark in Life. *Nature* **2011**, *469* (7330), 343–349. DOI:10.1038/nature09784.
48. Di Croce, L.; Helin, K. Transcriptional Regulation by Polycomb Group Proteins. *Nature Structural & Molecular Biology* **2013**, *20* (10), 1147–1155. DOI:10.1038/nsmb.2669.
49. Yap, D. B.; Chu, J.; Berg, T.; Schapira, M.; Cheng, S.-W. G.; Moradian, A.; Morin, R. D.; Mungall, A. J.; Meissner, B.; Boyle, M.; Marquez, V. E.; Marra, M. A.; Gascoyne, R. D.; Humphries, R. K.; Arrowsmith, C. H.; Morin, G. B.; Aparicio, S. A. Somatic Mutations at EZH2 Y641 Act Dominantly through a Mechanism of Selectively Altered PRC2 Catalytic Activity, to Increase H3K27 Trimethylation. *Blood* **2011**, *117* (8), 2451–2459. DOI:10.1182/blood-2010-11-321208.
50. Bedford, M. T.; Clarke, S. G. Protein Arginine Methylation in Mammals: Who, What, and Why. *Molecular Cell* **2009**, *33* (1), 1–13. DOI:10.1016/j.molcel.2008.12.013.
51. Yang, Y.; Bedford, M. T. Protein Arginine Methyltransferases and Cancer. *Nature Reviews Cancer* **2012**, *13* (1), 37–50. DOI:10.1038/nrc3409.
52. Blanc, R. S.; Richard, S. Arginine Methylation: The Coming of Age. *Molecular Cell* **2017**, *65* (1), 8–24. DOI:10.1016/j.molcel.2016.11.003.
53. Allis, C. D.; Berger, S. L.; Cote, J.; Dent, S.; Jenuwien, T.; Kouzarides, T.; Pillus, L.; Reinberg, D.; Shi, Y.; Shiekhhattar, R.; Shilatifard, A.; Workman, J.; Zhang, Y. New Nomenclature for Chromatin-Modifying Enzymes. *Cell* **2007**, *131* (4), 633–636. DOI:10.1016/j.cell.2007.10.039.
54. Sterner, D. E.; Berger, S. L. Acetylation of Histones and Transcription-Related Factors. *Microbiology and Molecular Biology Reviews* **2000**, *64* (2), 435–459. DOI:10.1128/mmb.64.2.435-459.2000.

55. Bannister, A. J.; Kouzarides, T. Regulation of Chromatin by Histone Modifications. *Cell Research* **2011**, *21* (3), 381–395. DOI:10.1038/cr.2011.22.
56. McGee-Lawrence, M. E.; Westendorf, J. J. Histone Deacetylases in Skeletal Development and Bone Mass Maintenance. *Gene* **2011**, *474* (1–2), 1–11. DOI:10.1016/j.gene.2010.12.003.
57. Guarente, L. Calorie Restriction and Sirtuins Revisited. *Genes & Development* **2013**, *27* (19), 2072–2085. DOI:10.1101/gad.227439.113.
58. Shi, Y.; Lan, F.; Matson, C.; Mulligan, P.; Whetstine, J. R.; Cole, P. A.; Casero, R. A.; Shi, Y. Histone Demethylation Mediated by the Nuclear Amine Oxidase Homolog LSD1. *Cell* **2004**, *119* (7), 941–953. DOI:10.1016/j.cell.2004.12.012.
59. Haberland, M.; Montgomery, R. L.; Olson, E. N. The Many Roles of Histone Deacetylases in Development and Physiology: Implications for Disease and Therapy. *Nature Reviews Genetics* **2009**, *10* (1), 32–42. DOI:10.1038/nrg2485.
60. Seto, E.; Yoshida, M. Erasers of Histone Acetylation: The Histone Deacetylase Enzymes. *Cold Spring Harbor Perspectives in Biology* **2014**, *6* (4). DOI:10.1101/cshperspect.a018713.
61. Choudhary, C.; Kumar, C.; Gnad, F.; Nielsen, M. L.; Rehman, M.; Walther, T. C.; Olsen, J. V.; Mann, M. Lysine Acetylation Targets Protein Complexes and Co-Regulates Major Cellular Functions. *Science* **2009**, *325* (5942), 834–840. DOI:10.1126/science.1175371.
62. Chalkiadaki, A.; Guarente, L. Sirtuins Mediate Mammalian Metabolic Responses to Nutrient Availability. *Nature Reviews Endocrinology* **2012**, *8* (5), 287–296. DOI:10.1038/nrendo.2011.225.
63. Houtkooper, R. H.; Pirinen, E.; Auwerx, J. Sirtuins as Regulators of Metabolism and Healthspan. *Nature Reviews Molecular Cell Biology* **2012**, *13* (4), 225–238. DOI:10.1038/nrm3293.
64. Imai, S.; Guarente, L. NAD⁺ and Sirtuins in Aging and Disease. *Trends in Cell Biology* **2014**, *24* (8), 464–471. DOI:10.1016/j.tcb.2014.04.002.
65. Lee, M. G.; Wynder, C.; Schmidt, D. M.; McCafferty, D. G.; Shiekhattar, R. Histone H3 Lysine 4 Demethylation Is a Target of Nonselective Antidepressive

- Medications. *Chemistry & Biology* **2006**, *13* (6), 563–567. DOI:10.1016/j.chembiol.2006.05.004.
66. Metzger, E.; Wissmann, M.; Yin, N.; Müller, J. M.; Schneider, R.; Peters, A. H.; Günther, T.; Buettner, R.; Schüle, R. LSD1 Demethylates Repressive Histone Marks to Promote Androgen-Receptor-Dependent Transcription. *Nature* **2005**, *437* (7057), 436–439. DOI:10.1038/nature04020.
67. Kooistra, S. M.; Helin, K. Molecular Mechanisms and Potential Functions of Histone Demethylases. *Nature Reviews Molecular Cell Biology* **2012**, *13* (5), 297–311. DOI:10.1038/nrm3327.
68. Lan, F.; Shi, Y. Epigenetic Regulation: Methylation of Histone and Non-Histone Proteins. *Science in China Series C: Life Sciences* **2009**, *52* (4), 311–322. DOI:10.1007/s11427-009-0054-z.
69. Filippakopoulos, P.; Knapp, S. Targeting Bromodomains: Epigenetic Readers of Lysine Acetylation. *Nature Reviews Drug Discovery* **2014**, *13* (5), 337–356. DOI:10.1038/nrd4286.
70. Wu, S.-Y.; Lee, A.-Y.; Lai, H.-T.; Zhang, H.; Chiang, C.-M. Phospho Switch Triggers BRD4 Chromatin Binding and Activator Recruitment for Gene-Specific Targeting. *Molecular Cell* **2013**, *49* (5), 843–857. DOI:10.1016/j.molcel.2012.12.006.
71. Montalvo-Casimiro, M.; González-Barrios, R.; Meraz-Rodriguez, M. A.; Juárez-González, V. T.; Arriaga-Canon, C.; Herrera, L. A. Epidrug Repurposing: Discovering New Faces of Old Acquaintances in Cancer Therapy. *Frontiers in Oncology* **2020**, *10*. DOI:10.3389/fonc.2020.605386.
72. El Omari, N.; Bakrim, S.; Bakha, M.; Lorenzo, J. M.; Rebezov, M.; Shariati, M. A.; Aboulaghras, S.; Balahbib, A.; Khayrullin, M.; Bouyahya, A. Natural Bioactive Compounds Targeting Epigenetic Pathways in Cancer: A Review on Alkaloids, Terpenoids, Quinones, and Isothiocyanates. *Nutrients* **2021**, *13* (11), 3714. DOI:10.3390/nu13113714.
73. Dubey, R.; Prabhakar, P. K.; Gupta, J. Epigenetics: Key to Improve Delayed Wound Healing in Type 2 Diabetes. *Molecular and Cellular Biochemistry* **2021**, *477* (2), 371–383. DOI:10.1007/s11010-021-04285-0.

74. Sultana, F.; Manasa, K. L.; Shaik, S. P.; Bonam, S. R.; Kamal, A. Zinc Dependent Histone Deacetylase Inhibitors in Cancer Therapeutics: Recent Update. *Current Medicinal Chemistry* **2020**, *26* (40), 7212–7280. DOI:10.2174/0929867325666180530094120.
75. Liu, J.; Lang, G.; Shi, J. Epigenetic Regulation of PDX-1 in Type 2 Diabetes Mellitus. *Diabetes, Metabolic Syndrome and Obesity: Targets and Therapy* **2021**, *Volume 14*, 431–442. DOI:10.2147/dmso.s291932.
76. Xu, S.; Pelisek, J.; Jin, Z. G. Atherosclerosis Is an Epigenetic Disease. *Trends in Endocrinology & Metabolism* **2018**, *29* (11), 739–742. DOI:10.1016/j.tem.2018.04.007.
77. Rabaneda-Bueno, R.; Mena-Montes, B.; Torres-Castro, S.; Torres-Carrillo, N.; Torres-Carrillo, N. M. Advances in Genetics and Epigenetic Alterations in Alzheimer's Disease: A Notion for Therapeutic Treatment. *Genes* **2021**, *12* (12), 1959. DOI:10.3390/genes12121959.
78. Neele, A. E.; Willemsen, L.; Chen, H.-J.; Dzobo, K. E.; de Winther, M. P. J. Targeting Epigenetics as Atherosclerosis Treatment: An Updated View. *Current Opinion in Lipidology* **2020**, *31* (6), 324–330. DOI:10.1097/mol.0000000000000711.
79. Kowluru, R. A.; Mohammad, G. Epigenetic Modifications in Diabetes. *Metabolism* **2022**, *126*, 154920. DOI:10.1016/j.metabol.2021.154920.
80. Punnia-Moorthy, G.; Hersey, P.; Emran, A. A.; Tiffen, J. Lysine Demethylases: Promising Drug Targets in Melanoma and Other Cancers. *Frontiers in Genetics* **2021**, *12*. DOI:10.3389/fgene.2021.680633.
81. van Dijk, S. J.; Molloy, P. L.; Varinli, H.; Morrison, J. L.; Muhlhausler, B. S.; Buckley, M.; Clark, S. J.; McMillen, I. C.; Noakes, M.; Samaras, K.; Tellam, R. L. Epigenetics and Human Obesity. *International Journal of Obesity* **2014**, *39* (1), 85–97. DOI:10.1038/ijo.2014.34.
82. Niculescu, M. D.; Zeisel, S. H. Diet, Methyl Donors and DNA Methylation: Interactions between Dietary Folate, Methionine and Choline. *The Journal of Nutrition* **2002**, *132* (8). DOI:10.1093/jn/132.8.2333s.
83. Ma, Y.; Olendzki, B. C.; Merriam, P. A.; Chiriboga, D. E.; Culver, A. L.; Li, W.; Hébert, J. R.; Ockene, I. S.; Griffith, J. A.; Pagoto, S. L. A Randomized Clinical

Trial Comparing Low-Glycemic Index versus ADA Dietary Education among Individuals with Type 2 Diabetes. *Nutrition* **2008**, *24* (1), 45–56. DOI:10.1016/j.nut.2007.10.008.

84. Niculescu, M. D. Nutritional Epigenetics. *ILAR Journal* **2012**, *53* (3–4), 270–278. DOI:10.1093/ilar.53.3-4.270.

85. Tiffon, C. The Impact of Nutrition and Environmental Epigenetics on Human Health and Disease. *International Journal of Molecular Sciences* **2018**, *19* (11), 3425. DOI:10.3390/ijms19113425.

86. Saldívar-González, F.; Prieto-Martínez, F. D.; Medina-Franco, J. L. Descubrimiento y Desarrollo de Fármacos: Un Enfoque Computacional. *Educación Química* **2017**, *28* (1), 51–58. DOI:10.1016/j.eq.2016.06.002.

87. Bunin, B. A.; Siesel, B.; Morales, G. A.; Bajorath, J. Chemoinformatics: Theory, Practice, & Products. *Springer Dordrecht* **2010**, *1* (1), 295. DOI:10.1007/1-4020-5001-1.

88. Heller, S. R.; McNaught, A.; Pletnev, I.; Stein, S.; Tchekhovskoi, D. InChI, the IUPAC International Chemical Identifier. *Journal of Cheminformatics* **2015**, *7* (1). DOI:10.1186/s13321-015-0068-4.

89. Weininger, D. Smiles, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *Journal of Chemical Information and Computer Sciences* **1988**, *28* (1), 31–36. DOI:10.1021/ci00057a005.

90. Gasteiger, J. Chemistry in Times of Artificial Intelligence. *ChemPhysChem* **2020**, *21* (20), 2233–2242. DOI:10.1002/cphc.202000518.

91. López-López, E.; Bajorath, J.; Medina-Franco, J. L. Informatics for Chemistry, Biology, and Biomedical Sciences. *Journal of Chemical Information and Modeling* **2020**, *61* (1), 26–35. DOI:10.1021/acs.jcim.0c01301.

92. Williams, A. J.; Ekins, S. A Quality Alert and Call for Improved Curation of Public Chemistry Databases. *Drug Discovery Today* **2011**, *16* (17–18), 747–750. DOI:10.1016/j.drudis.2011.07.007.

