



UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE QUÍMICA

**ENFOQUES COMPUTACIONALES PARA LA IDENTIFICACIÓN
DE PRODUCTOS NATURALES COMO INHIBIDORES DE
DIANAS EPIGENÉTICAS**

T E S I S

**QUE PARA OBTENER EL TÍTULO DE
QUÍMICA FARMACÉUTICA BIÓLOGA**

PRESENTA:

BEATRIZ ANGÉLICA PILÓN JIMÉNEZ

DIRECTOR DE TESIS

DR. JOSÉ LUIS MEDINA FRANCO

CIUDAD UNIVERSITARIA, Cd. Mx.

2019



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.

JURADO ASIGNADO:

PRESIDENTE: **Profesor: MARIA ISABEL AGUILAR LAURENTS**

VOCAL: **Profesor: FRANCISCO HERNÁNDEZ LUIS**

SECRETARIO: **Profesor: JOSE LUIS MEDINA FRANCO**

1er. SUPLENTE: **Profesor: JOSE FAUSTO RIVERO CRUZ**

2° SUPLENTE: **Profesor: MABEL CLARA FRAGOSO SERRANO**

SITIO DONDE SE DESARROLLÓ EL TEMA:

UNAM – FACULTAD DE QUÍMICA, EDIFICIO F, CUBÍCULO 309
DIFACQUIM (DISEÑO DE FÁRMACOS ASISTIDO POR COMPUTADORA)

ASESOR DEL TEMA:

DR. JOSÉ LUIS MEDINA FRANCO

SUSTENTANTE:

BEATRIZ ANGÉLICA PILÓN JIMÉNEZ

AGRADECIMIENTOS

Al Departamento de Superación Académica de la Facultad de Química de la Universidad Nacional Autónoma de México (UNAM) por el apoyo otorgado dentro del Subprograma 127 “Formación Básica en Investigación”.

A la Dirección General de Asuntos del Personal Académico (DGAPA) de la UNAM por el apoyo económico brindado dentro del Programa de Apoyo a Proyectos para la Innovación y Mejoramiento de la Enseñanza (PAPIME) con clave PE200118.

Al Consejo Nacional de Ciencia y Tecnología (CONACyT) por el apoyo económico brindado dentro del Programa de Ciencia Básica con clave 282785.

Los resultados de este proyecto se difundieron en las siguientes publicaciones y presentaciones en congresos. Se anexan los artículos al final del trabajo escrito (A3):

Artículos

Publicados

- Pilón-Jiménez BA, Saldívar-González FI, Díaz-Eufracio BI, Medina-Franco JL. BIOFACQUIM: A Mexican compound database of natural products. *BIOMOLECULES* **2019** 9:31
- Saldívar-González FI, Gómez-García A, Chávez-Ponce de León D.E, Sánchez-Cruz N, Ruiz-Rios J, Pilón-Jiménez BA, Medina-Franco J.L. 2018. Inhibitors of DNA methyltransferases from natural sources: A computational perspective. *FRONTIERS IN PHARMACOLOGY* **2018** 9:1144.

Enviados

Naveja JJ., Pilón-Jiménez BA, Bajorath J, Medina-Franco JL. A general approach for retrosynthetic molecular core analysis. *JOURNAL OF CHEMINFORMATICS* (**2019**), en revisión de pares.

Capítulo en libro

Saldívar-González FI, Pilón-Jiménez BA, and Medina-Franco JL. 2018. Chemical space of naturally occurring compounds *PHYSICAL SCIENCES REVIEW* **2019**, 4: 20180103.

Congreso

Pilón-Jiménez BA, Medina-Franco JL. BIOFACQUIM: A compound database of natural products from Mexico (CINF 6). 257th American Chemical Society Meeting, Orlando, FL, 31 Marzo - 4 Abril, 2019 (presentación oral).

ÍNDICE

ABREVIATURAS	7
ÍNDICE DE FIGURAS Y TABLAS	9
RESUMEN	12
1. INTRODUCCIÓN	13
2.1 Productos naturales	15
2.1.1 Importancia	15
2.1.2 Productos naturales como fármacos	16
2.2.1 Bases de datos de productos naturales	19
2.2.2 Quimioinformática de bases de datos	20
2.3 Dianas epigenéticas	21
2.1.3 Productos naturales y dianas epigenéticas	22
3. OBJETIVOS	26
3.1 Generales	26
3.2 Específicos	26
4. MATERIALES Y MÉTODOS	27
4.1 Base de datos BIOFACQUIM	27
4.1.1 Construcción	27
4.1.2 Curado	28
4.1.3 Análisis quimioinformático de contenido	28
4.1.3.1 Propiedades moleculares de relevancia farmacéutica	28
4.1.3.1.1 Conjuntos de datos de referencia	28
4.1.3.1.1 Distribución de propiedades	29
4.1.3.1.2 Espacio químico por análisis de componentes principales (PCA)	30
4.1.3.2 Núcleos base: contenido y diversidad	30
4.1.3.2.1 Núcleos base según Bemis y Murcko (<i>scaffolds</i>)	30
4.1.3.2.2 Núcleos putativos (<i>putative core</i>)	31
4.1.3.3 Huellas digitales: diversidad, espacio químico	32
4.1.3.3.1 Espacio químico por incrustación estocástica de vecino distribuido en t (t-SNE)	32
4.1.4 Diversidad “global”: análisis de diversidad de consenso	32
4.2 Perfil <i>in silico</i>	33

4.2.1 Búsqueda por similitud vs. dianas epigenéticas	33
4.2.1.1 Inhibidores de dianas epigenéticas de referencia	33
4.2.2.1.2 Espacio químico por PCA	34
4.2.2 Chemotargets (similitud vs. 4500 dianas biológicas)	35
4.2.2.1 Actividad conocida y predicha	36
5. RESULTADOS Y DISCUSIÓN	38
5.1 Base de datos BIOFACQUIM	38
5.1.1 Construcción y curado	38
5.1.2 Análisis quimioinformático de contenido	40
5.1.3.1 Distribución de propiedades	40
5.1.3.1.2 Espacio químico por análisis de PCA	43
5.1.3.2 Núcleos base: contenido y diversidad	44
5.1.3.2.1 Núcleos base según Bemis y Murcko (<i>scaffolds</i>)	44
5.1.3.2.2 Núcleos putativos (<i>putative core</i>)	45
5.1.3.3 Huellas digitales: diversidad y espacio químico	47
5.1.3.3.1 Espacio químico por t-SNE	47
5.1.4 Diversidad “global”: análisis de diversidad consenso	50
5.2 Perfil <i>in silico</i>	53
5.2.1 Búsqueda por similitud vs. dianas epigenéticas	53
5.2.1.1 Espacio químico por PCA	53
5.2.1.2 Análisis de similitud por huellas digitales moleculares	54
5.2.2 Chemotargets (similitud vs. 4500 dianas)	67
5.2.2.1 Actividad conocida	69
5.2.2.2 Actividad predicha	74
6. CONCLUSIONES	78
7. PERSPECTIVAS	80
8. REFERENCIAS	81
ANEXOS	87
A1. Ejemplo de contenido de la base de datos BIOFACQUIM.	87
A2. Resultados de Chemotargets para compuestos con actividad conocida.	88
A3. Resultados de Chemotargets para compuestos con actividad predicha.	99
A.4 Artículos publicados	111

ABREVIATURAS

AUC	Área bajo la curva
BrD(s)	Bromodominio (s)
CDplot	Gráfico de diversidad consenso
CP	Citocromos
DB	Base de datos
DNMT(s)	ADN metiltransferasa (s)
EC	Enzimas
ECFP	Huellas digitales moleculares de conectividad extendida / <i>Extended connectivity fingerprints</i>
ED	Distancia euclidiana
FDA	Agencia Regulatoria de Alimentos y Medicamentos de los Estados Unidos de América
FP (<i>Fingerprint</i>)	Huella digital molecular
GR	Receptores acoplados a proteínas G
HAT	Histona acetilasa
HBA	Átomos aceptores de puentes de Hidrógeno
HBD	Átomos donadores de puentes de Hidrógeno
HDAC(s)	Histona (s) desacetilasa (s)
HTS	Cribado de alto rendimiento
IC	Canales iónicos
IC₅₀	Concentración inhibitoria 50
KC	Cinasas
MACCS keys	Sistema de acceso molecular / <i>Molecular ACCess System</i>

MW	Peso molecular
NR	Receptores nucleares
OF	Otras familias
pAct_exp	Logaritmo de la actividad experimental
pAct_prd	Logaritmo de la actividad predicha
PC	Componente principal
PCA	Análisis de componentes principales
PCP	Propiedades fisicoquímicas
PUMA	Plataforma para el análisis molecular unificado
RB	Número de enlaces rotables
SAR	Relación estructura-actividad
<i>Scaffold</i>	Núcleo base
SlogP	Coefficiente de partición octanol / agua
SMILES	<i>Simplified Molecular Input Line Entry Specification</i>
T	Coefficiente de Tanimoto
t-SNE	Incrustación estocástica de vecino distribuido en t
TC	Transportadores
TPSA	Área de superficie polar topológica
UC	Sin clasificar

ÍNDICE DE FIGURAS Y TABLAS

FIGURAS

- Figura 1.** Estructuras químicas de un producto natural y derivados recientemente aprobados por la FDA para uso clínico.
- Figura 2.** Ejemplo de la tabla de análisis de huellas digitales respecto a la actividad conocida en blancos moleculares.
- Figura 3.** Distribución por año de los compuestos incluidos en la primera versión de BIOFACQUIM.
- Figura 4.** Compuestos seleccionados contenidos en BIOFACQUIM.
- Figura 5.** Diagramas de caja de las propiedades fisicoquímicas de BIOFACQUIM y los conjuntos de datos de referencia.
- Figura 6.** Representación visual 2D del espacio químico basado en las propiedades fisicoquímicas de ocho conjuntos de datos.
- Figura 7.** Núcleos base más frecuentes en BIOFACQUIM.
- Figura 8.** Ejemplos de solapamiento de núcleos base de BIOFACQUIM y NuBBEDB.
- Figura 9.** Representación visual 2D del espacio químico basado en huellas digitales topológicas.
- Figura 10.** Gráfico de diversidad consenso que compara la diversidad global de BIOFACQUIM con las bases de datos de referencia.
- Figura 11.** Representación visual 2D del espacio químico basado en las propiedades fisicoquímicas de once conjuntos de datos.

Figura 12. Porcentaje de compuestos con actividad, predicha y conocida, en al menos una diana biológica de cada familia.

Figura 13. Porcentaje de compuestos separados por tipo de actividad (predicha o conocida) en al menos una diana biológica de cada familia.

TABLAS

Tabla 1. Ejemplos de productos naturales con actividad en dianas epigenéticas.

Tabla 2. Bases de datos de referencia utilizadas para comparar BIOFACQUIM.

Tabla 3. Inhibidores epigenéticos de referencia utilizados para comparar BIOFACQUIM.

Tabla 4. Valores de los tres primeros componentes principales del espacio químico de las ocho bases de datos.

Tabla 5. Sobrelapamiento de núcleos putativos y núcleos base de Bemis y Murcko de BIOFACQUIM y NuBBEDB.

Tabla 6. Estadísticos de las representaciones moleculares utilizados para el *CDplot* de BIOFACQUIM y las bases de datos de referencia.

Tabla 7. Valores de los tres primeros componentes principales del espacio químico de los once conjuntos de datos.

Tabla 8. Compuestos seleccionados de BIOFACQUIM según el criterio de similitud con inhibidores de dianas epigenéticas.

Tabla 9. Resumen de dianas biológicas más comunes para presentar actividad por los compuestos en BIOFACQUIM.

Tabla 10. Número de compuestos analizados por Chemotargets.

- Tabla 11.** Datos estadísticos de cada familia de proteínas con actividad conocida.
- Tabla 12.** Número de moléculas de cada familia de proteínas con actividad conocida.
- Tabla 13.** Compuestos selectivos de BIOFACQUIM con actividad conocida.
- Tabla 14.** Compuestos promiscuos de BIOFACQUIM con actividad conocida.
- Tabla 15.** Datos estadísticos de cada familia de proteínas con actividad predicha.
- Tabla 16.** Número de moléculas de cada familia de proteínas con actividad predicha
- Tabla 17.** Compuestos selectivos de BIOFACQUIM con actividad predicha.
- Tabla 18.** Compuestos promiscuos de BIOFACQUIM con actividad predicha.

RESUMEN

Las bases de datos de productos naturales tienen un alto impacto en proyectos de descubrimiento de fármacos. El número de bases de datos públicas con moléculas de origen natural está aumentando. Aquí se muestran los avances para construir y analizar “BIOFACQUIM”, una base de datos constituida con productos naturales aislados en la Facultad de Química de la UNAM. Esta base de datos se construyó con base en una búsqueda bibliográfica en revistas indizadas y actualmente consta de 423 compuestos provenientes de plantas, hongos y propóleo mexicano. Se muestran los avances en la caracterización quimioinformática del contenido y la cobertura en el espacio químico de BIOFACQUIM comparada contra otras bases de datos de productos naturales y fármacos aprobados. También se presenta el perfil de propiedades fisicoquímicas, la diversidad estructural basada en diferentes representaciones moleculares, y el perfil multi-diana calculado *in silico*. Se discute el progreso en el perfil de BIOFACQUIM para identificar compuestos con potencial actividad como inhibidores de dianas epigenéticas con énfasis en ADN-metiltransferasas, histonas desacetilasas y bromodominios.

1. INTRODUCCIÓN

El término epigenética se refiere a cualquier cambio hereditario en la expresión de un gen que no está codificado en la secuencia del ADN por sí misma. Los cambios epigenéticos pueden encender o apagar genes y determinar qué proteínas son las que se transcribirán y por lo tanto se expresarán. En los humanos, hay tres mecanismos principales de modificación epigenética. El primero es la metilación de citosinas en las islas CpG del ADN que consiste en el enlace covalente de grupos metilo en el carbono 5' de la citosina cuyo proceso es catalizado por las enzimas llamadas ADN metiltransferasas (DNMTs, por sus siglas en inglés). El segundo mecanismo involucra modificaciones a las histonas, las cuales son proteínas que ayudan a la compactación del ADN. Para poder realizar el silenciamiento de genes, las histonas requieren estar desacetiladas para aumentar el grado de compactación. Este proceso se lleva a cabo por enzimas llamadas histonas desacetilasas (HDACs, por sus siglas en inglés). El tercer mecanismo es el reconocimiento de residuos de lisinas acetiladas en las colas de las histonas para el control de la transcripción. Esto se lleva a cabo por los bromodominios (BrD, por sus siglas en inglés) que están asociados a proteínas de cromatina y de transcripción. Si bien se requieren cambios epigenéticos para el desarrollo normal, también estos cambios pueden ser responsables de causar enfermedades a causa de una activación o silenciamiento anormal de los genes, tales como algún tipo de cáncer, inestabilidades cromosómicas y discapacidad intelectual.

A través de los años se han encontrado compuestos aislados de productos naturales que modulan diferentes dianas epigenéticas. Estos compuestos son del tipo polifenoles, flavonoides, antraquinonas y otras clases. Ejemplos de los primeros productos naturales descritos fue la curcumina, (-)-epigallocatequin-3-galato, mahanina, genisteína y quercetina.

Sin embargo, la mayoría de los productos naturales con actividad epigenética se han encontrado en forma fortuita. Una forma de facilitar la búsqueda sistemática de compuestos activos de origen natural como inhibidores de dianas epigenética o de cualquier otro blanco terapéutico de interés, es primero coleccionar, organizar, y sistematizar la información de productos naturales en bases de datos moleculares.

Dada la vasta biodiversidad de México, el trabajo que se hace en entidades de la UNAM para la recolección e identificación de productos naturales y, específicamente, la labor que realizan los investigadores en la Facultad de Química y sus equipos de trabajo para el aislamiento y purificación de compuestos encontrados en ellos surge la iniciativa de generar, analizar y hacer pública una base de datos molecular de estos compuestos. El análisis consiste en cuantificar la diversidad estructural y cobertura en el espacio químico de los compuestos. Otro elemento de la caracterización es estimar, *in silico*, el perfil multi-diana. Una de las primeras aplicaciones de la base de datos de productos naturales es identificar compuestos con potencial actividad moduladora de dianas epigenéticas con la guía de técnicas quimioinformáticas.

2. ANTECEDENTES

2.1 Productos naturales

2.1.1 Importancia

Los productos naturales son un recurso importante para el descubrimiento de fármacos [1]. Las plantas y hongos principalmente se encuentran en todos los ambientes habitables, y la mayoría se encuentran en la tierra. Frente a muchas tensiones y desafíos, aunado a que son seres sedentarios, estos productos naturales han desarrollado moléculas para evitar los ataques de los animales y las agresiones ambientales. Estas mismas moléculas confieren a las plantas su capacidad de desprender fragancias, colores y, de hecho, algunas de ellas están asociadas a su toxicidad.

Los vastos efectos farmacológicos de las plantas medicinales dependen de los componentes fitoquímicos. En general, los constituyentes fitoquímicos de las plantas se dividen en dos categorías según su papel en los procesos metabólicos básicos, a saber, los metabolitos primarios y secundarios. Los metabolitos primarios de las plantas están involucrados en las funciones básicas de la vida; por lo tanto, son más o menos similares en todas las células vivas. Por otro lado, los metabolitos secundarios son productos de vías subsidiarias como la vía del ácido shikímico [2]. Las propiedades terapéuticas de las plantas han sido reconocidas desde tiempos remotos y se ha reconocido que es esta área del metabolismo secundario la que proporciona la mayoría de los productos naturales farmacológicamente activos [3].

Muchas afecciones patológicas han sido tratadas con medicamentos derivados de plantas. Estos medicamentos se utilizan como brebajes o extractos de plantas concentrados sin aislamiento de compuestos activos. Sin embargo, el desarrollo de fármacos moderno requiere

el aislamiento y la purificación de uno o dos compuestos activos para ser evaluado durante la selección de candidatos a fármacos [4].

Los productos naturales de plantas y animales han sido la fuente principal de medicamentos, especialmente para los agentes anticancerígenos y antimicrobianos. En las últimas décadas se ha visto un aumento en el uso de plantas medicinales para la promoción de la salud y el tratamiento de enfermedades en muchos países, incluidos los países desarrollados.

A pesar del advenimiento de tecnologías eficientes como la química combinatoria y el cribado de alto rendimiento (HTS, por sus siglas en inglés), en los últimos años, los productos naturales han atraído nuevamente la atención de académicos e investigadores centrados en la química farmacéutica. Esto se debe a que los productos naturales han demostrado ser una fuente más prometedora de fármacos y estructuras nuevas que los compuestos obtenidos por la química combinatoria [5].

2.1.2 Productos naturales como fármacos

La Figura 1 muestra las estructuras químicas de algunos medicamentos provenientes de productos naturales aprobados en los últimos cuatro años para uso clínico por la agencia regulatoria de alimentos y medicamentos de Estados Unidos de América (FDA por sus siglas en inglés). Migalastat (Galafold[®]) es una molécula pequeña recientemente aprobada como medicamento por la FDA en Septiembre del 2018, que se aisló del hongo *Streptomyces lydicus PA-5726* [6]. Fue encontrado por la empresa Amicus Therapeutics y se usa para el tratamiento de la enfermedad de Fabry, restaurando la actividad de formas mutantes específicas de α -galactosidasa [7].

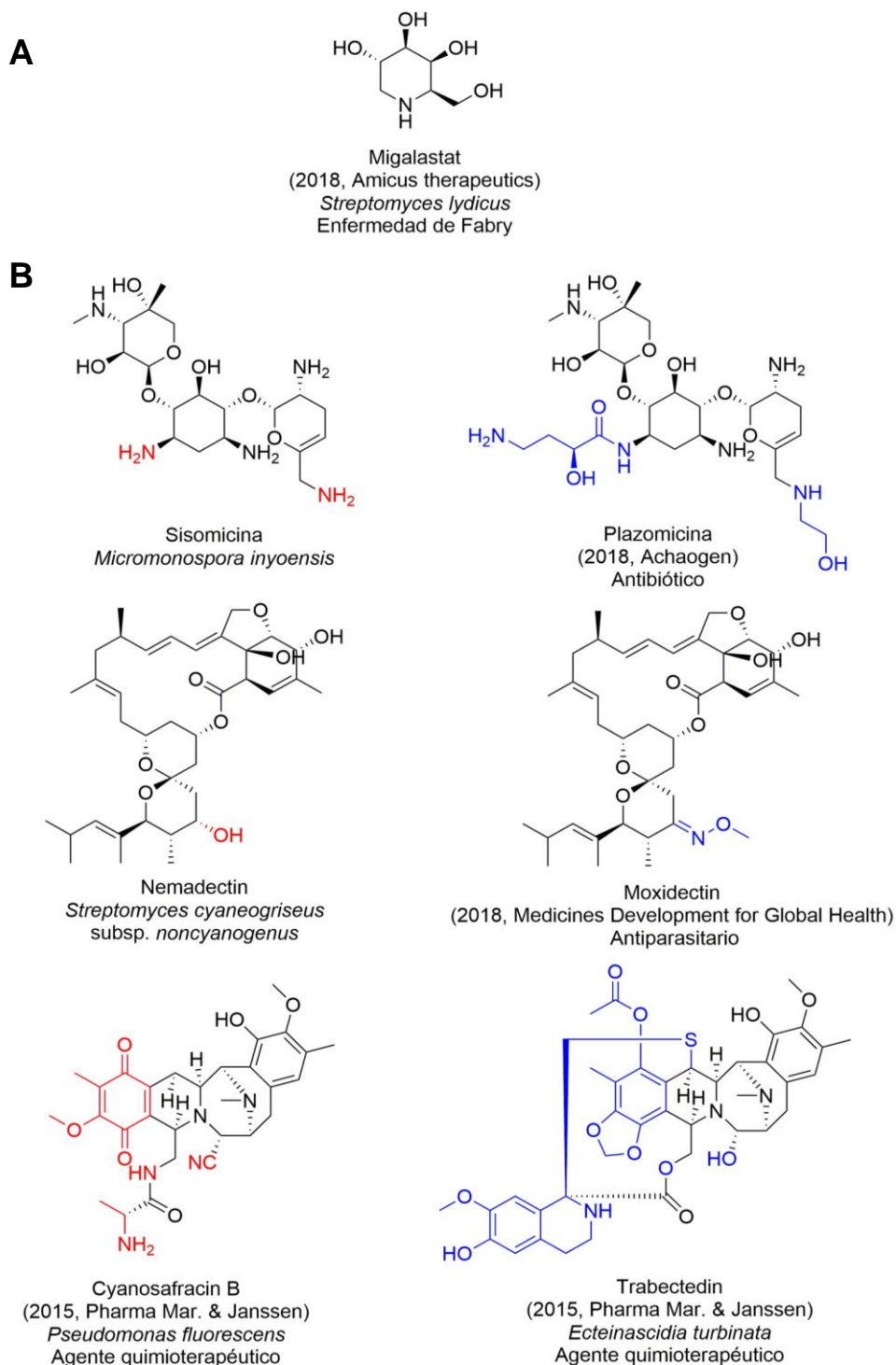


Figura 1. Estructuras químicas de A) un producto natural y B) derivados (derecha) de productos naturales (izquierda) recientemente aprobados para uso clínico por la FDA. Se indica la fuente, indicación terapéutica, año de aprobación y compañía farmacéutica. En azul se muestran las modificaciones que se hicieron a los compuestos aislados de productos naturales (izquierda).

Los procesos químicos de alto costo y de varios pasos para sintetizar migalastat llevaron al desarrollo de un proceso de fermentación sostenible y de bajo costo con *Streptomyces lydicus* PA-5726.

La plazomicina (Zemdri[®]) es un fármaco derivado de un producto natural. Fue desarrollado por Achaogen como un aminoglucósido de nueva generación que inhibe la síntesis de proteínas bacterianas. Se derivó sintéticamente de sisomicina, que se aísla de la fermentación aeróbica de *Micromonospora inyoensis*, añadiendo un sustituyente ácido hidroxiaminobutírico en la posición 1 y un sustituyente hidroxietilo en la posición 6' [8].