93. Irwin, J. J.; Tang, K. G.; Young, J.; Dandarchuluun, C.; Wong, B. R.; Khurelbaatar, M.; Moroz, Y. S.; Mayfield, J.; Sayle, R. A. Zinc20—a Free Ultralarge-Scale Chemical Database for Ligand Discovery. *Journal of Chemical*

- Information and Modeling* **2020**, *60* (12), 6065–6073.
DOI:10.1021/acs.jcim.0c00675.
94. Sorokina, M.; Merseburger, P.; Rajan, K.; Yirik, M. A.; Steinbeck, C. Coconut Online: Collection of Open Natural Products Database. *Journal of Cheminformatics* **2021**, *13* (1). DOI:10.1186/s13321-020-00478-9.
95. Gómez-García, A.; Jiménez, D. A.; Zamora, W. J.; Barazorda-Ccahuana, H. L.; Chávez-Fumagalli, M.; Valli, M.; Andricopulo, A. D.; Bolzani, V. da; Olmedo, D. A.; Solís, P. N.; Núñez, M. J.; Rodríguez Pérez, J. R.; Valencia Sánchez, H. A.; Cortés Hernández, H. F.; Medina-Franco, J. L. Navigating the Chemical Space and Chemical Multiverse of a Unified Latin American Natural Product Database: LANaPDB. *Pharmaceuticals* **2023**, *16* (10), 1388. DOI:10.3390/ph16101388.
96. <https://foodb.ca/>
97. Dimova, D.; Stumpfe, D.; Bajorath, J. Computational Design of New Molecular Scaffolds for Medicinal Chemistry, Part II: Generalization of Analog Series-Based Scaffolds. *Future Science OA* **2018**, *4* (2). DOI:10.4155/fsoa-2017-0102.
98. Mishra, N.; Tiwari, V. K.; Schmidt, R. R. Recent Trends and Challenges on Carbohydrate-Based Molecular Scaffolding: General Consideration toward Impact of Carbohydrates in Drug Discovery and Development. *Carbohydrates in Drug Discovery and Development* **2020**, 1–69. DOI:10.1016/b978-0-12-816675-8.00001-4.
99. Varnek, A.; Baskin, I. I. Chemoinformatics as a Theoretical Chemistry Discipline. *Molecular Informatics* **2011**, *30* (1), 20–32. DOI:10.1002/minf.201000100.
100. Bajorath, J. Chemoinformatics. Concepts, Methods, and Tools for Drug Discovery. *Methods in Molecular Biology* **2004**, *1*. DOI:10.1385/1592598021.
101. Osolodkin, D. I.; Radchenko, E. V.; Orlov, A. A.; Voronkov, A. E.; Palyulin, V. A.; Zefirov, N. S. Progress in Visual Representations of Chemical Space. *Expert Opinion on Drug Discovery* **2015**, *10* (9), 959–973. DOI:10.1517/17460441.2015.1060216.
102. Pezzotti, N.; Lelieveldt, B. P.; Maaten, L. van; Holtt, T.; Eisemann, E.; Vilanova, A. Approximated and User Steerable Tsne for Progressive Visual

- Analytics. *IEEE Transactions on Visualization and Computer Graphics* **2017**, *23* (7), 1739–1752. DOI:10.1109/tvcg.2016.2570755.
103. Xu, J.; Hagler, A. Chemoinformatics and Drug Discovery. *Molecules* **2002**, *7* (8), 566–600. DOI:10.3390/70800566.
104. Probst, D.; Reymond, J.-L. Visualization of Very Large High-Dimensional Data Sets as Minimum Spanning Trees. *Journal of Cheminformatics* **2020**, *12* (1). DOI:10.1186/s13321-020-0416-x.
105. Awale, M.; Visini, R.; Probst, D.; Arús-Pous, J.; Reymond, J.-L. Chemical Space: Big Data Challenge for Molecular Diversity. *CHIMIA* **2017**, *71* (10), 661. DOI:10.2533/chimia.2017.661.
106. Li, C.; Zheng, K. Methods, Progresses, and Opportunities of Materials Informatics. *InfoMat* **2023**, *5* (8). DOI:10.1002/inf2.12425.
107. Naveja, J. J.; Rico-Hidalgo, M. P.; Medina-Franco, J. L. Analysis of a Large Food Chemical Database: Chemical Space, Diversity, and Complexity. *F1000Research* **2018**, *7*, 993. DOI:10.12688/f1000research.15440.2.
108. Tropsha, A. Best Practices for QSAR Model Development, Validation, and Exploitation. *Molecular Informatics* **2010**, *29* (6–7), 476–488. DOI:10.1002/minf.201000061.
109. Cherkasov, A.; Muratov, E. N.; Fourches, D.; Varnek, A.; Baskin, I. I.; Cronin, M.; Dearden, J.; Gramatica, P.; Martin, Y. C.; Todeschini, R.; Consonni, V.; Kuz'min, V. E.; Cramer, R.; Benigni, R.; Yang, C.; Rathman, J.; Terfloth, L.; Gasteiger, J.; Richard, A.; Tropsha, A. QSAR Modeling: Where Have You Been? Where Are You Going To? *Journal of Medicinal Chemistry* **2014**, *57* (12), 4977–5010. DOI:10.1021/jm4004285.
110. Fourches, D.; Muratov, E.; Tropsha, A. Trust, but Verify: On the Importance of Chemical Structure Curation in Cheminformatics and QSAR Modeling Research. *Journal of Chemical Information and Modeling* **2010**, *50* (7), 1189–1204. DOI:10.1021/ci100176x.
111. Iqbal, J.; Vogt, M.; Bajorath, J. Activity Landscape Image Analysis Using Convolutional Neural Networks. *Journal of Cheminformatics* **2020**, *12* (1). DOI:10.1186/s13321-020-00436-5.

112. Thomas Scior; Philippe Bernard; Jose Luis Medina-Franco; Gerald. M. Maggiora. Large Compound Databases for Structure-Activity Relationships Studies in Drug Discovery. *Mini-Reviews in Medicinal Chemistry* **2007**, 7 (8), 851–860. DOI:10.2174/138955707781387858.
113. Iyer, P.; Stumpfe, D.; Vogt, M.; Bajorath, J.; Maggiora, G. M. Activity Landscapes, Information Theory, and Structure – Activity Relationships. *Molecular Informatics* **2013**, 32 (5–6), 421–430. DOI:10.1002/minf.201200120.
114. Maldonado, A. G.; Doucet, J. P.; Petitjean, M.; Fan, B.-T. Molecular Similarity and Diversity in Chemoinformatics: From Theory to Applications. *Molecular Diversity* **2006**, 10 (1), 39–79. DOI:10.1007/s11030-006-8697-1.
115. Sawada, R.; Kotera, M.; Yamanishi, Y. Benchmarking a Wide Range of Chemical Descriptors for Drug-target Interaction Prediction Using a Chemogenomic Approach. *Molecular Informatics* **2014**, 33 (11–12), 719–731. DOI:10.1002/minf.201400066.
116. Devillers, J.; Sartor, V.; Devillers, H. Predicting Mosquito Repellents for Clothing Application from Molecular Fingerprint-Based Artificial Neural Network SAR Models. *SAR and QSAR in Environmental Research* **2022**, 33 (9), 729–751. DOI:10.1080/1062936x.2022.2124014.
117. Bajusz, D.; Rácz, A.; Héberger, K. Why Is Tanimoto Index an Appropriate Choice for Fingerprint-Based Similarity Calculations? *Journal of Cheminformatics* **2015**, 7 (1). DOI:10.1186/s13321-015-0069-3.
118. Maggiora, G. M.; Shanmugasundaram, V. Molecular Similarity Measures. *Methods in Molecular Biology* **2010**, 39–100. DOI:10.1007/978-1-60761-839-3_2.
119. Willett, P. Similarity-Based Virtual Screening Using 2D Fingerprints. *Drug Discovery Today* **2006**, 11 (23–24), 1046–1053. DOI:10.1016/j.drudis.2006.10.005.
120. Pérez-Villanueva, J.; Méndez-Lucio, O.; Soria-Arteche, O.; Medina-Franco, J. L. Activity Cliffs and Activity Cliff Generators Based on Chemotype-Related Activity Landscapes. *Molecular Diversity* **2015**, 19 (4), 1021–1035. DOI:10.1007/s11030-015-9609-z.
121. Stumpfe, D.; Bajorath, J. Exploring Activity Cliffs in Medicinal Chemistry. *Journal of Medicinal Chemistry* **2012**, 55 (7), 2932–2942. DOI:10.1021/jm201706b.

122. Medina-Franco, J. L. Activity Cliffs: Facts or Artifacts? *Chemical Biology & Drug Design* **2013**, *81* (5), 553–556. DOI:10.1111/cbdd.12115.
123. Wawer, M.; Peltason, L.; Weskamp, N.; Teckentrup, A.; Bajorath, J. Structure–activity Relationship Anatomy by Network-like Similarity Graphs and Local Structure–activity Relationship Indices. *Journal of Medicinal Chemistry* **2008**, *51* (19), 6075–6084. DOI:10.1021/jm800867g.
124. PubMed. <https://pubmed.ncbi.nlm.nih.gov/>
125. Web of Science Core Collection. <https://clarivate.com/products/scientific-and-academic-research/research-discovery-and-workflow-solutions/web-of-science/web-of-science-core-collection/>
126. Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *Journal of Medicinal Chemistry* **1996**, *39* (15), 2887–2893. DOI:10.1021/jm9602928.
127. Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *Journal of Chemical Information and Modelling*. **2010**, *50* (5), 742–754. <https://doi.org/10.1021/ci100050t>.
128. Durant, J. L.; Leland, B. A.; Henry, D. R.; Nourse, J. G. Reoptimization of MDL Keys for Use in Drug Discovery. *Journal of Chemical Information and Modelling*. **2002**, *42* (6), 1273–1280. <https://doi.org/10.1021/ci010132r>.
129. The RDKit Book. https://www.rdkit.org/docs/RDKit_Book.html
130. Medina-Franco, J. L.; Martínez-Mayorga, K.; Bender, A.; Marín, R. M.; Giulianotti, M. A.; Pinilla, C.; Houghten, R. A. Characterization of Activity Landscapes Using 2D and 3D Similarity Methods: *Consensus Activity Cliffs*. *Journal of Chemical Information and Modeling* **2009**, *49* (2), 477–491. DOI:10.1021/ci800379q.
131. Medina-Franco, J. L. Scanning Structure–Activity Relationships with Structure–Activity Similarity and Related Maps: From *Consensus Activity Cliffs* to *Selectivity Switches*. *Journal of Chemical Information and Modeling* **2012**, *52* (10), 2485–2493. DOI:10.1021/ci300362x.
132. Waddell, J.; Medina-Franco, J. L. Bioactivity Landscape Modeling: Chemoinformatic Characterization of Structure–Activity Relationships of

Compounds Tested across Multiple Targets. *Bioorganic & Medicinal Chemistry* **2012**, *20* (18), 5443–5452. DOI:10.1016/j.bmc.2011.11.051.

133. Medina-Franco, J. L.; Navarrete-Vázquez, G.; Méndez-Lucio, O. Activity and Property Landscape Modeling Is at the Interface of Chemoinformatics and Medicinal Chemistry. *Future Medicinal Chemistry* **2015**, *7* (9), 1197–1211. DOI:10.4155/fmc.15.51.

134. Medina-Franco, J. L.; Martínez-Mayorga, K.; Bender, A.; Scior, T. Scaffold Diversity Analysis of Compound Data Sets Using an Entropy-based Measure. *QSAR; Combinatorial Science* **2009**, *28* (11–12), 1551–1560. DOI:10.1002/qsar.200960069.

135. Maaten, L. v. d., Hinton, G. Visualizing data using t-SNE. *Journal of Machine Learning Research*. **2008**;9(Nov):2579-2605.

136. Kris-Etherton, P. M.; Harris, W. S.; Appel, L. J. Omega-3 Fatty Acids and Cardiovascular Disease. *Arteriosclerosis, Thrombosis, and Vascular Biology* **2003**, *23* (2), 151–152. DOI:10.1161/01.atv.0000057393.97337.ae.

137. Simopoulos, A. P. The Importance of the Omega-6/Omega-3 Fatty Acid Ratio in Cardiovascular Disease and Other Chronic Diseases. *Experimental Biology and Medicine* **2008**, *233* (6), 674–688. DOI:10.3181/0711-mr-311.

138. Kris-Etherton, P.; Taylor, D. S.; Yu-Poth, S.; Huth, P.; Moriarty, K.; Fishell, V.; Hargrove, R. L.; Zhao, G.; Etherton, T. D. Polyunsaturated Fatty Acids in the Food Chain in the United States. *The American Journal of Clinical Nutrition* **2000**, *71* (1). DOI:10.1093/ajcn/71.1.179s.

139. Pérez-Jiménez, J.; Neveu, V.; Vos, F.; Scalbert, A. Identification of the 100 Richest Dietary Sources of Polyphenols: An Application of the Phenol-Explorer Database. *European Journal of Clinical Nutrition* **2010**, *64* (S3). DOI:10.1038/ejcn.2010.221.