La moxidectina (Moxidectin[®]) es otro caso de un medicamento aprobado derivado de un compuesto previamente aislado de un producto natural. Es una lactona macrocíclica derivada de la nemadectina que se aisló a partir de *Streptomyces cyaneogriseus subsp. noncyanogenus* pero Moxidectin tiene la adición de un resto metoxime en C-23 [9]. Se utiliza en el tratamiento de la ceguera del río [10] o la oncocercosis por *Onchocerca volvulus*. Curiosamente, el primer análisis quimioinformático de productos naturales se centró en macrociclos y macrólidos [11]. En esas publicaciones, los autores destacaron la importancia de los macrociclos biológicamente activos en el descubrimiento de fármacos. La trabedectina (ET-743, Yondelis[®]) fue desarrollada por PharmaMar [12]. Fue descubierta y aislada del Tunicado de Manglar *Ecteinascidia turbinata*. Este fármaco es un agente alquilante que ha mostrado un potencial de amplio espectro significativo como un fármaco de segunda línea, de un agente solo o en combinación, particularmente en el tratamiento de liposarcomas y leiomiomas [13].

PharmaMar buscó una fuente adecuada debido a los bajos rendimientos de Trabedectin [12] obtenidos de las granjas acuícolas mediterráneas, más el impacto económico de los procesos de extracción y purificación. La compañía desarrolló varios procesos de síntesis

pero, incluso con estas mejoras, no era adecuada para la fabricación de ET-743 a escala industrial. Sin embargo, PharmaMar encontró un proceso semisintético a partir de la cianosafracina B, un antibiótico obtenido por fermentación de la bacteria *Pseudomonas fluorescens* que obtuvo buenos rendimientos y fue económicamente rentable.

2.2 Bases de datos

La importancia de las bases de datos de compuestos químicos en los proyectos de descubrimiento de fármacos está aumentando continuamente. De hecho, las bases de datos y los conjuntos de datos químicos son una parte central en las compañías farmacéuticas y otros centros de investigación académicos y gubernamentales [14].

2.2.1 Bases de datos de productos naturales

Las bases de datos de productos naturales tienen un alto impacto en proyectos de descubrimiento de fármacos, moléculas “sonda” y otras áreas de investigación [15]. Como se revisó en la sección 2.1.2, hay varios medicamentos recientemente aprobados para uso clínico que son productos naturales o son análogos sintéticos de compuestos de éxito identificados inicialmente a partir de fuentes naturales.

No es sorprendente que el descubrimiento de fármacos basados en productos naturales se combine con otras estrategias importantes de descubrimiento de fármacos, como el cribado de alto rendimiento y el cribado virtual [16]. Es por estas técnicas que son necesarias las bases de datos para su mejor rendimiento.

Varias bases de datos de compuestos de productos naturales han sido construidas, curadas y actualizadas a menudo por académicos y otros grupos de investigación sin fines de lucro.

Ejemplos notables son la Base de Datos Universal de Productos Naturales (UNPD, por sus siglas en inglés) [17] y la base de datos de Medicina Tradicional China que contiene compuestos utilizados en la medicina tradicional de Taiwán (TCM, por sus siglas en inglés) [18]. Es de destacar que UNPD ya no está disponible en línea, pero representa un esfuerzo de un grupo académico para reunir una base de datos de productos naturales de gran tamaño.

Existen otras bases de datos compuestas que recopilan productos naturales de diferentes países y áreas geográficas específicas. Ejemplos de estas son la Base de Datos Núcleo de Bioensayos, Ecofisiología y Biosíntesis (NuBBEDB, por sus siglas en inglés) que contiene productos naturales de Brasil [19], AfroDB con los productos naturales de Camerún [20], TIPdb de Taiwán [21] y recientemente fue lanzado al público VIETHERB: Una base de datos para las especies herbáceas vietnamitas [22].

A pesar de que México también tiene una gran biodiversidad, hay esfuerzos limitados para reunir una base de datos compuesta de productos naturales. Un ejemplo es UNIIQUIM (<https://uniquim.iquimica.unam.mx/>), una base de datos de productos naturales aislados en el Instituto de Química de la UNAM, donde el último artículo registrado de compuestos data del año 1996 y del producto natural de 2014 [23].

2.2.2 Quimioinformática de bases de datos

La quimioinformática es una disciplina que consiste en el uso de técnicas computacionales para comprender y ayudar a resolver los problemas de la química con especial énfasis en la manipulación de la información estructural química [24]. El término de “quimioinformática” se propuso en 1998 por Frank Brown [25]. La disciplina teórica y altamente relacionada con “bioinformática” y “quimiometría” se centra en almacenar, indexar, buscar, recuperar y

aplicar información sobre compuestos químicos [26, 27]. La quimioinformática ayuda a encontrar asociaciones en datos complejos y a explotar rápidamente la creciente información disponible para el descubrimiento de medicamentos y otras áreas de investigación.

En quimioinformática, la construcción de bases de datos es una práctica fundamental para realizar diversos estudios computacionales, como el diseño de bibliotecas químicas, la caracterización y comparación del espacio químico, el estudio de las relaciones estructura-actividad (SAR, por sus siglas en inglés) y los estudios de detección virtual [28], entre otros.

Las aplicaciones más específicas del espacio químico incluyen evaluar la diversidad de diferentes conjuntos de datos, explorar las relaciones entre colecciones de compuestos, y evaluar el potencial para cubrir otras regiones en el espacio químico aún por explorar. Del mismo modo, el concepto del espacio químico es útil para diseñar bibliotecas de compuestos novedosos y en la selección de compuestos de bibliotecas existentes para la evaluación computacional y/o experimental [28].

2.3 Dianas epigenéticas

La epigenética es el estudio de los cambios hereditarios en la expresión génica que no requieren, o generalmente no implican, cambios en la secuencia del ADN genómico. Anteriormente la epigenética se refería únicamente a los fenómenos del desarrollo, pero, más recientemente, ha llegado a significar una relación con la acción del gen, mientras que la herencia epigenética significa la modulación de la expresión del gen sin modificar la secuencia del ADN [29].

Las alteraciones en los procesos epigenéticos se han asociado con un gran número de enfermedades que incluyen cáncer, diabetes, trastornos neurodegenerativos, enfermedades mediadas por el sistema inmunitario [30, 31], entre otras.

Las ADN metiltransferasas (DNMTs, por sus siglas en inglés) conforman una familia de enzimas responsables de la metilación del ADN, que es la adición de un grupo metilo en la posición C5 de la citosina [32]. Dado que la metilación del ADN tiene un papel esencial para la diferenciación y el desarrollo celular, las alteraciones en la función de las DNMT se han asociado con el cáncer y otras enfermedades.

De manera similar a la metilación del ADN, las modificaciones de las histonas postraduccionales no afectan la secuencia de nucleótidos del ADN, pero pueden modificar su disponibilidad para la maquinaria transcripcional. Las histonas desacetilasas (HDAC, por sus siglas en inglés) eliminan los grupos acetilo de los residuos de lisina de la cola de histonas y, por lo tanto, funcionan como represores de la expresión génica [33].

Los bromodominios (BrDs) son moduladores de las interacciones proteína-proteína esenciales para los mecanismos epigenéticos del control transcripcional. Estos son pequeños dominios de proteínas (aproximadamente 110 aminoácidos) con secuencia conservada que reconocen la lisina acetilada de las colas N-terminales de las histonas. Es por ello que desempeñan funciones importantes en la regulación de la transcripción de genes dirigidos por histonas y la remodelación de la cromatina [34].

2.1.3 Productos naturales y dianas epigenéticas

Se ha visto que la variabilidad epigenética, en los sitios específicos reguladores de la transcripción, es susceptible a la modulación por cambios nutricionales [35]. Es por ello que

los productos naturales y los productos químicos alimenticios han sido una fuente importante de compuestos activos contra estas dianas. Estos productos naturales son del tipo polifenoles, flavonoides, antraquinonas y otras clases. En la Tabla 1 se muestran ejemplos de compuestos provenientes de productos naturales, su fuente y la actividad que presentan en estas dianas epigenéticas [36].

Cabe destacar que la mayoría de los productos naturales con actividad en dianas epigenéticas han sido descubiertos de forma fortuita. La gran cantidad de productos naturales en nuestro país es una fuente atractiva para encontrar compuestos con actividad epigenética y que eventualmente puedan convertirse en epi-fármacos o en moléculas sonda para el estudio de procesos epigenéticos. Uno de los primeros pasos para buscar en forma sistemática productos naturales mexicanos con actividad epigenética es primero organizar, curar y caracterizar la información en una base de datos molecular. Un segundo paso es seleccionar compuestos de la base de datos con potencial actividad biológica. Una forma de hacer la selección es mediante un filtrado computacional, específicamente, estimar o generar hipótesis *in silico* de compuestos con potencial actividad utilizando modelos validados.

Tabla 1. Ejemplos de productos naturales con actividad en dianas epigenéticas.

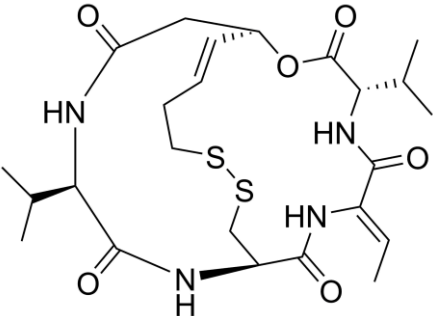
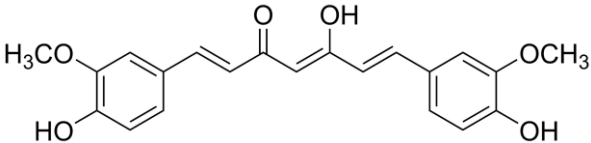
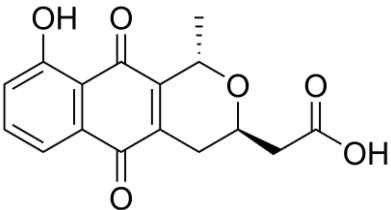
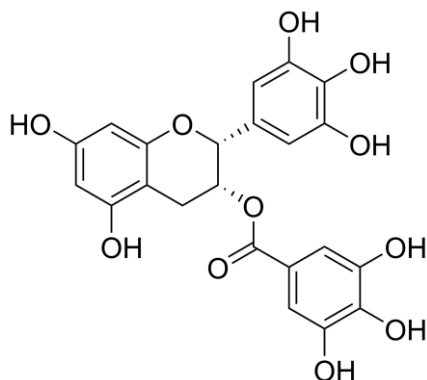
Compuesto	Fuente	Actividad
<p>Romidepsin</p> 	<p><i>Chromobacterium violaceum</i></p>	<p>Es uno de los cuatro inhibidores de HDAC, y el único proveniente de fuentes naturales aprobado por la FDA para el tratamiento del linfoma cutáneo y periférico de células T.</p>
<p>Curcumina</p> 	<p><i>Curcuma longa</i></p>	<p>Disminuye la expresión de la enzima DNMT1 en algunas líneas celulares tales como HCT116, HL-60, HEK293 y MV4-11.</p>
<p>Nanaomicin A</p> 	<p>Quinona aislada de <i>Streptomyces sp.</i></p>	<p>En las líneas celulares HCT116-colorrectal, A549-pulmón y HL60-leucemia inhibió el crecimiento celular y redujo los niveles de la metilación global del ADN.</p>

Tabla 1. Continuación

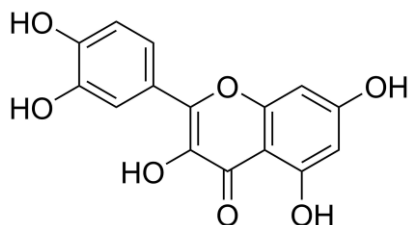
Epigallocatequin-3-galato



Es el polifenol principal que se encuentra en el té verde (*Camellia sinensis*).

Es un inhibidor de la enzima DNMT, y también inhibidor de HAT y activador del reclutamiento de HDAC3.

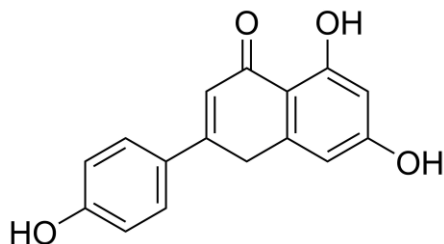
Quercetina



Compuesto tipo flavona extraído del té verde (*Camellia sinensis*).

Inhibidor de la metilación del promotor del gen P16^{INK4a}, el cual es un gen supresor de tumores.

Apigenina



Está presente en abundancia en frutas y verduras comunes como la toronja, el perejil, la manzanilla y trigo.

Mostró ser inhibidor de HDAC *in vitro*, especialmente para la clase I.

3. OBJETIVOS

3.1 Generales

- Introducir BIOFACQUIM como una de las primeras bases de datos compuesta de productos naturales aislados y caracterizados en México.
- Identificar compuestos encontrados en productos naturales como moduladores potenciales de dianas epigenéticas con énfasis en DNMTs, BrDs y HDACs, mediante el uso de métodos computacionales.

3.2 Específicos

- A. Iniciar la construcción de BIOFACQUIM: una base de datos de compuestos aislados de productos naturales en la Facultad de Química de la UNAM.
- B. Caracterizar la diversidad molecular y cobertura del espacio químico BIOFACQUIM respecto a moduladores de dianas epigenéticas y otros compuestos.
- C. Calcular el perfil epigenético y multi-diana potencial de compuestos en BIOFACQUIM mediante métodos quimioinformáticos.

4. MATERIALES Y MÉTODOS

4.1 Base de datos BIOFACQUIM

4.1.1 Construcción

La base de datos de productos naturales BIOFACQUIM, fue construida a partir de una búsqueda bibliográfica en revistas indizadas. Para la construcción de la primera versión de BIOFACQUIM se buscó en la base de datos *Scopus* (www.scopus.com) las palabras clave: "producto natural" y "Facultad de Química de la Universidad Nacional Autónoma de México (FQ, UNAM)". Esta búsqueda llevó a una lista de artículos científicos e investigadores que trabajan con productos naturales. Se seleccionaron ocho revistas en las que más habían contribuido hasta ahora:

- *Journal of Ethnopharmacology*,
- *Natural Products Research*,
- *Journal of Agricultural and Food Chemistry*,
- *Journal of Natural Products*,
- *Planta Medica*,
- *Phytochemistry*,
- *Natural Product Letters* y,
- *Molecules*

Como parte de la búsqueda, se utilizaron tres criterios de selección para la inclusión de los artículos de cada revista:

- a) Búsqueda por institución (FQ, UNAM).
- b) Búsqueda por año de publicación (2000-2018).
- c) El análisis detallado de los artículos, que contuviera el procedimiento para el aislamiento, purificación y caracterización de los compuestos a partir del producto natural.

Cabe destacar que esta es la primera versión de BIOFACQUIM. Las versiones futuras de la base de datos tendrán productos naturales de más años, más instituciones del país y revistas revisadas por pares para llegar al punto de tener una base de datos representativa de la biodiversidad mexicana.

4.1.2 Curado

La información contenida en la base de datos fue tratada en el programa *Molecular Operating Environment* (MOE) versión 2018 [37]. Con el módulo “lavado (*Wash*)” se realizó el curado de la base. El proceso se hizo para normalizar y recopilar la información más relevante de las moléculas. Para realizar el proceso de curación se siguió un protocolo establecido [38], que implicó:

- Eliminación de sales,
- El ajuste de los estados de protonación,
- La optimización de la geometría mediante la minimización de la energía,
- La eliminación de las moléculas duplicadas.

Se utilizaron los ajustes predeterminados del módulo "*Wash*".

4.1.3 Análisis quimioinformático de contenido

4.1.3.1 Propiedades moleculares de relevancia farmacéutica

4.1.3.1.1 Conjuntos de datos de referencia

Para caracterizar la diversidad de BIOFACQUIM y explorar su cobertura en el espacio químico, se utilizaron como referencia siete bases de datos de gran interés en el descubrimiento de fármacos. Las bases de datos de referencia utilizados en este trabajo se tomaron de comparaciones previas y análisis quimioinformáticos publicados de productos naturales [39]. A las estructuras de los

compuestos de referencia se les realizó el proceso de curado utilizando el mismo procedimiento descrito en la sección 4.1.2.

La Tabla 2 resume las bases de datos de referencia y el número de compuestos que compone a cada una. Cabe destacar que las colecciones de referencia incluyen cinco conjuntos de datos de productos naturales.

Tabla 2. Bases de datos de referencia utilizadas para comparar BIOFACQUIM.

Base de datos	Nombre	Tamaño^a
Fármacos aprobados por la FDA	Approved	1806
Metabolitos de cianobacterias	Cyanobacteria	473
Metabolitos de hongos	Fungi	206
Metabolitos de especies marinas	Marines	6253
Productos naturales comerciales	MEGx	4103
Semisintéticos	NATx	26318
Productos naturales de Brasil	NuBBE _{DB}	2214

^a Número único de compuestos después de la curación de los datos.

4.1.3.1.1 Distribución de propiedades

Una vez hecho el curado, la base de datos BIOFACQUIM se caracterizó calculando seis propiedades fisicoquímicas de interés terapéutico:

- a) Peso molecular (MW),
- b) Coeficiente de partición octanol/agua (SLogP),
- c) Área de superficie topológica superficial (TPSA),
- d) Número de enlaces rotables (RB),
- e) Número de átomos donadores de puente de hidrógeno (HBD) y,
- f) Número de átomos aceptores de puente de hidrógeno (HBA).

El análisis estadístico se realizó con el programa DataWarrior [40], calculando la media, mediana y desviación estándar de las propiedades calculadas. Basándose en estas estadísticas, BIOFACQUIM se comparó con otras bases de datos de productos naturales (NuBBEDB, Cyanobacteria, Fungi, Marines y MEGx), con fármacos aprobados y compuestos semisintéticos (NATx) (Tabla 2).

4.1.3.1.2 Espacio químico por análisis de componentes principales (PCA)

Para generar la representación visual del espacio químico de las propiedades de interés farmacéutico de BIOFACQUIM, se utilizó el método de visualización por análisis de componentes principales (PCA, por sus siglas en inglés). El PCA reduce la dimensión de los datos al proyectarlos geoméricamente en dimensiones más bajas llamadas componentes principales (PC). El primer PC se obtiene al minimizar la distancia total entre los datos y su proyección en el PC y para maximizar la varianza de los puntos proyectados [41]. Para ello se realizó un diagrama de trabajo en KNIME [42]. Se utilizó el nodo “*RDKit Descriptor Calculator*” para el cálculo de las propiedades fisicoquímicas. Se utilizó nodo “*Normalizer*”, que proporciona una transformación lineal de todos los valores, tomando el mínimo y el máximo de cada base de datos. Luego se aplicó el PCA para reducir la dimensionalidad de las seis propiedades fisicoquímicas calculadas y así comparar BIOFACQUIM con las bases de referencia (Tabla 2).

4.1.3.2 Núcleos base: contenido y diversidad

4.1.3.2.1 Núcleos base según Bemis y Murcko (*scaffolds*)

El análisis del contenido de núcleos base (*scaffolds*), permite identificar a los núcleos base más frecuentes en las bases de datos. Estos análisis también permiten identificar posibles núcleos base novedosos.

Los núcleos base más frecuentes de BIOFACQUIM se calcularon utilizando la definición de Bemis y Murcko [43], donde el núcleo base central se obtiene mediante la eliminación sistemática de las cadenas laterales de los compuestos. Para la obtención de estos, se utilizó la plataforma *Platform for Unified Molecular Analysis* (PUMA) [44]. Los núcleos base más frecuentes en BIOFACQUIM se compararon con los datos de la literatura.

4.1.3.2.2 Núcleos putativos (*putative core*)

Se realizó el análisis de superposición de núcleo putativo, en colaboración con el Dr. Jesús Naveja. En esta colaboración, un núcleo putativo se define como cualquier subestructura de una molécula que cumple con dos reglas:

- a) El tamaño del núcleo es una proporción significativa del tamaño de la molécula completa,
y
- b) Se puede sintetizar la subestructura desde la molécula original a través de una sucesión de las reglas de retrosíntesis [45].

Se implementó en *RDKit - Python* un algoritmo en el cual:

- a) Se lee la base de datos con la información de la estructura de los compuestos en formato SMILES y su identificador (ID).
- b) Se agrega una secuencia de comandos de "lavado" para eliminar las sales, retener el componente molecular más grande, generar SMILES canónicos y omitir la información estereoquímica de forma predeterminada.
- c) Cada molécula se fragmenta de forma independiente y solo se guardan los fragmentos que cumplen con la definición del núcleo putativo.

Se reporta la superposición de núcleos y se compara la información que arroja el núcleo base según Bemis & Murcko con el núcleo putativo aquí propuesto.

4.1.3.3 Huellas digitales: diversidad, espacio químico

4.1.3.3.1 Espacio químico por incrustación estocástica de vecino distribuido en t (t-SNE)

Para generar una representación visual del espacio químico de BIOFACQUIM por huellas digitales moleculares, se utilizó el método de incrustación de vecino estocástico distribuido en t (t-SNE por sus siglas en inglés).

t-SNE es una reducción de dimensión no lineal, donde las distribuciones de probabilidad gaussianas sobre el espacio de alta dimensión se construyen y se utilizan para optimizar una distribución *t de Student* en el espacio de baja dimensión. El espacio de baja dimensión mantiene la similitud de pares con el espacio de alta dimensión, lo que lleva a una agrupación en el espacio de incrustación sin perder información estructural significativa [46,47] [18, 19].

Para generar una visualización del espacio químico utilizando t-SNE se utilizaron subconjuntos de las bases de datos de referencia (Tabla 1), a denotar: 40% de los compuestos en las bases de metabolitos de especies marinas, MEGx y NuBBEDB (2501, 1641 y 886 compuestos, respectivamente). Para NATx y fármacos aprobados, se utilizaron 1000 moléculas para cada una. Para metabolitos de cianobacterias y de hongos y para BIOFACQUIM, se emplearon las bases de datos completas (473, 206 y 423 compuestos, respectivamente).

4.1.4 Diversidad “global”: análisis de diversidad de consenso

Dado que la diversidad química depende en gran medida de la representación de la estructura de los compuestos [48], se recomienda considerar representaciones múltiples para una evaluación global o completa. Con este fin, los gráficos de diversidad de consenso (CDplot) se han propuesto como simples gráficos bidimensionales que permiten comparar la diversidad de conjuntos de datos

compuestos utilizando cuatro conjuntos de representaciones de estructura [44]; típicamente, huellas digitales moleculares, núcleos base, propiedades moleculares, y número de compuestos. Los CDplot se han utilizado para comparar la diversidad de productos naturales y otros conjuntos de datos compuestos [49].

En un CDplot típico, la diversidad de *scaffolds* y huellas digitales moleculares se representa a lo largo de los ejes Y y X, respectivamente. La diversidad basada en propiedades moleculares de interés terapéutico se representa con una escala de color continua y el número de compuestos se mapea en el gráfico utilizando diferentes tamaños de puntos para cada base de datos.

Para generar el CDplot en este proyecto, se utilizó en el eje Y el área bajo la curva de recuperación del sistema cíclico [50]. Para el eje X, se empleó la mediana de la diversidad basada en huellas digitales moleculares calculada con MACCS keys (166-bits) y el coeficiente de Tanimoto. Ambas son métricas establecidas y representativas de la diversidad basada en núcleos base y huellas digitales moleculares, respectivamente. Se tomaron subconjuntos de las bases de datos de referencia (Tabla 1) considerando el tamaño de estas; para NATx, Marines, MEGx, NuBBEDB y fármacos aprobados se utilizaron 2000, 1500, 1000, 800 y 700 moléculas, respectivamente. Para metabolitos de cianobacterias, hongos y BIOFACQUIM, se emplearon todas las moléculas (473, 206 y 423 compuestos, respectivamente).

4.2 Perfil *in silico*

4.2.1 Búsqueda por similitud vs. dianas epigenéticas

4.2.1.1 Inhibidores de dianas epigenéticas de referencia

Para caracterizar el posible potencial de los compuestos en BIOFACQUIM como inhibidores de dianas epigenéticas, explorar su cobertura en el espacio químico y su similitud estructural con estos, se utilizó la base de datos D-DATABASE (D-DB) [51]. D-DB es una base de datos curada

de inhibidores epigenéticos, en la cual se han integrado varias bases de datos públicas que proporcionan información de la actividad o evidencia de interacción directa (por ejemplo, de estructuras co-cristalizadas).

La Tabla 3 resume las dianas epigenéticas exploradas y el número de inhibidores que tienen actividad contra cada una.