140. Birt DF, Hendrich S, Wang W. Dietary agents in cancer prevention: flavonoids and isoflavonoids. *Pharmacology and Therapeutics*. **2001**, *90* (2-3): 157-177. DOI:10.1016/s0163-7258(01)00152-8.

141. Holick MF. Vitamin D deficiency. *The New England Journal of Medicine*. **2007**; *357* (3): 266-281. DOI:10.1056/NEJMra070553.

142. Mensink RP, Zock PL, Kester AD, Katan MB. Effects of dietary fatty acids and carbohydrates on the ratio of serum total to HDL cholesterol and on serum lipids and apolipoproteins: a meta-analysis of 60 controlled trials. *The American Journal of Clinical Nutrition*. **2003**; 77 (5): 1146-1155. DOI:10.1093/ajcn/77.5.1146.
143. Fernandez ML, Webb D. The LDL to HDL cholesterol ratio as a valuable tool to evaluate coronary heart disease risk. *Journal of the American College of Nutrition*. **2008**; 27 (1): 1-5. DOI:10.1080/07315724.2008.10719652.
144. Curis E, Crenn P, Cynober L. Citrulline and the gut. *Current Opinion in Clinical Nutrition and Metabolic Care*. **2007**; 10 (5): 620-626. DOI:10.1097/MCO.0b013e32829fb38d.
145. Figueroa A, Trivino JA, Sanchez-Gonzalez MA, Vicil F. Watermelon extract supplementation reduces ankle blood pressure and carotid augmentation index in obese adults with prehypertension or hypertension. *American Journal of Hypertension*. **2012**; 25 (6): 640-643. DOI:10.1038/ajh.2012.20.
146. Juneja LR, Chu D-C, Okubo T, Nagato Y, Yokogoshi H. L-Theanine—A unique amino acid of green tea and its relaxation effect in humans. *Trends in Food Science and Technology*. **1999**; 10 (6-7): 199-204. DOI:10.1016/s0924-2244(99)00044-8
147. Kakuda T. Neuroprotective effects of the green tea components theanine and catechins. *Biol Pharm Bull*. **2002**; 25 (12): 1513-1518. DOI:10.1248/bpb.25.1513.
148. Gattuso G, Barreca D, Gargiulli C, Leuzzi U, Caristi C. Flavonoid Composition of Citrus Juices. *Molecules*. **2007**; 12 (8): 1641-1673. DOI:10.3390/12081641.
149. Gao Z, Huang K, Yang X, Xu H. Free Radical Scavenging and Antioxidant Activities of Flavonoids Extracted from the Radix of *Scutellaria baicalensis* Georgi. *Biochimica et Biophysica Acta*. **1999**; 1472 (3): 643-650. DOI:10.1016/s0304-4165(99)00165-3.
150. Cao G, Prior RL. Anthocyanins are detected in human plasma after oral administration of an elderberry extract. *Clinical Chemistry*. **1999**; 45 (4): 574-576. DOI:10.1093/clinchem/45.4.574.

151. Wu X, Prior RL. Systematic identification and characterization of anthocyanins by HPLC-ESI-MS/MS in common foods in the United States: fruits and berries. *J Agric Food Chem.* **2005**; 53 (7): 2589-2599. DOI:10.1021/jf048068b.
152. Zhishen J, Mengcheng T, Jianming W. The determination of flavonoid contents in mulberry and their scavenging effects on superoxide radicals. *Food Chemistry.* **1999**; 64 (4): 555-559. DOI:10.1016/s0308-8146(98)00102-2.
153. Giusti MM, Wrolstad RE. Characterization and measurement of anthocyanins by UV-visible spectroscopy. *Current Protocols in Food Analytical Chemistry.* **2001**; F1.2.1-F1.2.13. DOI:10.1002/0471142913.faf0102s00.
154. Saldívar-González, F. I.; Naveja, J. J.; Palomino-Hernández, O.; Medina-Franco, J. L. Getting Smart in Drug Discovery: Chemoinformatics Approaches for Mining Structure–Multiple Activity Relationships. *RSC Advances* **2017**, 7 (2), 632–641. DOI:10.1039/c6ra26230a.

12. Anexos

Contenido		Disponible en:
<i>Epi Food Chemical Database</i>	Base de datos de sustancias químicas alimentarias reportadas con actividad epigenética en formato html.	https://DIFACQUIM.github.io/Epi_food_Chemical_Database/EpiFoodChemicalDatabase.html
	Base de datos de sustancias químicas alimentarias reportadas con actividad epigenética en formato csv.	https://github.com/DIFACQUIM/Epi_food_Chemical_Database/blob/main/EpiFoodChemicalDatabase.csv
Tabla S1. Enfermedades y genes	Tabla con el listado de enfermedades y genes obtenidos en la búsqueda bibliográfica en formato csv.	https://github.com/DIFACQUIM/Epi_food_Chemical_Database/blob/main/Tabl eS1_Diseasesandgenes.csv
Tabla S2. Artículos de investigación	Lista de los 436 artículos de investigación utilizados para crear la base de datos <i>Epi Food Chemical Database</i> .	https://github.com/DIFACQUIM/Epi_food_Chemical_Database/blob/main/Tabl eS2_Researchpapers.csv
Tabla S3. Descriptores moleculares	Lista de descriptores moleculares calculados y empleados en la generación del t-SNE.	Páginas 120-131
Mapa SAS con ECFP4	Mapa SAS interactivo de los compuestos en la <i>Epi Food Chemical Database</i> usando la huella digital molecular ECFP de radio 2.	https://DIFACQUIM.github.io/Epi_food_Chemical_Database/SAS_maps/SAS_map_ECFP4.html
Mapa SAS con ECFP6	Mapa SAS interactivo de los compuestos en la <i>Epi Food Chemical Database</i> usando la huella digital molecular ECFP de radio 3.	https://DIFACQUIM.github.io/Epi_food_Chemical_Database/SAS_maps/SAS_map_ECFP6.html

Mapa SAS con MACCS Keys	Mapa SAS interactivo de los compuestos en la <i>Epi Food Chemical Database</i> usando la huella digital molecular MACCS Keys.	https://DIFACQUIM.github.io/Epi_food_Chemical_Database/SAS_maps/SAS_map_MACCSKeys.html
Mapa SAS con RDKit	Mapa SAS interactivo de los compuestos en la <i>Epi Food Chemical Database</i> usando la huella digital molecular RDKit.	https://DIFACQUIM.github.io/Epi_food_Chemical_Database/SAS_maps/SAS_map_RDKit.html
Preprint del proyecto	Preprint con los resultados de esta investigación: https://doi.org/10.26434/chemrxiv-2023-drxvm-v2	Páginas 132-152

Tabla S3. Descriptores moleculares de *RDKit* empleados en la generación del t-SNE.

ID	Descriptor	Descripción
1	MaxAbsEStateIndex	Índice máximo absoluto de estado de electronegatividad. Representa la máxima electronegatividad absoluta entre todos los átomos en una molécula.
2	MaxEStateIndex	Índice máximo de estado de electronegatividad. Representa la máxima electronegatividad entre todos los átomos en una molécula.
3	MinAbsEStateIndex	Índice mínimo absoluto de estado de electronegatividad. Representa la mínima electronegatividad absoluta entre todos los átomos en una molécula.
4	MinEStateIndex	Índice mínimo de estado de electronegatividad. Representa la mínima electronegatividad entre todos los átomos en una molécula.
5	qed	Cantidad de energía deslocalizada. Indica la extensión de la deslocalización de la energía en la molécula.
6	MolWt	Peso molecular. Es la masa total de una molécula, sumando las masas atómicas de todos sus átomos.
7	HeavyAtomMolWt	Peso molecular de átomos pesados. Representa la masa total de una molécula excluyendo los átomos de hidrógeno.
8	ExactMolWt	Peso molecular exacto. Es la masa exacta de una molécula, considerando los pesos isotópicos exactos de los átomos.
9	NumValenceElectrons	Número de electrones de valencia. Indica la suma de los electrones en la capa de valencia de todos los átomos en la molécula.
10	NumRadicalElectrons	Número de electrones radicales. Representa el número total de electrones no apareados en la molécula.
11	MaxPartialCharge	Carga parcial máxima. Indica la máxima carga parcial en la molécula.
12	MinPartialCharge	Carga parcial mínima. Indica la mínima carga parcial en la molécula.
13	MaxAbsPartialCharge	Carga parcial máxima absoluta. Representa el valor absoluto de la máxima carga parcial en la molécula.
14	MinAbsPartialCharge	Carga parcial mínima absoluta. Representa el valor absoluto de la mínima carga parcial en la molécula.
15	FpDensityMorgan1	Densidad de fragmentos de Morgan 1. Descripción de la densidad de fragmentos moleculares basada en el primer nivel del algoritmo de Morgan.
16	FpDensityMorgan2	Densidad de fragmentos de Morgan 2. Descripción de la densidad de fragmentos moleculares basada en el segundo nivel del algoritmo de Morgan.

17	FpDensityMorgan3	Densidad de fragmentos de Morgan 3. Descripción de la densidad de fragmentos moleculares basada en el tercer nivel del algoritmo de Morgan.
18	BCUT2D_MWHI	Índice de fragmentación molecular bidimensional ponderado por masa, alto peso. Descriptor que cuantifica la estructura molecular basado en la matriz de adyacencia y la masa atómica.
19	BCUT2D_MWLOW	Índice de fragmentación molecular bidimensional ponderado por masa, bajo peso. Descriptor similar al anterior con peso bajo.
20	BCUT2D_CHGHI	Índice de fragmentación molecular bidimensional ponderado por carga, alto peso. Descriptor que cuantifica la estructura molecular basado en la matriz de adyacencia y la carga atómica.
21	BCUT2D_CHGLO	Índice de fragmentación molecular bidimensional ponderado por carga, bajo peso. Descriptor similar al anterior con peso bajo.
22	BCUT2D_LOGPHI	Índice de fragmentación molecular bidimensional logarítmico de la fase, alto peso. Descriptor logarítmico relacionado con la estructura molecular.
23	BCUT2D_LOGPLOW	Índice de fragmentación molecular bidimensional logarítmico de la fase, bajo peso. Descriptor logarítmico con peso bajo.
24	BCUT2D_MRHI	Índice de fragmentación molecular bidimensional de Ritz ponderado por masa, alto peso. Descriptor basado en la matriz de adyacencia y la masa atómica con peso alto.
25	BCUT2D_MRLOW	Índice de fragmentación molecular bidimensional de Ritz ponderado por masa, bajo peso. Descriptor similar al anterior con peso bajo.
26	AvgIpc	IPC promedio. Índice de Path Count promedio, que cuantifica la conectividad de una molécula.
27	BalabanJ	Índice de Balaban J. Descriptor topológico que cuantifica la insaturación y la ramificación en una molécula.
28	BertzCT	Número de enlaces de resonancia de Bertz. Cuantifica el número de enlaces en una molécula que pueden participar en resonancia.
29	Chi0	Índice de Chi0. Descriptor topológico relacionado con la conectividad molecular.
30	Chi0n	Índice de Chi0n. Descriptor topológico con peso en la conectividad de átomos de nitrógeno.
31	Chi0v	Índice de Chi0v. Descriptor topológico con peso en la conectividad de átomos de oxígeno y otros átomos pesados.
32	Chi1	Índice de Chi1. Descriptor topológico de primer orden.
33	Chi1n	Índice de Chi1n. Descriptor topológico de primer orden con peso en la conectividad de átomos de nitrógeno.

34	Chi1v	Índice de Chi1v. Descriptor topológico de primer orden con peso en la conectividad de átomos de oxígeno y otros átomos pesados.
35	Chi2n	Índice de Chi2n. Descriptor topológico de segundo orden con peso en la conectividad de átomos de nitrógeno.
36	Chi2v	Índice de Chi2v. Descriptor topológico de segundo orden con peso en la conectividad de átomos de oxígeno y otros átomos pesados.
37	Chi3n	Índice de Chi3n. Descriptor topológico de tercer orden con peso en la conectividad de átomos de nitrógeno.
38	Chi3v	Índice de Chi3v. Descriptor topológico de tercer orden con peso en la conectividad de átomos de oxígeno y otros átomos pesados.
39	Chi4n	Índice de Chi4n. Descriptor topológico de cuarto orden con peso en la conectividad de átomos de nitrógeno.
40	Chi4v	Índice de Chi4v. Descriptor topológico de cuarto orden con peso en la conectividad de átomos de oxígeno y otros átomos pesados.
41	HallKierAlpha	Alfa de Hall-Kier. Descriptores topológicos relacionados con la forma y el tamaño molecular.
42	lpc	IPC (Índice de Conectividad de Padua). Descriptor que cuantifica la conectividad molecular.
43	Kappa1	Kappa1. Descriptor relacionado con la forma molecular y la primera dispersión angular.
44	Kappa2	Kappa2. Descriptor relacionado con la forma molecular y la segunda dispersión angular.
45	Kappa3	Kappa3. Descriptor relacionado con la forma molecular y la tercera dispersión angular.
46	LabuteASA	Área de superficie de Labute. Cuantifica el área superficial molecular según el método de Labute.
47	PEOE_VSA1	Descriptor que cuantifica la superficie accesible por el método PEOE, enfocado en la región 1.
48	PEOE_VSA10	Descriptor que cuantifica la superficie accesible por el método PEOE, enfocado en la región 10.
49	PEOE_VSA11	Descriptor que cuantifica la superficie accesible por el método PEOE, enfocado en la región 11.
50	PEOE_VSA12	Descriptor que cuantifica la superficie accesible por el método PEOE, enfocado en la región 12.
51	PEOE_VSA13	Descriptor que cuantifica la superficie accesible por el método PEOE, enfocado en la región 13.