Tabla 3. Inhibidores epigenéticos de referencia utilizados para comparar BIOFACQUIM.

Diana epigenética	Tamaño (número de compuestos)
HDAC1	4885
HDAC2	149
HDAC3	336
BrD2	276
BrD3	26
BrD4	1200
BrD9	35
DNMT1	201
DNMT3A	36
DNMT3B	7
BIOFACQUIM	423

4.2.2.1.2 Espacio químico por PCA

Se generó la representación visual del espacio químico de las propiedades de interés farmacéutico de BIOFACQUIM con la de los inhibidores de dianas epigenéticas de referencia (Tabla 3). Para este fin se utilizó el método de visualización de PCA, mismo descrito en la sección 4.1.3.1.2.

4.2.2.1.3 Análisis de similitud por huellas digitales moleculares (*fingerprints*)

Se creó un diagrama de trabajo en KNIME [42] para analizar la similitud por dos huellas digitales, MACCS keys y *Extended Connectivity Fingerprint 4* (ECFP4). Para ello, tanto de los compuestos en BIOFACQUIM como del respectivo grupo de inhibidores de cada diana epigenética (Tabla 3), se obtuvieron los dos tipos de *fingerprints*.

Para cada *fingerprint* se utilizó el nodo “Similitud por *fingerprints* (*Fingerprint similarity*)”, una herramienta del grupo CDK. En este nodo se introducen los *fingerprints* de los compuestos de las dos bases de datos a comparar. Se obtuvieron tres aproximaciones:

- Matriz de similitud: Se obtuvo la matriz en donde se representa cada compuesto de BIOFACQUIM y la similitud que tiene contra cada inhibidor de cada diana epigenética, calculada con el coeficiente de Tanimoto. Los resultados se expresaron en mapas de calor.
- Promedio de similitud: Para cada compuesto en BIOFACQUIM se calculó el promedio de similitud del coeficiente de Tanimoto, contra todos los compuestos inhibidores de dianas epigenéticas.
- Máximo de similitud: Para cada compuesto en BIOFACQUIM se determinó el inhibidor de la diana epigenética que tuvo el máximo de similitud calculado con el coeficiente de Tanimoto.

Se realizó un consenso con los compuestos de BIOFACQUIM que tuvieron un alto nivel de similitud calculado con MACCS keys y ECFP4. Para ello, se aplicó un filtro para ambos casos (media y máximo), donde a los valores de los compuestos de BIOFACQUIM, se obtuvieron la media, desviación estándar y se seleccionaron a los compuestos que cumplieran con el requisito que fueran mayor o iguales a la suma de la media más dos veces la desviación estándar. Se identificaron a las estructuras que son potenciales para presentar actividad, en cuanto a similitud estructural con los inhibidores.

4.2.2 Chemotargets (similitud vs. 4500 dianas biológicas)

Para realizar el perfil *in silico* basado en ligando, se usó de la plataforma Chemotargets CLARITY, en colaboración con el Dr. Jordi Mestres (Hospital del Mar y Universitat Pompeu Fabra, Barcelona, España). Esta plataforma [52] es un *software* estadístico que utiliza seis enfoques

independientes para identificar dianas biológicas potenciales y modos de acción de moléculas pequeñas mediante la detección de más de 2,000 mecanismos de acción asociados con la actividad terapéutica y las condiciones de seguridad. Estos enfoques utilizan una variedad de información estructural tanto del compuesto y de las dianas biológicas para la predicción de la diana biológica, el mecanismo de acción y la afinidad de unión. Los métodos de predicción incluyen grupos de farmacóforos, relaciones cuantitativas estructura-actividad y técnicas de *machine learning* [53].

Chemotargets CLARITY da como salida el archivo con las interacciones conocidas y las predicciones respecto a las 4500 dianas biológicas programadas. Específicamente para los compuestos de la base de datos de BIOFACQUIM, proporcionó la información de:

- Identificador (ID) del compuesto de BIOFACQUIM
- Si el valor de actividad es conocido/predicho
- Valor de actividad (pAct_exp/pAct_prd)
- Valor de confianza
- Familia de la diana biológica
- Nombre de la proteína

Se realizó un primer filtro de la información y se retuvo la información de los compuestos con afinidad cuantitativa > 5 (10 μM) y valor de confianza > 0.30 (al menos 30 % de confianza). La primera aproximación realizada fue denotar el número de compuestos analizados, los blancos moleculares más comunes y el rango de actividades, tanto valores predictivos como ya conocidos.

4.2.2.1 Actividad conocida y predicha

Se calculó la actividad máxima, mínima, media y desviación estándar de los compuestos de cada familia de proteínas. Se establecieron criterios heurísticos para la selección de los compuestos con mayor actividad y reducir el número de moléculas por familia. Se realizaron gráficos divididos en

actividad conocida y predicha (aplicado el criterio según Tabla 11 y 15 respectivamente, de la sección 5.2.2), esto con el fin de identificar a los compuestos multi-diana (*multitarget*) y a los selectivos. Para cada familia de proteínas se hicieron dos gráficos (Anexos A1 y A2):

- Gráfico de número de moléculas por diana biológica.
- Gráfico de actividad por compuesto en cada diana biológica.

De la información obtenida se realizó un análisis de huellas digitales de los blancos moleculares tanto para la actividad conocida como predicha. En la Figura 2 se muestra un fragmento representativo de la tabla generada ilustrando la estrategia. En las filas se colocaron los compuestos de BIOFACQUIM y en las columnas los diferentes blancos moleculares (*targets*) analizados. Si el compuesto tiene una actividad conocida o predicha frente a un blanco molecular se le colocó un “1”, mientras que si no presenta en la predicción o no está reportada una actividad se le colocó un “0”. Se hizo la suma de los “1” y “0” y se dividió entre el número de actividades totales analizadas y se obtuvo un porcentaje para obtener un valor aproximado de la selectividad y promiscuidad de cada compuesto. Este análisis se hizo por separado tanto a las actividades conocidas como predichas. En la Figura 2 se muestra el ejemplo con las actividades conocidas.

ID	Targets						Núm. de targets	% Selectividad
	Cytochrome P450 1A2	Arginase	Adenosine receptor A1	[...]	Carbonic anhydrase 6	Protein disulfide-isomerase		
FQNP1	0	0	0		0	0	15	92.82
FQNP173	0	0	0		1	0	8	96.17
FQNP195	0	1	0		0	1	11	94.74
FQNP196	0	1	0		0	1	11	94.74
FQNP20	0	0	0		0	0	1	99.52
[...]								
FQNP261	0	0	0		1	0	9	95.69
FQNP263	0	0	0		1	0	12	94.26
FQNP274	0	1	0		1	0	14	93.3
FQNP317	1	0	1		0	0	41	80.38
FQNP418	0	0	0		0	0	2	99.04
FQNP85	0	0	0		0	0	1	99.52
Núm. total de targets analizados							209	

Figura 2. Ejemplo de la tabla de análisis de huellas digitales respecto a la actividad conocida en blancos moleculares.

5. RESULTADOS Y DISCUSIÓN

5.1 Base de datos BIOFACQUIM

5.1.1 Construcción y curado

Como se describe en la sección de Materiales y métodos, después de la primera revisión en *Scopus* con los nombres de los investigadores de la FQ, UNAM, se aplicaron tres filtros en las ocho revistas seleccionadas. Cada uno de los 92 artículos científicos seleccionados fue analizado individualmente para extraer la información de los productos naturales.

La base de datos contiene (Anexo A1):

- El número de identificación (ID)
- Nombre del compuesto, según lo reportado en el artículo
- Estructura del compuesto en forma de SMILES
- Referencia del artículo, que contiene:
 - Nombre de la revista
 - Año de publicación
 - Número de identificador de objeto digital (DOI, por sus siglas en inglés)
- Información del producto natural
 - Reino (*Plantae* o *Fungi*)
 - Género
 - Especie
 - Lugar de recolección
- Actividad biológica, si se reporta en la publicación
 - IC₅₀
 - Nombre del ensayo realizado
 - Bioactividad

Después de hacer el curado, se recopiló la información de 423 compuestos en total: 316 compuestos fueron aislados de 49 géneros diferentes de plantas, 98 de 19 géneros de hongos y 9 compuestos de propóleo mexicano (producto pegajoso de color oscuro recolectado por abejas de fuentes vegetales vivas).

La Figura 3 muestra la distribución de compuestos por año reportada desde el año 2000 como se encuentra en la primera versión de la base de datos. Los compuestos publicados en 2018 contenidos en la base de datos no se incluyen en la Figura 3.



Figura 3. Distribución de los compuestos incluidos, desde 2000 a 2017, en la primera versión de BIOFACQUIM.

La Figura 4 muestra a las estructuras químicas de los compuestos representativos de la primera versión de BIOFACQUIM, mismas que se mencionan en la sección 5.1.3.3.1.

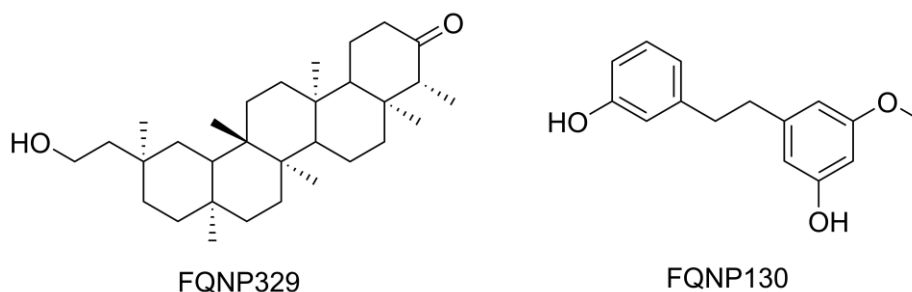


Figura 4. Compuestos seleccionados contenidos en BIOFACQUIM. Ver Figura 9b.

5.1.2 Análisis quimiinformático de contenido

5.1.3.1 Distribución de propiedades

La Figura 5 muestra diagramas de caja de la distribución de las seis propiedades fisicoquímicas calculadas para BIOFACQUIM. Para realizar la comparación, los diagramas de caja incluyen la distribución de las mismas propiedades de los siete conjuntos de datos de referencia (Tabla 1). Las tres propiedades moleculares principales de tamaño, flexibilidad y polaridad molecular están descritas por MW; RB; y SlogP, TPSA, HBA y HBD, respectivamente. En estas gráficas, las cajas encierran los datos con valores dentro del primer y tercer cuartil; la línea que divide la caja denota la mediana de las distribuciones, y las líneas arriba y abajo indican los valores adyacentes superior e inferior. Los asteriscos rojos indican los puntos de datos con valores más allá de los valores adyacentes superior e inferior. La figura también incluye una tabla debajo de cada diagrama de caja con el máximo, la mediana, la media, la desviación estándar y los valores mínimos para cada propiedad y cada base de datos.

De acuerdo con la Figura 5, con base a la media de RB, los compuestos BIOFACQUIM tienen una flexibilidad comparable a la de los fármacos aprobados. La figura también muestra que, a excepción de los metabolitos de las cianobacterias, todas las bases de datos tienen una mediana de hasta 5 enlaces rotables (incluidos los fármacos aprobados). La media y mediana de la MW de BIOFACQUIM son 340.5 y 412 g / mol, respectivamente. En particular, BIOFACQUIM y

NuBBE_{DB} tienen el perfil de MW más similar en comparación con los fármacos aprobados. BIOFACQUIM tiene una mediana de 4 HBA, el mismo número que los datos de NuBBE_{DB} y metabolitos de especies marinas. Además, BIOFACQUIM tiene un perfil muy similar de HBA en comparación con MEG_x. Comparando HBD, BIOFACQUIM, NuBBE_{DB}, NAT_x y los metabolitos de cianobacterias tienen valores similares de mediana con el perfil de los fármacos aprobados, aunque con una desviación estándar más alta que estos. Con respecto a TPSA, los compuestos en BIOFACQUIM son aquellos que comparten los valores más cercanos a los fármacos aprobados. Cabe señalar que el conjunto de metabolitos de las cianobacterias tiene la mayor distribución y los valores medios más altos de TPSA, siendo el doble de la media de los fármacos aprobados. La distribución de los valores de SlogP indica que, en general, los productos naturales son ligeramente más hidrofóbicos que los fármacos aprobados.