52	PEOE_VSA14	Descriptor que cuantifica la superficie accesible por el método PEOE, enfocado en la región 14.
53	PEOE_VSA2	Descriptor que cuantifica la superficie accesible por el método PEOE, enfocado en la región 2.
54	PEOE_VSA3	Descriptor que cuantifica la superficie accesible por el método PEOE, enfocado en la región 3.
55	PEOE_VSA4	Descriptor que cuantifica la superficie accesible por el método PEOE, enfocado en la región 4.
56	PEOE_VSA5	Descriptor que cuantifica la superficie accesible por el método PEOE, enfocado en la región 5.
57	PEOE_VSA6	Descriptor que cuantifica la superficie accesible por el método PEOE, enfocado en la región 6.
58	PEOE_VSA7	Descriptor que cuantifica la superficie accesible por el método PEOE, enfocado en la región 7.
59	PEOE_VSA8	Descriptor que cuantifica la superficie accesible por el método PEOE, enfocado en la región 8.
60	PEOE_VSA9	Descriptor que cuantifica la superficie accesible por el método PEOE, enfocado en la región 9.
61	SMR_VSA1	Descriptor que cuantifica la superficie móvil rápida accesible por el método SMR, enfocado en la región 1.
62	SMR_VSA10	Descriptor que cuantifica la superficie móvil rápida accesible por el método SMR, enfocado en la región 10.
63	SMR_VSA2	Descriptor que cuantifica la superficie móvil rápida accesible por el método SMR, enfocado en la región 2.
64	SMR_VSA3	Descriptor que cuantifica la superficie móvil rápida accesible por el método SMR, enfocado en la región 3.
65	SMR_VSA4	Descriptor que cuantifica la superficie móvil rápida accesible por el método SMR, enfocado en la región 4.
66	SMR_VSA5	Descriptor que cuantifica la superficie móvil rápida accesible por el método SMR, enfocado en la región 5.
67	SMR_VSA6	Descriptor que cuantifica la superficie móvil rápida accesible por el método SMR, enfocado en la región 6.
68	SMR_VSA7	Descriptor que cuantifica la superficie móvil rápida accesible por el método SMR, enfocado en la región 7.
69	SMR_VSA8	Descriptor que cuantifica la superficie móvil rápida accesible por el método SMR, enfocado en la región 8.
70	SMR_VSA9	Descriptor que cuantifica la superficie móvil rápida accesible por el método SMR, enfocado en la región 9.

71	SlogP_VSA1	Descriptor que cuantifica la contribución de la lipofilidad a la superficie accesible, enfocado en la región 1.
72	SlogP_VSA10	Descriptor que cuantifica la contribución de la lipofilidad a la superficie accesible, enfocado en la región 10.
73	SlogP_VSA11	Descriptor que cuantifica la contribución de la lipofilidad a la superficie accesible, enfocado en la región 11.
74	SlogP_VSA12	Descriptor que cuantifica la contribución de la lipofilidad a la superficie accesible, enfocado en la región 12.
75	SlogP_VSA2	Descriptor que cuantifica la contribución de la lipofilidad a la superficie accesible, enfocado en la región 2.
76	SlogP_VSA3	Descriptor que cuantifica la contribución de la lipofilidad a la superficie accesible, enfocado en la región 3.
77	SlogP_VSA4	Descriptor que cuantifica la contribución de la lipofilidad a la superficie accesible, enfocado en la región 4.
78	SlogP_VSA5	Descriptor que cuantifica la contribución de la lipofilidad a la superficie accesible, enfocado en la región 5.
79	SlogP_VSA6	Descriptor que cuantifica la contribución de la lipofilidad a la superficie accesible, enfocado en la región 6.
80	SlogP_VSA7	Descriptor que cuantifica la contribución de la lipofilidad a la superficie accesible, enfocado en la región 7.
81	SlogP_VSA8	Descriptor que cuantifica la contribución de la lipofilidad a la superficie accesible, enfocado en la región 8.
82	SlogP_VSA9	Descriptor que cuantifica la contribución de la lipofilidad a la superficie accesible, enfocado en la región 9.
83	TPSA	Área de Superficie Polar Total. Cuantifica la superficie polar total de una molécula.
84	EState_VSA1	Descriptor que cuantifica la superficie accesible por el método EState, enfocado en la región 1.
85	EState_VSA10	Descriptor que cuantifica la superficie accesible por el método EState, enfocado en la región 10.
86	EState_VSA11	Descriptor que cuantifica la superficie accesible por el método EState, enfocado en la región 11.
87	EState_VSA2	Descriptor que cuantifica la superficie accesible por el método EState, enfocado en la región 2.
88	EState_VSA3	Descriptor que cuantifica la superficie accesible por el método EState, enfocado en la región 3.
89	EState_VSA4	Descriptor que cuantifica la superficie accesible por el método EState, enfocado en la región 4.

90	EState_VSA5	Descriptor que cuantifica la superficie accesible por el método EState, enfocado en la región 5.
91	EState_VSA6	Descriptor que cuantifica la superficie accesible por el método EState, enfocado en la región 6.
92	EState_VSA7	Descriptor que cuantifica la superficie accesible por el método EState, enfocado en la región 7.
93	EState_VSA8	Descriptor que cuantifica la superficie accesible por el método EState, enfocado en la región 8.
94	EState_VSA9	Descriptor que cuantifica la superficie accesible por el método EState, enfocado en la región 9.
95	VSA_EState1	Descriptor que cuantifica la contribución de la superficie accesible al método EState, enfocado en la región 1.
96	VSA_EState10	Descriptor que cuantifica la contribución de la superficie accesible al método EState, enfocado en la región 10.
97	VSA_EState2	Descriptor que cuantifica la contribución de la superficie accesible al método EState, enfocado en la región 2.
98	VSA_EState3	Descriptor que cuantifica la contribución de la superficie accesible al método EState, enfocado en la región 3.
99	VSA_EState4	Descriptor que cuantifica la contribución de la superficie accesible al método EState, enfocado en la región 4.
100	VSA_EState5	Descriptor que cuantifica la contribución de la superficie accesible al método EState, enfocado en la región 5.
101	VSA_EState6	Descriptor que cuantifica la contribución de la superficie accesible al método EState, enfocado en la región 6.
102	VSA_EState7	Descriptor que cuantifica la contribución de la superficie accesible al método EState, enfocado en la región 7.
103	VSA_EState8	Descriptor que cuantifica la contribución de la superficie accesible al método EState, enfocado en la región 8.
104	VSA_EState9	Descriptor que cuantifica la contribución de la superficie accesible al método EState, enfocado en la región 9.
105	FractionCSP3	Fracción de átomos de carbono con hibridación sp ³ en la molécula.
106	HeavyAtomCount	Número total de átomos pesados en la molécula.
107	NHOHCount	Número de grupos funcionales amida e hidroxilo en la molécula.
108	NOCCount	Número de grupos funcionales nitroso y óxido en la molécula.
109	NumAliphaticCarbocycles	Número de carbociclos alifáticos en la molécula.
110	NumAliphaticHeterocycles	Número de heterociclos alifáticos en la molécula.

111	NumAliphaticRings	Número total de anillos alifáticos en la molécula.
112	NumAromaticCarbocycles	Número de carbociclos aromáticos en la molécula.
113	NumAromaticHeterocycles	Número de heterociclos aromáticos en la molécula.
114	NumAromaticRings	Número total de anillos aromáticos en la molécula.
115	NumHAcceptors	Número de aceptores de enlace de hidrógeno en la molécula.
116	NumHDonors	Número de donantes de enlace de hidrógeno en la molécula.
117	NumHeteroatoms	Número total de átomos hetero en la molécula.
118	NumRotatableBonds	Número de enlaces rotativos en la molécula.
119	NumSaturatedCarbocycles	Número de carbociclos saturados en la molécula.
120	NumSaturatedHeterocycles	Número de heterociclos saturados en la molécula.
121	NumSaturatedRings	Número total de anillos saturados en la molécula.
122	RingCount	Número total de anillos en la molécula.
123	MolLogP	Logaritmo del coeficiente de partición octanol-agua de la molécula.
124	MolMR	Masa molar relativa de la molécula.
125	fr_Al_COO	Número de grupos funcionales alilo y carboxilo en la molécula.
126	fr_Al_OH	Número de grupos funcionales alilo e hidroxilo en la molécula.
127	fr_Al_OH_noTert	Número de grupos funcionales alilo e hidroxilo excluyendo los grupos terciarios en la molécula.
128	fr_ArN	Número de grupos funcionales amina aromática en la molécula.
129	fr_Ar_COO	Número de grupos funcionales ácido carboxílico aromático en la molécula.
130	fr_Ar_N	Número de grupos funcionales nitrógeno aromático en la molécula.
131	fr_Ar_NH	Número de grupos funcionales amina aromática en la molécula.
132	fr_Ar_OH	Número de grupos funcionales hidroxilo aromático en la molécula.
133	fr_COO	Número de grupos funcionales ácido carboxílico en la molécula.
134	fr_COO2	Número de grupos funcionales diácido carboxílico en la molécula.
135	fr_C_O	Número de enlaces dobles carbono-oxígeno en la molécula.

136	fr_C_O_noCOO	Número de enlaces dobles carbono-oxígeno excluyendo los grupos carboxilo en la molécula.
137	fr_C_S	Número de enlaces dobles carbono-azufre en la molécula.
138	fr_HOCCN	Número de grupos funcionales hidroxilo, carbonilo y ciano en la molécula.
139	fr_Imine	Número de grupos funcionales imina en la molécula.
140	fr_NH0	Número de grupos funcionales amina primaria en la molécula.
141	fr_NH1	Número de grupos funcionales amina secundaria en la molécula.
142	fr_NH2	Número de grupos funcionales amina terciaria en la molécula.
143	fr_N_O	Número de grupos funcionales nitrógeno-oxígeno en la molécula.
144	fr_Ndealkylation1	Número de grupos funcionales desalquilación de nitrógeno en la molécula.
145	fr_Ndealkylation2	Número de grupos funcionales desalquilación de nitrógeno en la molécula.
146	fr_Nhpyrrole	Número de grupos funcionales pirrol de nitrógeno en la molécula.
147	fr_SH	Número de grupos funcionales tiol en la molécula.
148	fr_aldehyde	Número de grupos funcionales aldehído en la molécula.
149	fr_alkyl_carbamate	Número de grupos funcionales carbamato alquílico en la molécula.
150	fr_alkyl_halide	Número de grupos funcionales haluro alquílico en la molécula.
151	fr_allylic_oxid	Número de grupos funcionales óxido alílico en la molécula.
152	fr_amide	Número de grupos funcionales amida en la molécula.
153	fr_amidine	Número de grupos funcionales amidina en la molécula.
154	fr_aniline	Número de grupos funcionales anilina en la molécula.
155	fr_aryl_methyl	Número de grupos funcionales metilo arílico en la molécula.
156	fr_azide	Número de grupos funcionales azida en la molécula.
157	fr_azo	Número de grupos funcionales azo en la molécula.
158	fr_barbitur	Número de grupos funcionales barbitúrico en la molécula.
159	fr_benzene	Número de grupos funcionales benceno en la molécula.
160	fr_benzodiazepine	Número de grupos funcionales benzodiazepina en la molécula.

161	fr_bicyclic	Número de grupos funcionales bicíclico en la molécula.
162	fr_diazo	Número de grupos funcionales diazo en la molécula.
163	fr_dihydropyridine	Número de grupos funcionales dihidropiridina en la molécula.
164	fr_epoxide	Número de grupos funcionales epóxido en la molécula.
165	fr_ester	Número de grupos funcionales éster en la molécula.
166	fr_ether	Número de grupos funcionales éter en la molécula.
167	fr_furan	Número de grupos funcionales furano en la molécula.
168	fr_guanido	Número de grupos funcionales guanido en la molécula.
169	fr_halogen	Número de átomos de halógeno en la molécula.
170	fr_hdrzine	Número de grupos funcionales hidrazina en la molécula.
171	fr_hdrzone	Número de grupos funcionales hidrazona en la molécula.
172	fr_imidazole	Número de grupos funcionales imidazol en la molécula.
173	fr_imide	Número de grupos funcionales imida en la molécula.
174	fr_isocyan	Número de grupos funcionales isocianato en la molécula.
175	fr_isothiocyan	Número de grupos funcionales isotiocianato en la molécula.
176	fr_ketone	Número de grupos funcionales cetona en la molécula.
177	fr_ketone_Topliss	Número de grupos funcionales cetona según la clasificación de Topliss en la molécula.
178	fr_lactam	Número de grupos funcionales lactama en la molécula.
179	fr_lactone	Número de grupos funcionales lactona en la molécula.
180	fr_methoxy	Número de grupos funcionales metoxi en la molécula.
181	fr_morpholine	Número de grupos funcionales morfolina en la molécula.
182	fr_nitrile	Número de grupos funcionales nitrilo en la molécula.
183	fr_nitro	Número de grupos funcionales nitro en la molécula.
184	fr_nitro_ arom	Número de grupos funcionales nitro en anillos aromáticos en la molécula.
185	fr_nitro_ arom_ nonortho	Número de grupos funcionales nitro en anillos aromáticos no orto en la molécula.
186	fr_nitroso	Número de grupos funcionales nitroso en la molécula.
187	fr_oxazole	Número de grupos funcionales oxazol en la molécula.

188	fr_oxime	Número de grupos funcionales oxima en la molécula.
189	fr_para_hydroxylation	Número de grupos funcionales hidroxilo en posición para en la molécula.
190	fr_phenol	Número de grupos funcionales fenol en la molécula.
191	fr_phenol_noOrthoHbond	Número de grupos funcionales fenol sin enlace de hidrógeno en posición orto en la molécula.
192	fr_phos_acid	Número de grupos funcionales ácido fosfónico en la molécula.
193	fr_phos_ester	Número de grupos funcionales éster fosfónico en la molécula.
194	fr_piperdine	Número de grupos funcionales piperidina en la molécula.
195	fr_piperzine	Número de grupos funcionales piperazina en la molécula.
196	fr_priamide	Número de grupos funcionales amida primaria en la molécula.
197	fr_prisulfonamd	Número de grupos funcionales sulfonamida primaria en la molécula.
198	fr_pyridine	Número de grupos funcionales piridina en la molécula.
199	fr_quatN	Número de átomos de nitrógeno cuaternarios en la molécula.
200	fr_sulfide	Número de grupos funcionales sulfuro en la molécula.
201	fr_sulfonamd	Número de grupos funcionales sulfonamida en la molécula.
202	fr_sulfone	Número de grupos funcionales sulfona en la molécula.
203	fr_term_acetylene	Número de grupos funcionales acetileno terminal en la molécula.
204	fr_tetrazole	Número de grupos funcionales tetrazol en la molécula.
205	fr_thiazole	Número de grupos funcionales tiazol en la molécula.
206	fr_thiocyan	Número de grupos funcionales tiocianato en la molécula.
207	fr_thiophene	Número de grupos funcionales tiofeno en la molécula.
208	fr_unbrch_alkane	Número de cadenas alquílicas no ramificadas en la molécula.
209	fr_urea	Número de grupos funcionales urea en la molécula.

Juárez-Mercado, K. E.; Avellaneda-Tamayo, J. F.; Villegas-Quintero, H.; Chávez-Hernández, A. L.; López-López, C. D.; Medina-Franco, J. L. *Food Chemicals and Epigenetic Targets: Assembling an Epi Food Chemical Database*. *ChemRxiv* **2023**. DOI: 10.26434/chemrxiv-2023-drxvm-v2

Food Chemicals and Epigenetic Targets: Assembling an Epi Food Chemical Database

K. Eurídice Juárez-Mercado,¹ Juan F. Avellaneda-Tamayo,¹ Hassan Villegas-Quintero,¹ Ana L. Chávez-Hernández,¹ Claudia Daniela López-López,² José L. Medina-Franco^{1,*}

¹ DIFACQUIM research group. Department of Pharmacy, School of Chemistry, Universidad Nacional Autónoma de México.

² PECEM, School of Medicine, Universidad Nacional Autónoma de México.

* Correspondence: medinajl@unam.mx; Tel. +52-55-5622-3899

Abstract: There is increasing awareness of epigenetics's importance in understanding disease etiologies and developing novel therapeutics. An increasing number of publications in the past few years reflect the renewed interest in epigenetic processes and their relationship with food chemicals. However, there needs to be a recent study that accounts for the most recent advances in the area associating the chemical structures of the food and natural product components with their biological activity. Here, we analyze the status of food chemicals and their intersection with natural products in epigenetic research. Using cheminformatics tools, we compared quantitatively chemical contents, structural diversity, and coverage in the chemical space of food chemicals with reported epigenetic activity. As part of this work, we built and curated a compound database of food and natural product chemicals annotated with structural information, epigenetic target activity profile, and main source of the food chemical or natural product, among other relevant features. The compounds are cross-linked with identifiers from other major public databases such as FooDB and the COLleCtion of Open Natural ProdUcTs, COCONUT. The compound database, the "Epi Food Chemical Database" is accessible at https://github.com/DIFACQUIM/Epi_food_Chemical_Database

Keywords: chemical space; databases; epigenetics; food chemicals; foodinformatics; natural products.

1. Introduction

The concept of epigenetics has changed since it was first introduced in the 1940s by Conrad Waddington to describe "the branch of biology which studies the causal interactions between genes and their products which bring the phenotype into being".¹ Nowadays, the meaning of epigenetics is accepted as the study of the heritable changes in the gene expression profile that do not entail a change in DNA sequence but

1

modifies on the accessibility of the code via DNA methylation, modifications of amino acids on the amino-terminal tail of histones and non-coding RNAs.¹⁻³ It has been proposed that these changes could be classified into three types: direct epigenetics, which occurs in the lifespan of a person; indirect epigenetics defines the ones that occur inside the womb due to events during gestation; and across indirect epigenetics, referred to those changes that affected the individual predecessors and somehow, maybe through changes in the gametes or intrauterine environment setting, are transmitted across generations.² The immense interest in the field led to many studies showing the link between epigenetic changes and certain diseases such as diabetes, heart failure, cancer, inflammatory bowel diseases, and neurodegenerative diseases.⁴⁻⁷

Certain enzymes have been described as having a key role in these epigenetic modifications: DNA methyltransferases (DNMTs), in charge of the covalent addition of a methyl group to the DNA leading to the repression of certain genes; histone acetyltransferases (HATs) with the function of the acetylation of histone proteins, allowing the chromatin structure to open and become more transcriptionally active,⁸ and histone deacetylases (HDACs), which regulate the deacetylation of histones, leading to a hypoacetylation towards heterochromatin and gene suppression.⁹ Thus, the search for molecules that could hit these targets began, and the term "epidrugs" was coined to describe chemical compounds that alter DNA and chromatin structure, promoting the disruption of transcriptional and post-transcriptional modifications by the inhibition of DNMTs and HDACs, mainly. As of 2022, several compounds have been approved by the Food and Drug Administration of the USA for clinical use, while other compounds are chemical probes. Examples of representative epidrugs and epidrug candidates are azacytidine (DNMT1 inhibitor), 5-aza-2'-deoxycytidine (DNMTs and HDACs inhibitor), procaine (DNMTs inhibitor), hydralazine (DNMTs inhibitor), vorinostat (HDACs inhibitor), romidepsin (HDACs inhibitor), panobinostat (HDACs inhibitor), and belinostat (HDACs inhibitor). Nanaomycin A is a promising probe molecule (DNMT3b inhibitor).¹⁰⁻¹⁴ The chemical structures are shown in Figure 1.

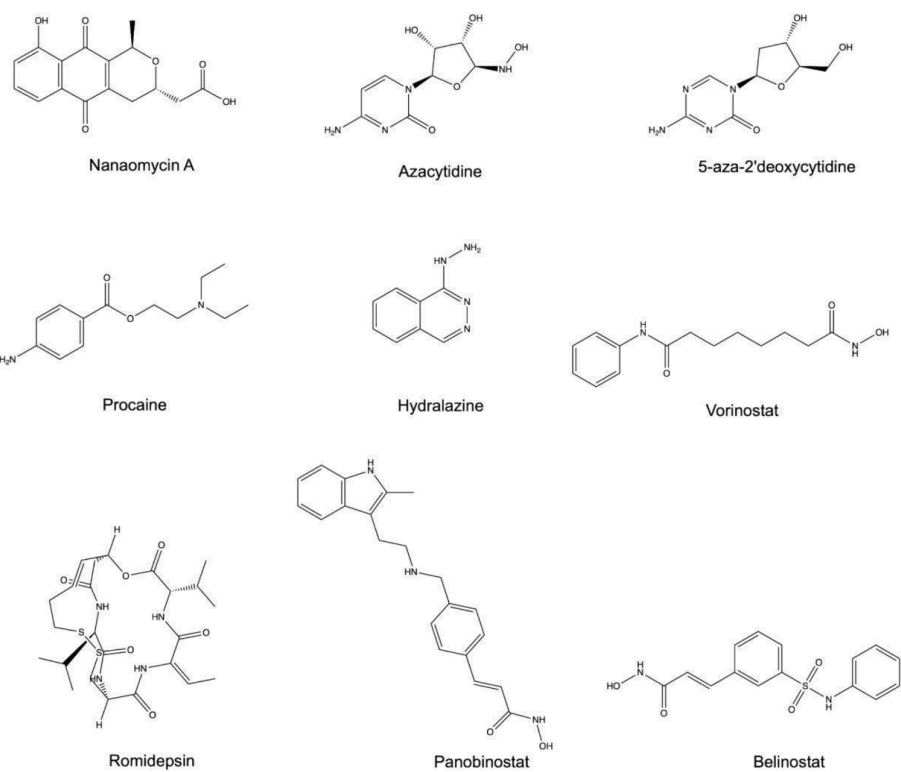


Figure 1. Chemical structures of representative epi drugs and epi drug candidates.

One of the most promising areas of this search is the field of nutriepigenomics, focused on the study of the interaction between food nutrients and genome through epigenetic mechanisms, modulating the overexpression or silencing of specific genes and metabolic responses.¹⁵⁻¹⁷ The interaction between nutrition, epigenetic targets, and the development of certain diseases such as type I and type II diabetes, inflammatory diseases, liver fibrosis, and cancer have been discussed in the last few years, leading to new alternatives to mitigate the damage or prevent such conditions.^{4,5,9,15,18-21}

Using chemoinformatics to analyze natural products²² and food chemical data sets is increasingly widespread. The term foodinformatics, coined in 2014,²³ captures chemical information's application to food science. Several studies focused on the contents and diversity of food chemicals have been published, yielding useful information to organize and mine chemical information associated with food chemicals, which, ultimately, is at the core of informatics applications in chemistry.²⁴ Similarly,

3

chemoinformatics has a growing interest in natural product research,²⁵ giving rise to the sub-field of natural products informatics.²² Notable examples of the applications of chemoinformatics to food chemistry and natural product research are the development of large compound databases such as FooDB²⁶ and the Collection of Open Natural Products (COCONUT).²⁷ Despite the increasing evidence of the effect of food and natural products chemicals on epigenetic targets, there needs to be a comprehensive survey of the effect of food molecules on different epigenetic targets rather than focusing on a specific disease or a specific epigenetic target family.

This study aimed to analyze the recent progress of research on food chemicals and food components acting with epigenetic targets, building a compound database that integrates the information on the chemical structure of food chemicals and other natural products with the epigenetic activity profile. The scientific papers and compound database were analyzed using chemoinformatics, data mining, and visualization approaches to identify the most frequent epigenetic targets and related therapeutic areas associated with food chemicals reported so far, the food chemicals and other natural products most studied, and their epigenetic activity profile. The chemical structure contents, diversity, and coverage in the chemical space of the compounds in the molecular database were evaluated using quantitative methods and data visualization techniques. Since a compound data set's chemical diversity and chemical space depend on the structure representation, we explored the chemical multiverse, e.g., chemical space generated with multiple structure representations.²⁸ As part of the analysis, we explored the relationships between the chemical structures and the epigenetic activity profile using the structure-property landscapes concept.²⁹

2. Methods

2.1. Literature search and analysis

We conducted a meta-analysis of the literature of research papers published between January 2017 and March 2023 in peer-reviewed journals with digital object identifier (DOI) numbers, documenting the research of food chemicals interacting with epigenetic targets with potential therapeutic applications or disease prevention. The literature search was done in PubMed³⁰ and Web of Science Core Collection³¹ databases using the following search terms: (("epigenetics" AND "food chemical(s)") OR ("*epigenetics*"

4

AND "natural products") OR ("epigenetics" AND "therapeutic application") OR ("epigenetics" AND "disease") OR ("epigenetics" AND "drug discovery") OR ("epigenetics" AND "drug development") OR "epigenetic targets" OR "epigenetic therapy" OR "epigenetic mechanisms" OR "epigenetic regulation" OR "epigenetic modifiers" OR "epidrugs" OR "nutritional epigenetics" OR "nutrigenetics"). As part of the analysis, the dietary compounds were determined in the abstract of the selected papers. Then, the most common therapeutic indications associated with these compounds were selected in the related papers. Additional analyses were performed after assembling and annotating a compound database described in Section 2.2.