Considerando los resultados del perfil general de la distribución de propiedades, se puede concluir que la versión actual de BIOFACQUIM es, en general, la más similar a los conjuntos de datos de NuBBE_{DB} y metabolitos de hongos. Este resultado está de acuerdo con los hallazgos que se reunieron al construir BIOFACQUIM, ya que al analizar en detalle el origen de los productos, resultó que la mayoría de los compuestos fueron aislados en su mayoría de plantas y hongos.

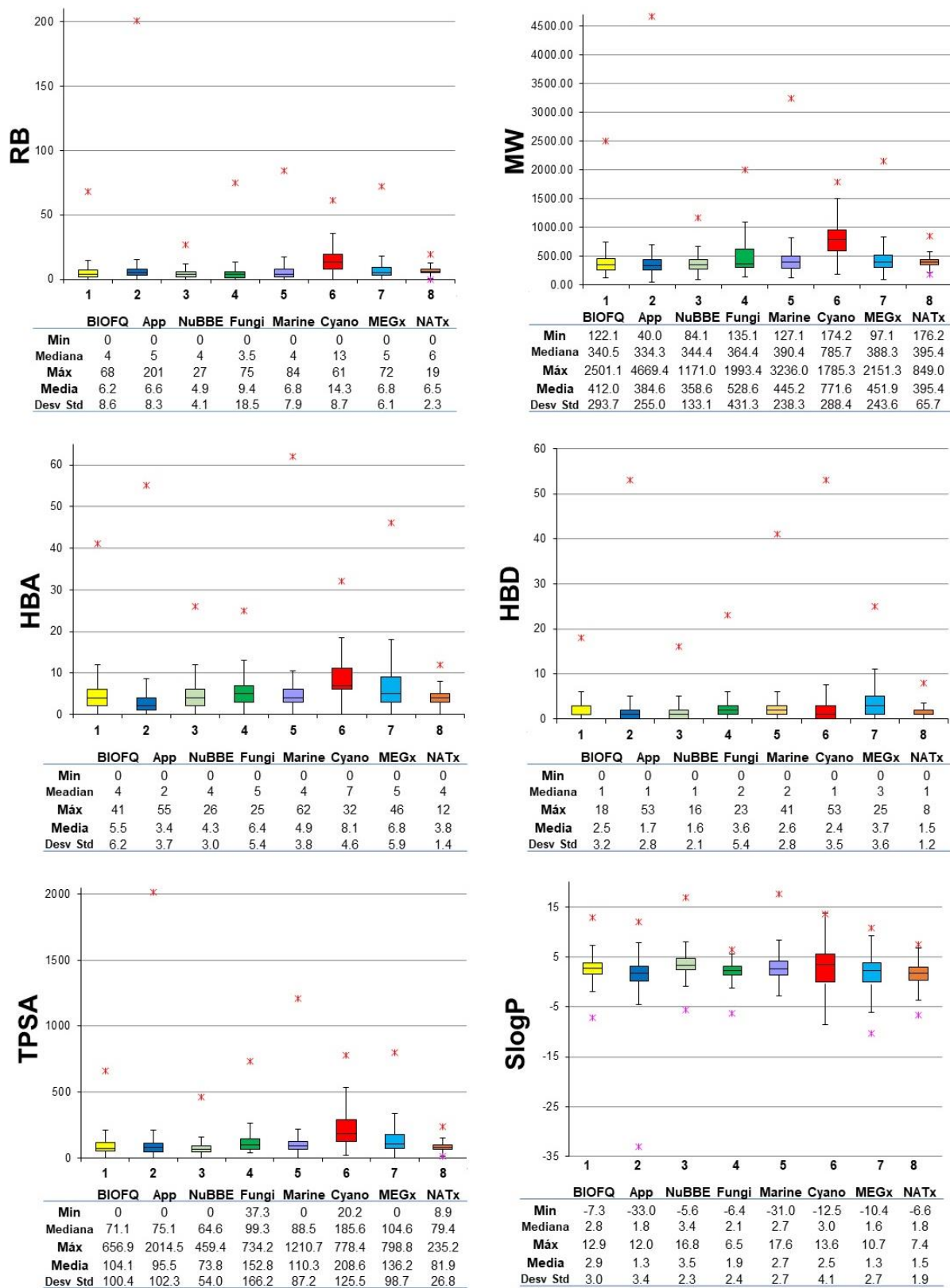


Figura 5. Diagramas de caja de las propiedades fisicoquímicas de BIOFACQUIM (BIOFQ) y los conjuntos de datos de referencia (Tabla 1). Las cajas encierran puntos de datos con valores dentro del primer y tercer cuartil. Los asteriscos rojos indican valores atípicos. Las estadísticas de resumen se incluyen debajo de cada gráfico.

5.1.3.1.2 Espacio químico por análisis de PCA

La Figura 6 muestra una representación visual del espacio químico basado en propiedades. La Tabla 4 resume los valores correspondientes para las tres primeras PC. Las dos primeras PC capturan el 84 % de la varianza, mientras que las tres primeras recuperan el 92 % de la varianza. La Tabla 4 muestra que para el primer componente las propiedades que tienen más peso o contribución corresponden a SlogP, seguidas de RB. Para el segundo componente la contribución más grande corresponde a HBD.

La representación visual del espacio químico en la Figura 6 indica que algunos de los compuestos de productos naturales ocupan el mismo espacio que los fármacos ya aprobados. También muestra que hay moléculas en BIOFACQUIM y del conjunto de metabolitos de especies marinas que cubren las regiones desatendidas del espacio químico que cubre actualmente el de fármacos aprobados. Finalmente, la Figura 6 sugiere que BIOFACQUIM comparte el espacio químico de casi todos los metabolitos de hongos y de NuBBEDB.

Tabla 4. Valores de los tres primeros componentes principales del espacio químico de las ocho bases de datos. Se muestran los valores correspondientes por propiedad fisicoquímica.

Componente principal	PC1	PC2	PC3
Eigenvalor	1.98	1.05	0.71
Eigenvalor acumulado (%)	65.58	83.85	92.15
SlogP	0.18	-0.86	0.23
TPSA	-0.49	0.04	0.21
MW	-0.45	-0.31	0.13
HBA	-0.45	-0.04	0.47
HBD	-0.44	0.23	-0.08
RB	-0.37	-0.33	-0.81

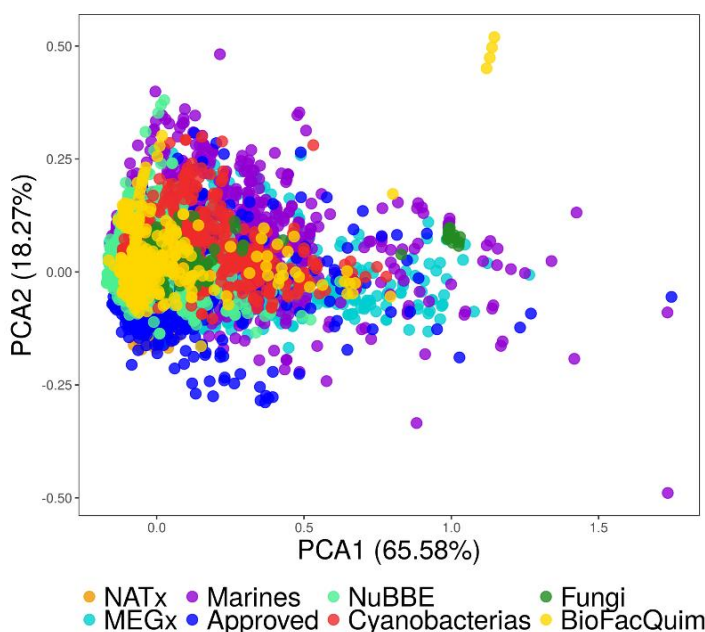


Figura 6. Representación visual 2D del espacio químico basado en las propiedades fisicoquímicas de ocho conjuntos de datos. BIOFACQUIM (423 compuestos, amarillo); Metabolitos de especies de hongos (206 compuestos, verde); Metabolitos de cianobacterias (473 compuestos, rojo); NuBBE_{DB} (2214 compuestos, verde claro); NATx (26318 compuestos, naranja); MEGx (4103 compuestos, azul); Metabolitos de especies marinas (6253 compuestos, lila); Fármacos aprobados por la FDA (1806 compuestos, azul oscuro).

5.1.3.2 Núcleos base: contenido y diversidad

5.1.3.2.1 Núcleos base según Bemis y Murcko (*scaffolds*)

La Figura 7 muestra las estructuras químicas de los 27 núcleos base más frecuentes presentes en BIOFACQUIM que incluyen la mitad (50.6 %) de los 423 compuestos que componen la base de datos. Además del benceno, que también es muy frecuente en otras bases de datos de compuestos [21], el segundo núcleo base más frecuente es un núcleo base relacionado con las flavonas (5 %), seguido de 1,3-benzodioxol y el núcleo dibencílico (2.4 %). Es interesante notar que estos tres últimos núcleos base frecuentes en BIOFACQUIM no son los más frecuentes en otras bases de datos de productos naturales [39].

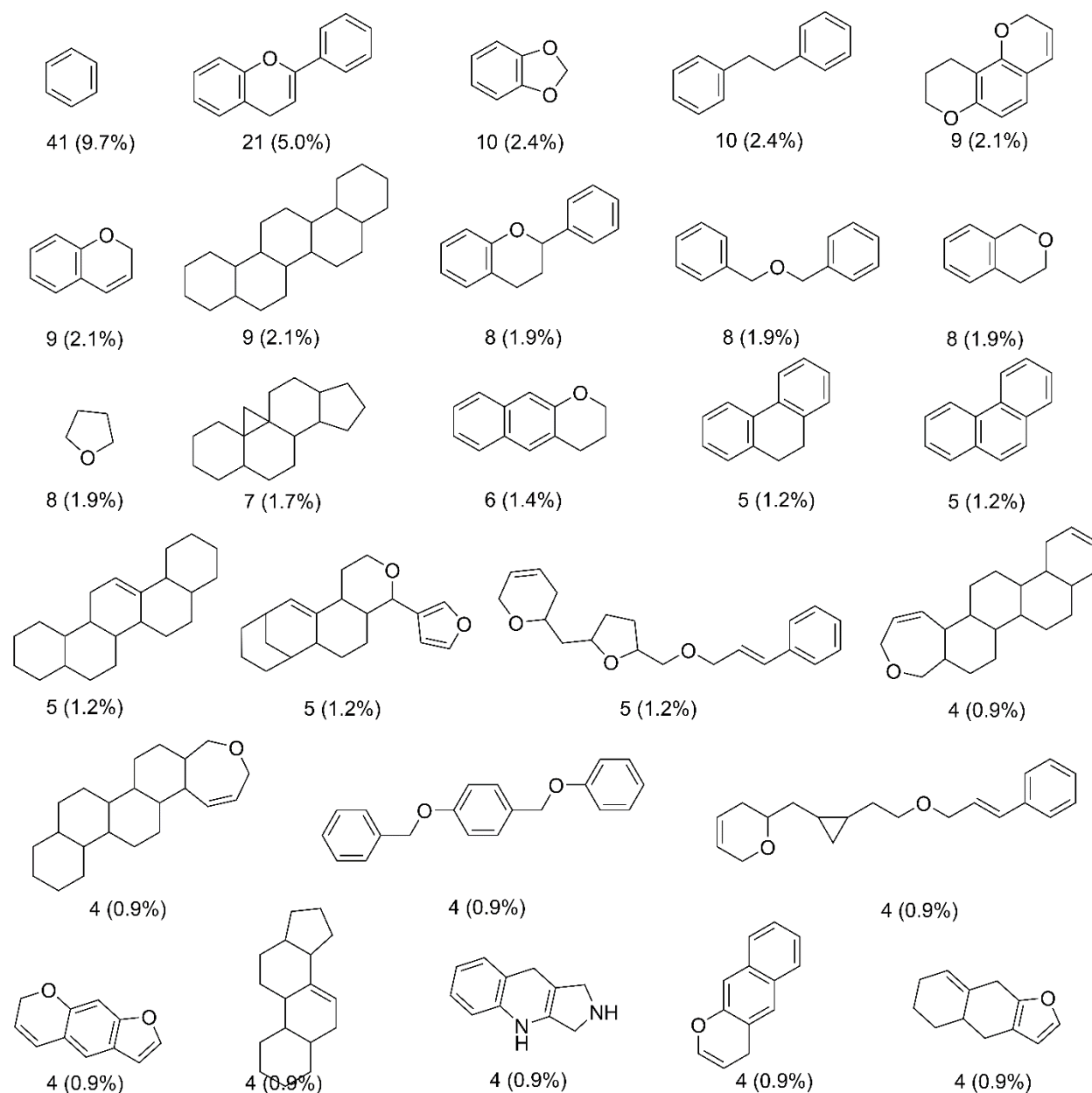


Figura 7. Núcleos base más frecuentes en BIOFACQUIM. Se muestran la frecuencia y el porcentaje que representan en la base de datos. Los 27 núcleos base moleculares que se muestran en la figura contienen la mitad del total de compuestos contenidos en la base de datos (50.6 %).