2.2. Compound database of food and natural product chemicals annotated with epigenetic activity

Based on the literature search and analysis described in Section 2.1, a compound database herein termed "Epi Food Chemical Database" was assembled using Google Sheets. The chemical structures were represented using the linear notation Simplified Molecular-Input Line-Entry System (SMILES).³² The compound database was annotated with the following information: compound name; the International Chemical Identifier (InChI); the hashed version of InChI (InChIKey); main food source; if available, link of the compound to the FooDB or COCONUT databases (using the corresponding identifiers in those public databases); reference to the peer-reviewed paper using the DOI number; activity profile with the epigenetic targets for which the given compound has reported activity. To facilitate subsequent analysis and rapidly identify trends in the data, the activity profile was represented as a vector of "1"s and "0"s to indicate if the compound has or has not reported activity with a given epigenetic target, respectively.

2.3. Chemoinformatic analysis of the chemical database

The content and diversity of the chemical structures of the 187 compounds in the Epi Food Chemical Database were analyzed under three main types of analysis: a) scaffolds and chemical diversity using structural fingerprints and chemical scaffolds, b) distribution in chemical space, and c) descriptive structure-activity relationships based on the concept of activity, or more general, property landscapes.³³ Each of the three types of analysis is described below.

2.3.1. Chemical content and diversity analysis

The scaffold content analysis was based on the definition of Bemis and Murcko,³⁴ which considers a scaffold as the rings in a molecule and the connectors of them, the analysis was performed using an in-house code in Python with the modules MurckoScaffold of RDKit library. Also, the chemical structures of the compound database were analyzed using well-established protocols and broadly used to characterize or assess the chemical diversity, namely scaffold contents and structural diversity using four molecular fingerprints: Molecular ACCES System (MACCS) Keys (166-bits); Extended Connectivity Fingerprints (ECFP) radius 2 and 3; and RDKit fingerprints. The similarity analysis was calculated using the Jaccard-Tanimoto index.³⁵

2.3.2. Visualization of the chemical space

To visualize the chemical space of the compounds in the Epi Food Chemical Database, we generated a t-distributed stochastic neighbor embedding (t-SNE). This technique involves nonlinearly reducing dimensions by creating Gaussian probability distributions across high-dimensional space and then utilizing them to enhance a Student t-distribution within a lower-dimensional space through optimization. The lower dimensional space conserves pairwise similarities from the original higher dimensional space, resulting in clustering within the embedding space without a notable loss of the structural information.³⁶⁻³⁷

2.3.3. Structure-epigenetic activity profile

We computed all pairwise fingerprint-based and epigenetic activity profile similarities for the 187 Epi Food Chemical Database compounds. In both cases, we used the Jaccard-Tanimoto coefficient. The fingerprint-based similarity was calculated with four different fingerprints: ECFP4, ECFP6, MACCS Keys, and RDKit fingerprints.³⁸ In total, 17,590 pairwise comparisons were computed for each fingerprint (including self-comparisons) and 17,430 pairwise comparisons for each fingerprint (excluding self-comparisons). The structure vs. epigenetic activity profile similarity was plotted in a scatter plot reminiscent of the structure-activity similarity (SAS) maps.³⁹⁻⁴² Figure 2 shows a prototype plot of a SAS map where the epigenetic activity profile similarity is plotted on the Y-axis while the fingerprint-based structural similarity is plotted on the X-axis. An SAS map can be roughly divided into four regions as described in Figure 2; in Region I are

6

pairs of compounds with very similar activity profiles but very different structural similarities. In Region II are pairs of compounds with high structural similarity as similar activity profiles. Region III identifies pairs of compounds with low structural similarity and very different activity profiles. In Region IV, there are pairs of compounds with high structure similarity but very different epigenetic activity profiles.

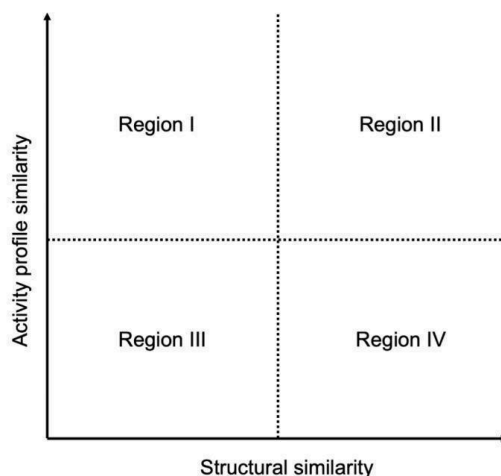


Figure 2. Prototype plot of a structure-activity similarity (SAS) map. Pairs of compounds in regions I and III have low structural similarity, while those in regions II and IV have high structural similarity. Pairs of compounds in regions I and II have a high similarity in their epigenetic activity profiles, although the chemical compounds in regions III and IV hold very different epigenetic activity profiles.

3. Results and discussion

3.1. Literature analysis

The literature search revealed that the number of peer-reviewed papers found in PubMed and Web of Science using the search terms described in the Methods section was 7,430 and 5,960, respectively; of which 4,484 were in both databases and 2,946 were unique for PubMed and 1,476 were unique for Web of Science. Table 1 summarizes the major twenty types of diseases associated with epigenetics, and chemical compounds present in the food or natural products identified in the current search are listed. Table S1 in the Supporting Information summarizes the complete list associated with the respective related genes and epigenetic targets.

Table 1. Top twenty types of diseases associated with food epigenetic compounds.

Associated diseases	Epigenetic target
Breast cancer	DNMT1, DNMT3a, DNMT3b, HDAC1, HDAC2, HDAC3, HDAC 4, HDAC6, SIRT1, SIRT 2, SIRT 3, SIRT 4, SIRT5, SIRT6, KDM1B, KDM2A, KDM3A, KDM4A, KDM4B, KDM5A, KDM6B, KDM7A, KDM8
Lung cancer	DNMT1, DNMT3a, HDAC4, HDAC5, HDAC6, HDAC8, HDAC9, SIRT2, KDM1A, KDM3B
Prostate cancer	DNMT1, HDAC, HDAC4, HDAC5, HDAC6, KDM1A, KDM2B, SIRT1
Colorectal cancer	DNMT, HDAC7, KDM6B
Bladder cancer	HDAC6, LSD1, KDM6A
Melanoma	HDAC2, HDAC5, KDM5A, KDM6A
Oral cancer	HDAC6, HDAC8, KDM1A
Hepatocellular carcinoma	DNMT3a, HDAC10, KDM1A, KDM2A
Alzheimer's	DNMT, HDAC3, SIRT1
Endometrial cancer	DNMT, DNMT1, HDAC 3, KDM4A
Non-small cell lung cancer	DNMT3a, HDAC1, HDAC2, KDM6B
Gastric cancer	DNMT1, DNMT3a, HDAC 2, KDM2A, KDM2B
Cervical cancer	DNMT1, HAT/Ep300, HAT2B/Ep300, KDM5C
Colon cancer	DNMT3b, HDAC 1, HDAC 3, HDAC 7, KDM4C, KDM5A, KDM6B
Diabetes mellitus type 2	DNMT, HDAC, SIRT1
Glioblastoma	LSD1, KDM1A
Obesity and metabolic diseases	DNMT, HDAC1, SIRT1
Esophageal carcinoma	DNMT, HAT2B/Ep300
Squamous cell carcinoma	HDAC, HDAC5
Atherosclerosis	DNMT, HDAC7, SIRT1

3.2. Compound database

A total of 436 papers out of 8,906 unique papers from both databases (PubMed and Web of Science) were used as the basis to build and curate the compound data set introduced in this work. The current data set version contains 187 unique compounds, of which 121 compounds have reported specific activity against

8

at least one of the targets, and 66 compounds have reported general activity against at least one target family. The Epi Food Chemical Database contains ten columns with general information plus forty-nine columns that encode the epigenetic activity profile of the compounds across forty-six epigenetic targets. The general information is comprised of structural data in three linear notations, namely SMILES, InChi, and InChi keys; chemical name, source of the compound, DOI of the peer-reviewed reference reporting the epigenetic activity, and links to FooDB and COCONUT databases through hyperlinks using the corresponding ID's on these two public databases.

The epigenetic activity profile is encoded as bit vectors of 0 and 1, indicating the absence or presence of reported activity for each of the 46 targets (see the Methods section 2.2 for details). The epigenetic targets are ordered and arranged into three main groups: writers, erasers, and readers as follows: 8 writers (DNMT1, DNMT3a, DNMT3b, HAT/Ep300, HAT2B/Ep300, HAT3B/p300, EZH2, PRMT1); 37 erasers (HDAC1, HDAC2, HDAC3, HDAC4, HDAC5, HDAC6, HDAC7, HDAC8, HDAC9, HDAC10, HDAC11, SIRT1, SIRT2, SIRT3, SIRT4, SIRT5, SIRT6, SIRT7, LSD1, KDM1A, KDM1B, KDM2A, KDM2B, KDM3A, KDM3B, KDM4A, KDM4B, KDM4C, KDM4D, KDM5A, KDM5B, KDM5C, KDM5D, KDM6A, KDM6B, KDM7A, KDM8) and 1 reader (BET/BRD4). The primary sources of the food chemicals in the Epi Food Chemical Database are meat, legumes, whole grains, grapes, poultry, acorn, acerola, strawberries, and nuts. Additional food sources are listed in the full Epi Food Chemical Database.

The 15 most frequent targets with reported activity of the compounds in the database are shown in Figure 3. We can see that the most frequent target is DNMT1 (63), followed by DNMT3B (35) and DNMT3A (34), HDAC6 (31), and HDAC1 (28).

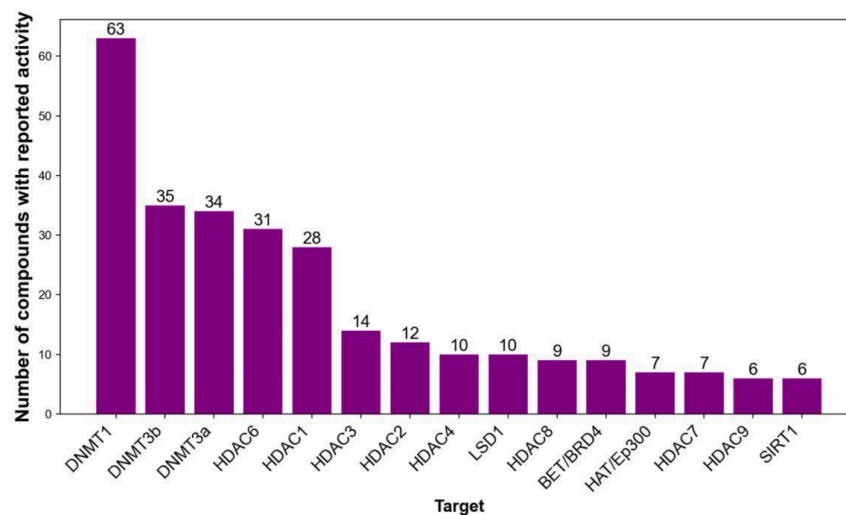


Figure 3. Histogram showing the 15 most frequent epigenetic targets.

There are 58 compounds with reported specific activity for only one target, being DNMT1 and HDAC6 the most frequent epigenetic targets with 18 compounds each, followed by LSD1 with eight compounds, BET/BRD4 with four compounds, and DNMT3a, DNMT3b, HAT/Ep300, KDM4a, with activity vs. two compounds in any case. Furthermore, three epigenetic targets are associated with specific reported activity vs. only one compound each: HDAC1 with phenethyl isothiocyanate (PEITC), SIRT1 with pterostilbene, and SIRT 5 with glutamate. The five compounds identified in the search with activity vs. the largest number of epigenetic targets were: biotin (27 targets), berberine (15 targets), alpha-ketoglutarate (13 targets), trichostatin (12 targets), and butein (11 targets). These and additional compounds are shown in Figure 4, including the chemical structure and the number of targets in parenthesis.

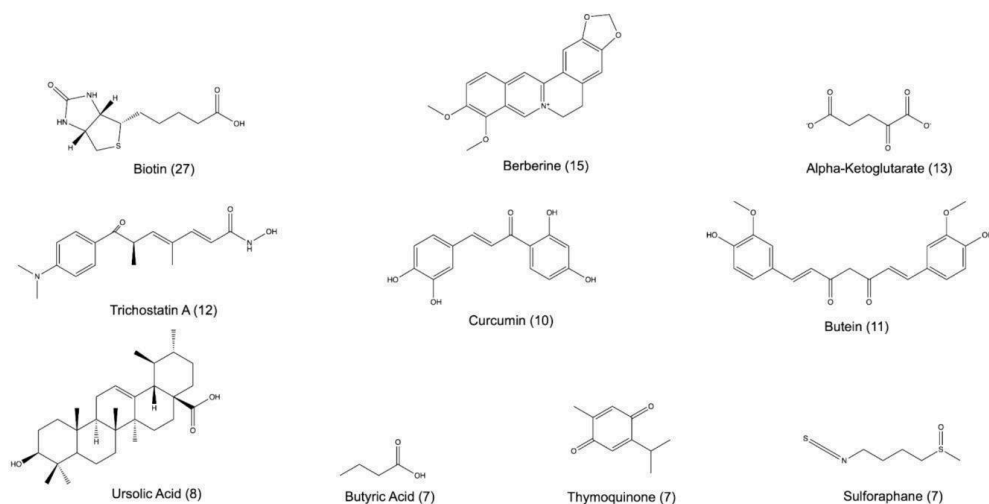


Figure 4. Top ten chemical compounds in Epi Food Chemical Database with epigenetic activity.