5.1.3.2.2 Núcleos putativos (*putative core*)

Se presenta un análisis de superposición de núcleos utilizando dos conjuntos de datos de productos naturales, BIOFACQUIM y NuBBEDB, que contienen información sobre productos naturales mexicanos y brasileños, respectivamente.

La motivación de realizar un análisis de superposición de núcleos base es identificar quimiotipos comunes y únicos en estas bases de datos. Como se muestra en la Tabla 5, NuBBE_{DB} y BIOFACQUIM comparten 49 (~5 %) núcleos base según la definición de Bemis y Murcko y alrededor de 106 (~1 %) de núcleos putativos. Por diseño y en la forma en la que se definen, el número de núcleos base únicos de Bemis y Murcko solo puede ser tan alto como el número total de moléculas únicas, mientras que este es el número mínimo de núcleos putativos que se pueden encontrar. Esto explica por qué se encuentran más núcleos putativos que los núcleos de Bemis-Murcko. Notablemente, si un núcleo se comparte entre dos bases de datos, se podría construir una serie analógica para ese núcleo (Figura 8a). Por otro lado, un andamio Bemis y Murcko compartido podría no representar una serie analógica significativa (Figura 8b).

Tabla 5. Sobrelapamiento de núcleos putativos y núcleos base de Bemis y Murcko de BIOFACQUIM y NuBBE_{DB}.

Núcleo analizado		BIOFACQUIM	NuBBE _{DB}	Ambos
	Moléculas únicas Intra DB	399	2018	2417
	Moléculas únicas Inter DB	344	1963	2362 (55 compartidos)
	Núcleos intra DB	1356	15,758	17,114
Núcleos putativos	Núcleos únicos intra DB	1153	11,738	12,289
	Núcleos únicos inter DB	1047	11,632	12,785 (106 compartidos)
Núcleos base según Bemis y Murcko	<i>Scaffolds</i> intra DB	396	1921	2317
	<i>Scaffolds</i> únicos Intra DB	176	754	930
	<i>Scaffolds</i> únicos Inter DB	127	705	881 (49 compartidos)

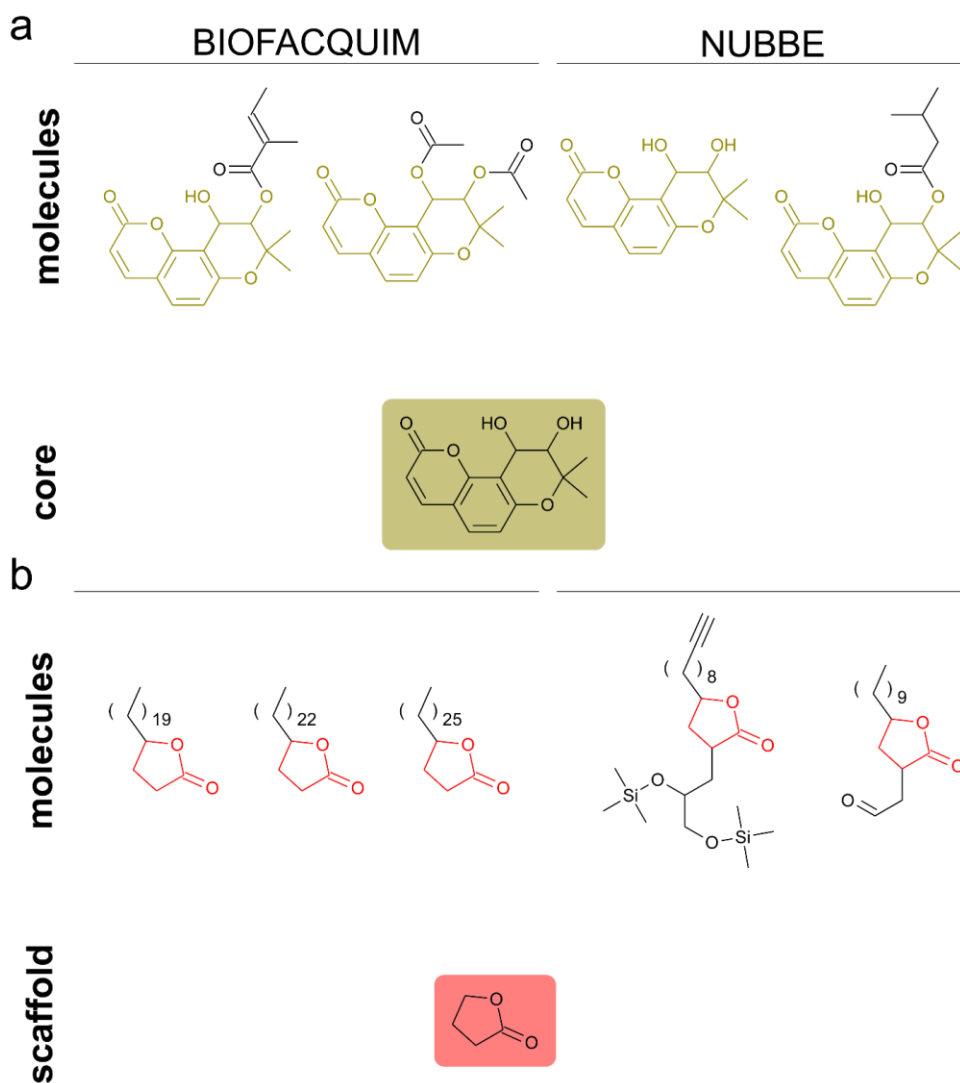


Figura 8. Ejemplos de superposición de núcleos base de BIOFACQUIM y NuBBE_{DB}. a) Para cualquier núcleo superpuesto, se puede encontrar una serie análoga con el mismo núcleo putativo; b) Este no es necesariamente el caso de la superposición de núcleos base Bemis y Murcko.

5.1.3.3 Huellas digitales: diversidad y espacio químico

5.1.3.3.1 Espacio químico por t-SNE

La Figura 9 muestra una representación visual del espacio químico de la versión actual de BIOFACQUIM basada en huellas digitales moleculares utilizando t-SNE. La Figura 9a compara BIOFACQUIM con todos los otros conjuntos de datos de referencia (Tabla 1). La Figura 9b

muestra una comparación de BIOFACQUIM con fármacos aprobados. La Figura 9a muestra tres grupos o *clusters* principales en los que todas las bases de datos tienen compuestos. Los grupos indican que el método de visualización y las huellas digitales pueden distinguir tres estructuras centrales principales que tendrían variaciones detalladas en la estructura. La Figura 9b indica que hay compuestos en BIOFACQUIM con una alta similitud estructural con los fármacos aprobados. Ejemplos notables son los compuestos **FQNP329** (estructura química en la Figura 4), que es similar al etinilestradiol (App_75), y **FQNP130** a la colina (App_878).

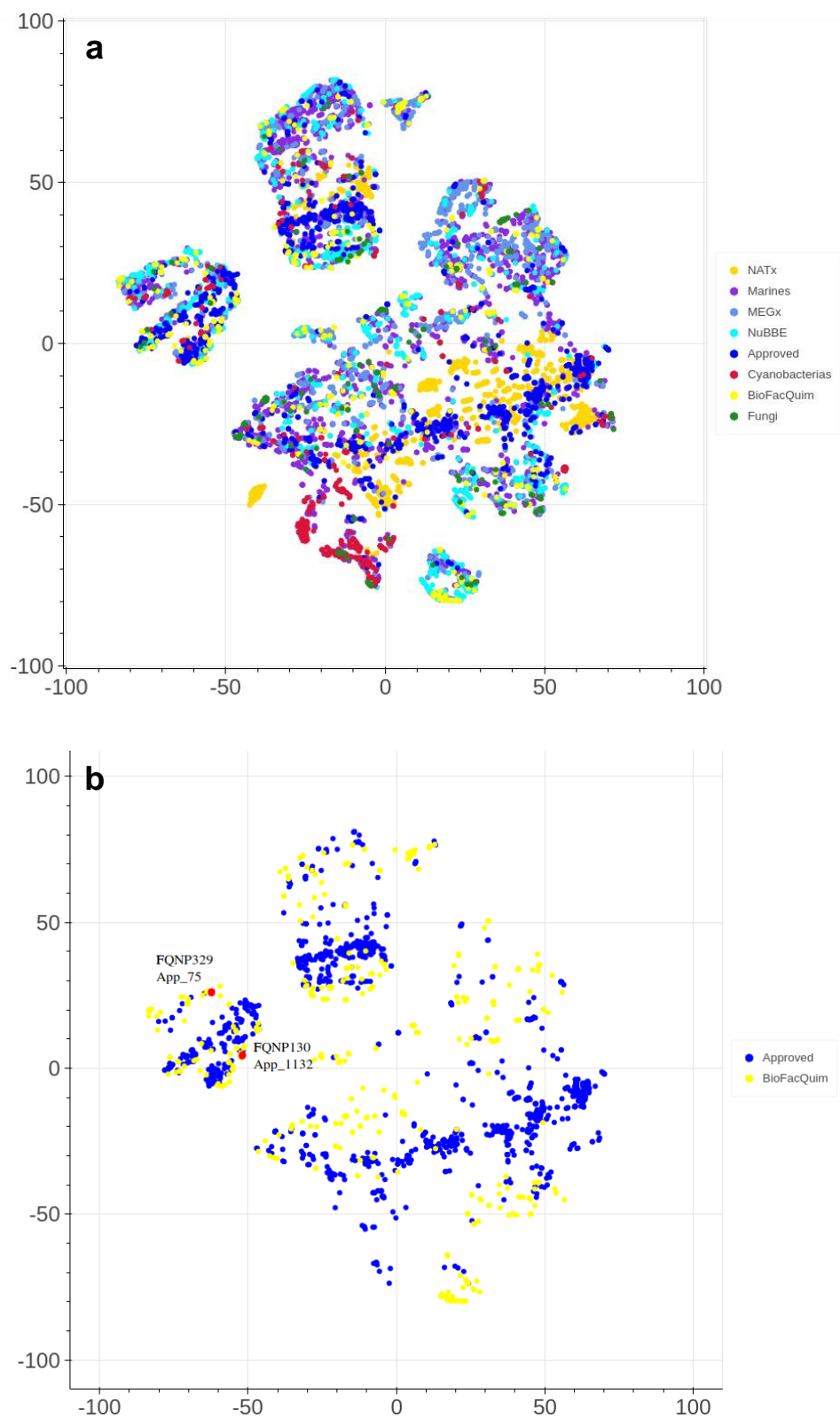


Figura 9. Representación visual 2D del espacio químico basado en huellas digitales topológicas de BIOFACQUIM en comparación con: a) Todos los conjuntos de datos de referencia; b) Fármacos aprobados por la FDA. La visualización se generó utilizando t-SNE.

5.1.4 Diversidad “global”: análisis de diversidad consenso

Tal como se explicó en la sección de Materiales y Métodos sección (4.1.4), se utilizó un CDplot para comparar la diversidad de BIOFACQUIM con la diversidad de los conjuntos de datos de referencia basados en huellas digitales moleculares, núcleos base, y propiedades fisicoquímicas en forma simultánea. La Figura 10 muestra el CDplot que representa en el eje X la diversidad de huellas digitales moleculares con MACCS keys/similitud de Tanimoto. Aquí, los valores más bajos indican una mayor diversidad (Tabla 6). El eje Y del gráfico CD representa la diversidad de núcleos base según la definición de Bemis y Murcko (*scaffolds*) donde los valores más bajos (del área bajo la curva de recuperación de núcleos base Tabla 6) indican una mayor diversidad de núcleo. La diversidad basada en propiedades de BIOFACQUIM y cada base de datos se calculó como la distancia euclidiana de las seis propiedades fisicoquímicas escaladas, descritas anteriormente. Los valores se representaron en el gráfico utilizando una escala de color continua: el color más oscuro representa una diversidad menor, mientras que el color más claro representa una diversidad más alta. Finalmente, el tamaño relativo de las bases de datos se representa con diferentes tamaños de los puntos: los puntos más pequeños indican conjuntos de datos con menor número de moléculas.