3.3. Chemoinformatic analysis

3.3.1. Diversity analysis

The total number of unique scaffolds for the 187 compounds was 90. Figure 5 shows the ten most frequent chemical scaffolds, along with the frequency and percent proportion, which represent 35.54% of the total distribution. The most frequent scaffolds were benzene (10.37%), followed by flavone (5.93%) and flavylum (2.96%). Other frequent scaffolds were indole (2.96%), pyridine (2.22%), hexane (2.22%), and isoflavone (1.48%).

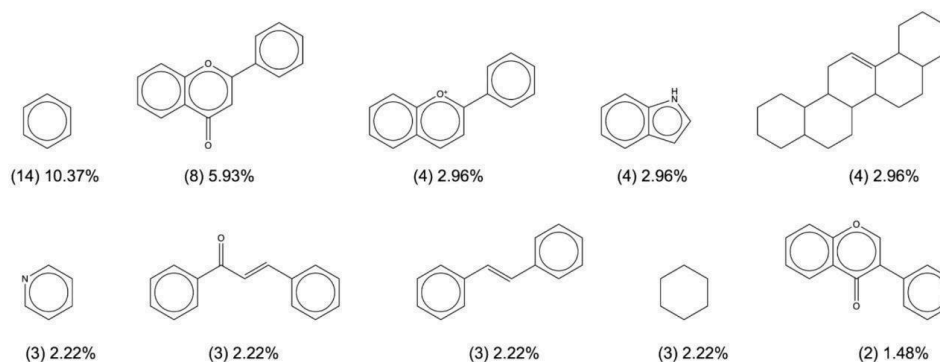


Figure 5. The ten most frequent scaffolds in the Epi Food Chemical Database.

Figure 6 shows the cyclic system recovery (CSR) curve for the scaffold diversity in the Epi Food Chemical Database. This curve illustrates the proportion of molecules within a dataset that belong to a specific fraction of scaffolds. In a dataset with high diversity, each molecule in the library would correspond to a different scaffold, resulting in a diagonal with an area under the curve (AUC) of 0.5. As the range of scaffold diversity diminishes, the curve will deviate from the diagonal orientation. Otherwise, the nadir of diversity would show in a dataset wherein all compounds share the same chemical scaffold; in such an instance, the CSR curve would appear as a vertical line, accompanied by an AUC of 1.0.⁴³ The shape of the CSR curve in Figure 6 indicates a large scaffold diversity of the Epi Food Chemical Database, with an AUC of 0.75.

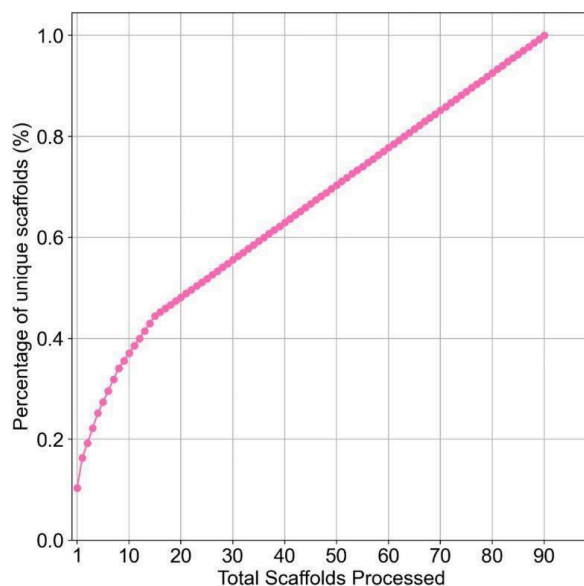


Figure 6. Cyclic system recovery curve of Bemis & Murcko scaffold diversity.

3.3.2. Visualization of the chemical space

The chemical space of the Epi Food Chemical Database was visualized in a graphical t-SNE representation. FoodDB (52,856 compounds) was included as a reference. The t-SNE was performed based

12

on the 209 descriptors in the module *MoleculeDescriptors* of RDKit. The descriptors include molecular weight, octanol/water coefficient (logP), number of hydrogen donor atoms (HBD), number of hydrogen acceptor atoms (HBA), topological polar surface area (TPSA), number of aromatic heterocycles, number of aromatic rings, number of heteroatoms, and the number of rotatable bonds. The visual representation of the chemical space indicates an overall diversity of the newly developed database, as compared to the space of the entire FooDB.

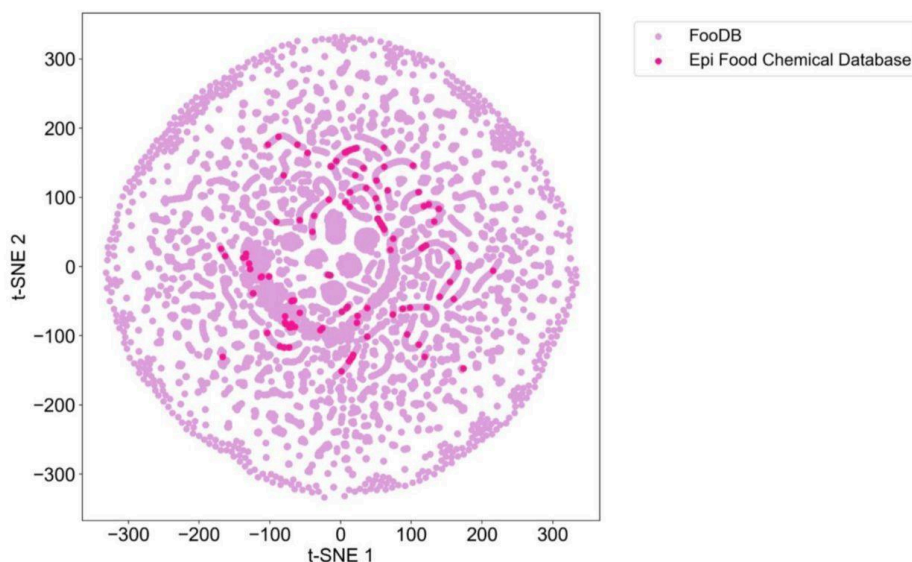


Figure 7. Visual representation of the Epi Food Chemical Database's chemical space (deep pink) compared to the chemical space covered by FooDB (lilac).

3.4. Structure-epigenetic target activity relationships

Figure 8 shows the SAS maps for the 187 chemical compounds in the Epi Food Chemical Database with the four different fingerprints: A) ECFP4, B) ECFP6, C) MACCS Keys, and D) RDKit fingerprint. The four interactive plots of the SAS maps are available in the Supporting information in html format.

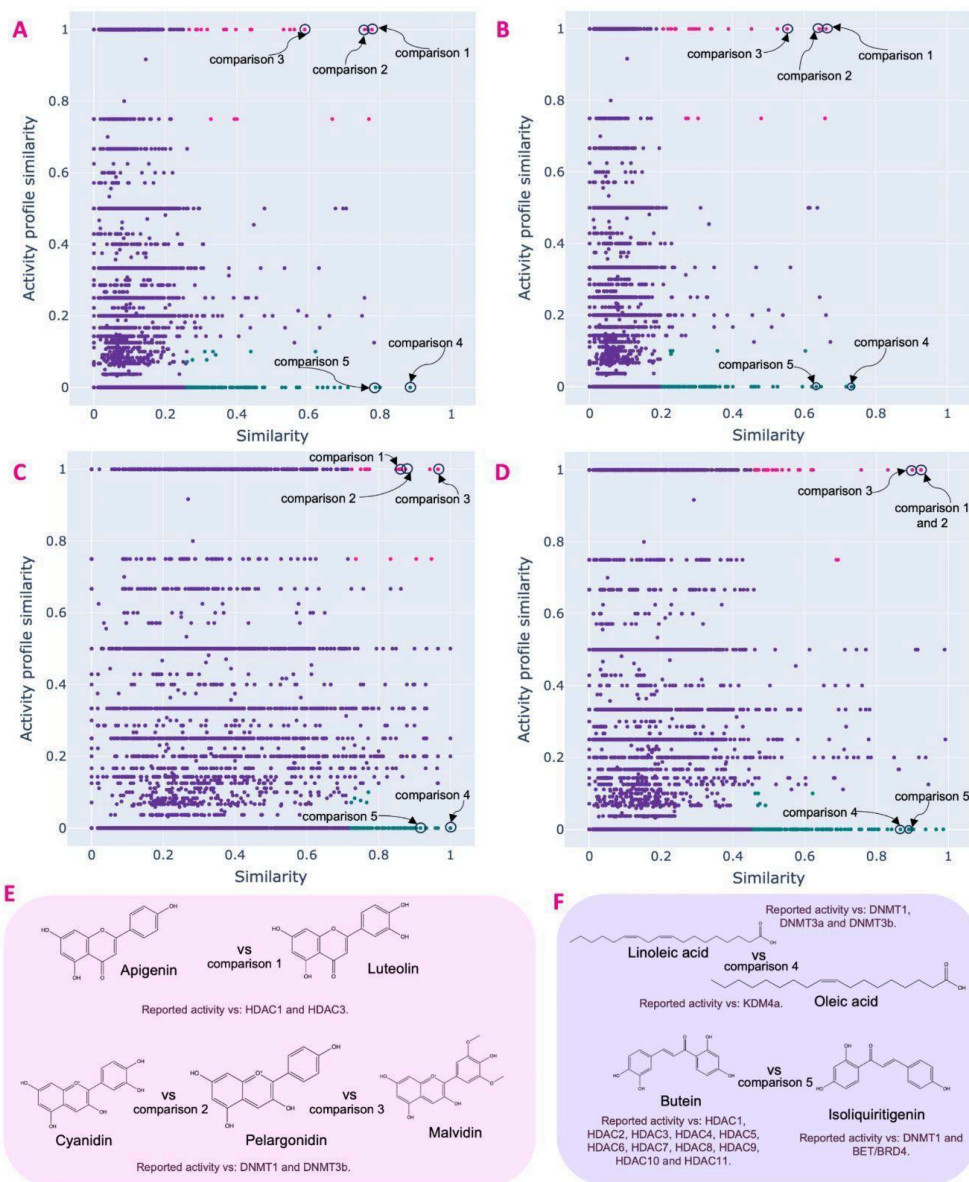


Figure 8. Structure activity-similarity (SAS) map of the Epi food Chemical Database. In pink are compound pairs in region II: similar structures and similar activity profiles; in green are compound pairs in IV region: similar chemical structures but very different epigenetic activity profiles (activity cliffs). Maps generated with A) ECFP4, B) ECFP6, C) MACCS Keys, D) RDKit fingerprint, E) examples of common compound pairs in region II (pink points) of all maps, F) examples of common compound pairs in region IV (green points) of all maps.

14

The pink data points represent the pair of chemical compounds in region II of the SAS maps, which correspond to compounds very similar in structure as in profile activity. An example of this compound pair that is common in the SAS maps of the four fingerprints is apigenin vs. luteolin (comparison 1 in Figure 8F). These compounds have reported activity vs. HDAC1 and HDAC3, and some of the principal sources of both compounds are parsley, celery, onions, and pepper. Other examples of compounds in this region of the SAS maps are the comparisons between cyanidin vs. malvidin vs. pelargonidin (comparisons 2 and 3, respectively, in Figure 8); in this case, the compounds have reported activity vs. DNMT1 and DNMT3b, and some of the principal common sources of the three compounds are blackberries, cherries, strawberries, and raspberries.

In contrast, the green data points in the SAS maps represent pairs of compounds in region IV, corresponding to compounds with similar activity profiles but very different chemical structures. Examples of these pairs of compounds present in region IV of all SAS maps for all the fingerprints are linoleic acid with reported activity vs. DNMT1, DNMT3a, and DNMT3b and oleic acid with reported activity vs. KDM4; their main sources are avocado, nuts, vegetable oils, and seeds. Another pair of compounds with very similar chemical structure but very different epigenetic activity profile (Figure 8F) is butein with reported activity vs. HDAC1, HDAC2, HDAC3, HDAC4, HDAC5, HDAC6, HDAC7, HDAC8, HDAC9, HDAC10 and HDAC 11 vs. isoliquiritigenin with reported activity vs. DNMT1 and BET/BRD4 in which their main sources are soybeans, peanuts, strawberries, and raspberries. It is important to emphasize that the pairwise epigenetic activity comparisons of the compounds of this work are based on the data published in the literature. For this reason, it is better to call them “pseudo activity cliffs” or pro-activity cliffs⁴⁴ instead of activity cliffs to the compounds in region IV. This is because some pairs of compounds may have very similar activity profiles but have not been fully tested yet. Examples of these compounds are apigenin and luteolin vs. chrysin. With current data reported in the literature, it is concluded that apigenin and luteolin are compounds that have similar structures with the same activity profile with reported activity vs. HDAC1 and HDAC3, but both compounds are pseudo activity cliffs vs. chrysin, which have activity reported vs. HDAC6. So it is probable that chrysin could have activity to HDAC1 and HDAC3 but also that apigenin and luteolin could also have activity vs. HDAC6.

4. Conclusions

Herein, we report constructing and curating the Epi Food Chemical Database, which contains 187 chemical compounds from dietary and natural products. The database includes structural information and the epigenetic activity profile obtained from the literature vs. 46 epigenetic targets. Breast cancer is by far the disease discussed in the literature with the largest number of epigenetic targets that are dysregulated. We used chemoinformatic tools to compare and analyze the structural content, diversity, and chemical space. Scaffold analysis revealed that the most frequent scaffolds were benzene, followed by flavone and flavylum. Diversity analysis and coverage in chemical space showed that the compounds in the Epi Food Chemical Database have an overall large diversity compared to compounds in FooDB. In addition, we identified two main groups of compounds; the first, with continuous structure-activity relationships, aka, fulfill the similarity principle: compounds with similar chemical structures have similar epigenetic activity profiles. The second group of compounds can be considered pseudo-activity cliffs (similar structures but very different epigenetic activity profiles). This work serves as a justification for further experimental testing of the compounds that form pseudo-activity cliffs. They may have similar activity to their analogous compounds. This work contributes to the further advancement of a systematic analysis of food and natural product chemicals with epigenetic activity using chemoinformatic approaches.

Supporting Information

The supporting information is available at <https://github.com/EuridiceJuarez/EpiFoodChemicalDatabase>. It contains the annotated compound database of food chemicals reported with epigenetic activity (Epi Food Chemical Database) in CSV format; Table S1 with the list of diseases/genes obtained in the literature search; Table S2 summarizes the list of 436 research papers used to build the Epi Food Chemical Database, and the interactive SAS maps plots of compounds in the Epi Food Chemical Database.

Acknowledgments

KE J-M, JF A-T, AL C-H thank *Consejo Nacional de Humanidades, Ciencias y Tecnologías* (CONAHCyT), Mexico, for the postgraduate scholarships 893849, 1270553, 847870. H V-Q thanks UNAM-HUAWEI for the scholarship under the project no. 7, “Desarrollo y aplicación de algoritmos de inteligencia artificial para el diseño de fármacos aplicables al tratamiento de diabetes mellitus y cáncer”.

16

Funding

We thank DGAPA, UNAM, Programa de Apoyo a Proyectos de Investigación e Innovación Tecnológica (PAPIIT), grants No. IN201321.

References

1. Dupont C, Armant DR, Brenner CA. Epigenetics: definition, mechanisms and clinical perspective. *Seminars in Reproductive Medicine*. 2009;27(5):351-357. DOI: 10.1055/s-0029-1237423
2. Lacal I, Ventura R. Epigenetic Inheritance: Concepts, Mechanisms and Perspectives. *Frontiers in Molecular Neuroscience*. 2018;11. doi:10.3389/fnmol.2018.00292
3. Li Y. Modern epigenetics methods in biological research. *Methods*. 2021;187:104-113. DOI: 10.1016/j.ymeth.2020.06.022
4. al Theyab A, Almutairi T, Al-Suwaidi AM, Bendriss G, McVeigh C, Chaari A. Epigenetic Effects of Gut Metabolites: Exploring the Path of Dietary Prevention of Type 1 Diabetes. *Frontiers in Nutrition*. 2020;7. DOI: 10.3389/fnut.2020.563605
5. Ramírez-Alarcón K, Victoriano M, Mardones L, et al. Phytochemicals as Potential Epidrugs in Type 2 Diabetes Mellitus. *Frontiers in Endocrinology*. 2021;12. DOI: 10.3389/fendo.2021.656978
6. Carlos-Reyes Á, López-González JS, Meneses-Flores M, et al. Dietary Compounds as Epigenetic Modulating Agents in Cancer. *Frontiers in Genetics*. 2019;10. DOI: 10.3389/fgene.2019.00079
7. Aleksandrova K, Romero-Mosquera B, Hernandez V. Diet, Gut Microbiome and Epigenetics: Emerging Links with Inflammatory Bowel Diseases and Prospects for Management and Prevention. *Nutrients*. 2017;9(9):962. DOI: 10.3390/nu9090962
8. Milagro FI, Martinez JA. Dietary and Metabolic Compounds Affecting Covalent Histone Modifications. In: *Handbook of Epigenetics*. Elsevier; 2017:307-322. DOI: 10.1016/B978-0-12-805388-1.00020-1
9. Evans LW, Stratton MS, Ferguson BS. Dietary natural products as epigenetic modifiers in aging-associated inflammation and disease. *Natural Product Reports*. 2020;37(5):653-676. DOI: 10.1039/C9NP00057G

10. Miranda Furtado CL, dos Santos Luciano MC, Silva Santos R da, Furtado GP, Moraes MO, Pessoa C. Epidrugs: targeting epigenetic marks in cancer treatment. *Epigenetics*. 2019;14(12):1164-1176. DOI: 10.1080/15592294.2019.1640546
11. Jin, Y.; Liu, T.; Luo, H.; Liu, Y.; Liu, D. Targeting Epigenetic Regulatory Enzymes for Cancer Therapeutics: Novel Small-Molecule Epidrug Development. *Frontiers in Oncology* 2022, 12. DOI: 10.3389/fonc.2022.848221
12. Arce, C.; Segura-Pacheco, B.; Perez-Cardenas, E.; Taja-Chayeb, L.; Candelaria, M.; Dueñas-Gonzalez, A. Hydralazine Target: From Blood Vessels to the Epigenome. *Journal of Translational Medicine* 2006, 4 (1). DOI: 10.1186/1479-5876-4-10
13. Yoon, S.; Eom, G. H. HDAC and HDAC Inhibitor: From Cancer to Cardiovascular Diseases. *Chonnam Medical Journal* 2016, 52 (1), 1. DOI: 10.4068/cmj.2016.52.1.1
14. Bennett, R. L.; Licht, J. D. Targeting Epigenetics in Cancer. *Annual Review of Pharmacology and Toxicology* 2018, 58 (1), 187–207. DOI: 10.1146/annurev-pharmtox-010716-105106
15. Meroni M, Longo M, Rustichelli A, Dongiovanni P. Nutrition and Genetics in NAFLD: The Perfect Binomium. *International Journal of Molecular Sciences*. 2020;21(8):2986. DOI: 10.3390/ijms21082986
16. González-Becerra K, Ramos-Lopez O, Barrón-Cabrera E, et al. Fatty acids, epigenetic mechanisms and chronic diseases: a systematic review. *Lipids in Health and Disease*. 2019;18(1):178. DOI: 10.1186/s12944-019-1120-6
17. Ali A, Hamzaid NH, Ismail NAS. The Interplay of Nutriepigenomics, Personalized Nutrition and Clinical Practice in Managing Food Allergy. *Life*. 2021;11(11):1275. DOI: 10.3390/life11111275
18. Li Y, Buckhaults P, Li S, Tollefsbol T. Temporal efficacy of a sulforaphane-based broccoli sprout diet in prevention of breast cancer through modulation of epigenetic mechanisms. *Cancer Prevention Research*. 2018;11(8):451-464. DOI: 10.1158/1940-6207
19. Regal P, Fente C, Cepeda A, Silva E. Food and omics: unraveling the role of food in breast cancer development. *Current Opinion in Food Science*. 2021;39:197-207. DOI: 10.1016/j.cofs.2021.03.008

20. Cuyàs E, Castillo D, Llorach-Parés L, et al. Computational de-orphanization of the olive oil biophenol oleacein: Discovery of new metabolic and epigenetic targets. *Food and Chemical Toxicology* 2019;131:110529. DOI: 10.1016/j.fct.2019.05.037
21. Taniguchi T, Tischkowitz M, Ameziane N, et al. Disruption of the Fanconi anemia–BRCA pathway in cisplatin-sensitive ovarian tumors. *Nature Medicine* 2003 9:5. 2003;9(5):568-574. DOI: 10.1038/nm852
22. Medina-Franco, J. L.; Saldívar-González, F. I. Cheminformatics to Characterize Pharmacologically Active Natural Products. *Biomolecules* 2020, 10 (11), 1566. DOI: 10.3390/biom10111566.
23. Martínez-Mayorga, K.; Medina-Franco, J. L. Foodinformatics. Applications of Chemical Information to Food Chemistry 2014. DOI: 10.1007/978-3-319-10226-9
24. López-López, E.; Bajorath, J.; Medina-Franco, J. L. Informatics for Chemistry, Biology, and Biomedical Sciences. *Journal of Chemical Information and Modeling* 2020, 61 (1), 26–35. DOI: 10.1021/acs.jcim.0c01301
25. Kirchmair, J. Molecular Informatics in Natural Products Research. *Molecular Informatics* 2020, 39 (11), 2000206. DOI: 10.1002/minf.202000206
26. <http://www.foodb.ca/>
27. Sorokina, M.; Merseburger, P.; Rajan, K.; Yirik, M. A.; Steinbeck, C. Coconut Online: Collection of Open Natural Products Database. *Journal of Cheminformatics* 2021, 13 (1). DOI: 10.1186/s13321-020-00478-9
28. Medina-Franco, J. L.; Chávez-Hernández, A. L.; López-López, E.; Saldívar-González, F. I. Chemical Multiverse: An Expanded View of Chemical Space. *Molecular Informatics* 2022, 41 (11). DOI: 10.1002/minf.202200116
29. Maggiora, G.; Medina-Franco, J. L.; Iqbal, J.; Vogt, M.; Bajorath, J. From Qualitative to Quantitative Analysis of Activity and Property Landscapes. *Journal of Chemical Information and Modeling* 2020, 60 (12), 5873–5880. DOI: 10.1021/acs.jcim.0c01249
30. PubMed. <https://pubmed.ncbi.nlm.nih.gov/>
31. Web of Science Core Collection. <https://clarivate.com/products/scientific-and-academic-research/research-discovery-and-workflow-solutions/web-of-science/web-of-science-core-collection/>

32. Weininger, D. Smiles, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *Journal of Chemical Information and Computer Sciences* 1988, 28 (1), 31–36. DOI: 10.1021/ci00057a005
33. Maggiora, G.; Medina-Franco, J. L.; Iqbal, J.; Vogt, M.; Bajorath, J. From Qualitative to Quantitative Analysis of Activity and Property Landscapes. *Journal of Chemical Information and Modeling* 2020, 60 (12), 5873–5880. DOI: 10.1021/acs.jcim.0c01249
34. Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *Journal of Medicinal Chemistry* 1996, 39 (15), 2887–2893. DOI: 10.1021/jm9602928
35. Chung, N. C.; Miasojedow, B.; Startek, M.; Gambin, A. Jaccard/Tanimoto Similarity Test and Estimation Methods for Biological Presence-Absence Data. *BMC Bioinformatics* 2019, 20 (S15). DOI: 10.1186/s12859-019-3118-5
36. Kang, B.; García García, D.; Lijffijt, J.; Santos-Rodríguez, R.; De Bie, T. Conditional T-SNE: More Informative T-SNE Embeddings. *Machine Learning* 2020, 110 (10), 2905–2940. DOI: 10.1007/s10994-020-05917-0
37. Zhou, Y.; Sharpee, T. O. Using Global T-SNE to Preserve Intercluster Data Structure. *Neural Computation* 2022, 34 (8), 1637–1651. DOI: 10.1162/neco_a_01504
38. Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *Journal of Chemical Information and Modeling* 2010, 50 (5), 742–754. DOI: 10.1021/ci100050t
39. Medina-Franco, J. L.; Martínez-Mayorga, K.; Bender, A.; Marín, R. M.; Giulianotti, M. A.; Pinilla, C.; Houghten, R. A. Characterization of Activity Landscapes Using 2D and 3D Similarity Methods: Consensus Activity Cliffs. *Journal of Chemical Information and Modeling* 2009, 49 (2), 477–491. DOI: 10.1021/ci800379q
40. Medina-Franco, J. L. Scanning Structure–Activity Relationships with Structure–Activity Similarity and Related Maps: From Consensus Activity Cliffs to Selectivity Switches. *Journal of Chemical Information and Modeling* 2012, 52 (10), 2485–2493. DOI: 10.1021/ci300362x
41. Waddell, J.; Medina-Franco, J. L. Bioactivity Landscape Modeling: Chemoinformatic Characterization of Structure–Activity Relationships of Compounds Tested across Multiple Targets. *Bioorganic & Medicinal Chemistry* 2012, 20 (18), 5443–5452. DOI: 10.1016/j.bmc.2011.11.051

20

42. Medina-Franco, J. L.; Navarrete-Vázquez, G.; Méndez-Lucio, O. Activity and Property Landscape Modeling Is at the Interface of Chemoinformatics and Medicinal Chemistry. *Future Medicinal Chemistry* 2015, 7 (9), 1197–1211. DOI: 10.4155/fmc.15.51
43. Lipkus, A. H.; Watkins, S. P.; Gengras, K.; McBride, M. J.; Wills, T. J. Recent Changes in the Scaffold Diversity of Organic Chemistry as Seen in the CAS Registry. *Journal of Organic Chemistry* 2019, 84 (21), 13948–13956. DOI: 10.1021/acs.joc.9b02111
44. Waddell, J.; Medina-Franco, J. L. Bioactivity Landscape Modeling: Chemoinformatic Characterization of Structure–Activity Relationships of Compounds Tested across Multiple Targets. *Bioorganic & Medicinal Chemistry* 2012, 20 (18), 5443–5452. DOI: 10.1016/j.bmc.2011.11.051