Analizando el gráfico en la Figura 10 se observa que BIOFACQUIM y los metabolitos de cianobacterias se encuentran en el área que representa una baja diversidad de *scaffolds* y huellas digitales moleculares. Esto se puede atribuir al hecho de que esta es la primera versión de la base de datos y cuenta con pocos compuestos, en comparación con las otras bases de datos. Con respecto a la diversidad basada en las propiedades fisicoquímicas, se observa que los metabolitos de las cianobacterias tienen una mayor diversidad (un punto de datos azul más claro en la Figura 10) en comparación con BIOFACQUIM. Esto es consistente con el análisis

de los diagramas de caja discutidos en la sección 5.1.3.1. El gráfico de diversidad consenso en la Figura 10 también indica que los fármacos aprobados tienen una alta diversidad de núcleos base y huellas digitales moleculares, lo que es consistente con reportes anteriores [49,54].

Tabla 6. Estadísticos de las representaciones moleculares utilizado para el CDplot de BIOFACQUIM y las bases de datos de referencia.*

DB	PCP EDmediana	FP Tmediana	Scaffold AUC	Tamaño relativo
Approved	1.96	0.32	0.59	699
BIOFACQUIM	1.74	0.45	0.72	423
Cyanobacteria	2.64	0.50	0.74	473
Fungi	1.39	0.44	0.66	206
MEGx	2.28	0.43	0.60	1000
Marines	1.93	0.40	0.58	1500
NATx	3.04	0.51	0.55	2000
NuBBE	2.51	0.39	0.67	1000

* Bases de datos (DB), diversidad fisicoquímica (PCP), distancia Euclidiana (ED), *fingerprint*/huellas digitales moleculares (FP), coeficiente de Tanimoto (T), área bajo la curva (AUC).

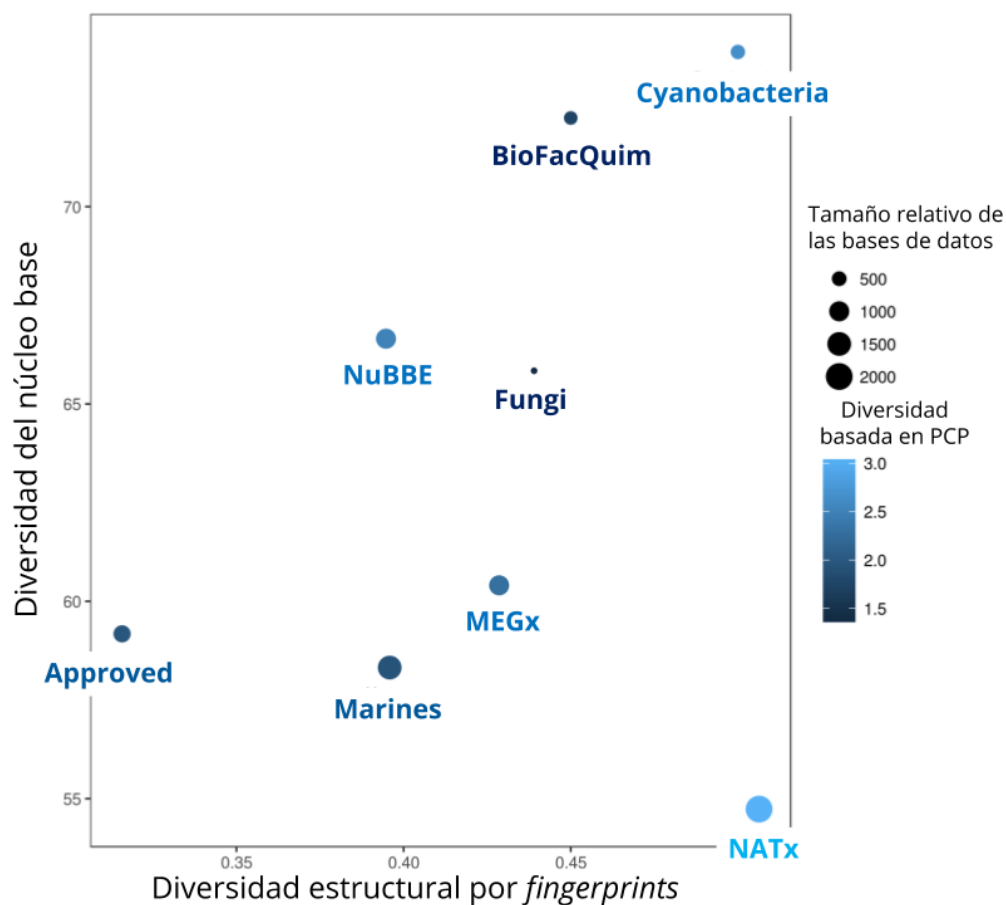


Figura 10. Gráfico de diversidad consenso que compara la diversidad global de BIOFACQUIM con las bases de datos de referencia (Tabla 1). La diversidad estructural (diversidad de huellas digitales moleculares) se calculó con la mediana del coeficiente de Tanimoto de MACCS keys y se representan en el eje X. La diversidad de *scaffolds* de cada base de datos se definió como el área bajo la curva (AUC) de las respectivas curvas de recuperación de *scaffolds*, y se representa en el eje Y. La diversidad basada en las propiedades fisicoquímicas (PCP) se calculó con la distancia euclidiana de seis propiedades escaladas (SlogP, TPSA, MW, RB, HBD y HBA) y se muestra en una escala de colores. La distancia se representa con una escala de color continua desde azul claro (más diverso) hasta azul oscuro (menos diverso). El tamaño relativo del conjunto de datos se representa con el tamaño del punto: los puntos de datos más pequeños indican conjuntos de datos compuestos con menos moléculas.

5.2 Perfil *in silico*

5.2.1 Búsqueda por similitud vs. dianas epigenéticas

5.2.1.1 Espacio químico por PCA

Como primera estrategia se hizo la comparación del espacio químico de BIOFACQUIM con inhibidores de DNMT1, DNMT3A, DNMT3B, HDAC1, HDAC2, HDAC3, BRD2, BRD3, BRD4 y BRD9 (Tabla 3).

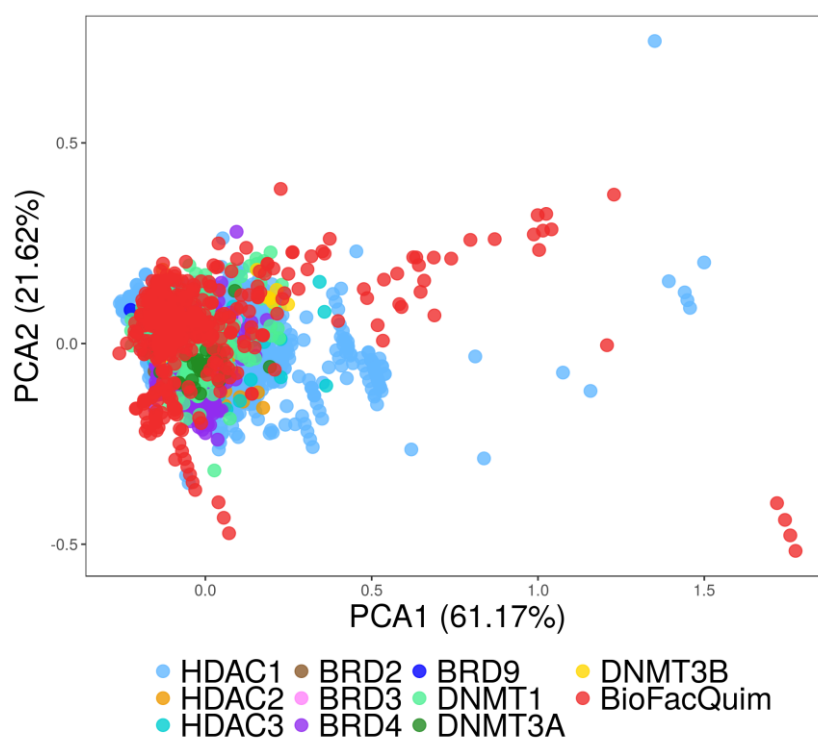


Figura 11. Representación visual 2D del espacio químico basado en las propiedades fisicoquímicas de once conjuntos de datos. BIOFACQUIM (423 compuestos, rojo); inhibidores HDAC1 (4885 compuestos azul claro); inhibidores HDAC2 (149 compuestos, naranja); inhibidores HDAC3 (226 compuestos, turquesa); inhibidores BrD2 (276 compuestos, café); inhibidores BrD3 (26 compuestos, rosa); inhibidores BrD4 (1200 compuestos, morado); inhibidores BrD9 (35 compuestos, azul oscuro); DNMT1 (201 compuestos, aguamarina); DNMT3A (36 compuestos, verde); DNMT3B (7 compuestos, amarillo).

La Figura 11 muestra una representación visual del espacio químico basado en las seis propiedades fisicoquímicas. Los primeros dos PC capturan el 82.79 % de la varianza. Al igual que en la sección 5.1.3.1.2, donde se evaluó el espacio químico contra fármacos aprobados y otros productos naturales, las propiedades con mayor contribución al primer componente principal fueron SlogP, seguido por el RB; en el segundo componente la mayor contribución corresponde al HBD. La visualización del espacio químico indica que los compuestos en BIOFACQUIM tienen un perfil de propiedades fisicoquímicas similar al de compuestos activos conocidos y pudieran presentar actividad contra estas dianas epigenéticas.

Tabla 7. Valores de los tres primeros componentes principales del espacio químico de los once conjuntos de datos. Se muestran los valores correspondientes por propiedad fisicoquímica.

Componente principal	PC1	PC2	PC3
Eigenvalor	1.92	1.14	0.77
Eigenvalor acumulado (%)	61.17	82.79	92.74
SlogP	0.03	-0.85	0.05
TPSA	-0.50	0.18	-0.05
MW	-0.46	-0.30	-0.27
HBA	-0.45	0.08	-0.61
HBD	-0.41	0.25	0.60
RB	-0.41	-0.28	0.44

5.2.1.2 Análisis de similitud por huellas digitales moleculares

Se realizó el perfil de similitud según las huellas digitales moleculares para determinar si estructuralmente los compuestos de cada grupo son parecidos y así tener otro criterio para predecir una posible actividad.

Se hizo la comparación pareada con el coeficiente de Tanimoto usando las huellas digitales MACCS keys y ECFP4. De allí se realizó un consenso con los compuestos que tenían una similitud mayor a su media más una desviación estándar.

La Tabla 8 resume a los compuestos seleccionados por consenso de similitud. En la segunda columna se indica el compuesto seleccionado de BIOFACQUIM, la similitud con el compuesto hallado en D-DB (como referencia o “plantilla”) y, en caso de haber cumplido el filtro, el valor de similitud medio con los demás compuestos de D-DB.

Tabla 8. Compuestos seleccionados de BIOFACQUIM según el criterio de similitud con inhibidores de dianas epigenéticas.

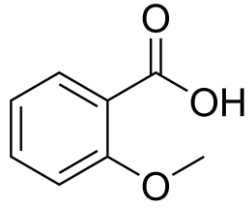
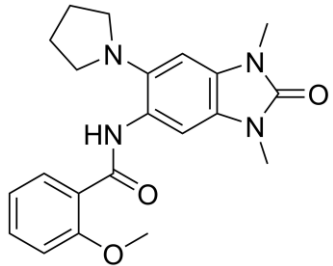
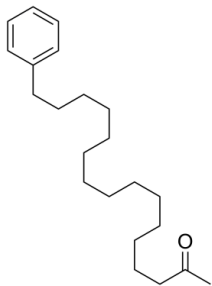
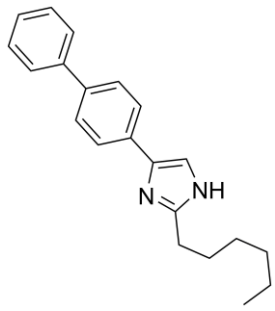
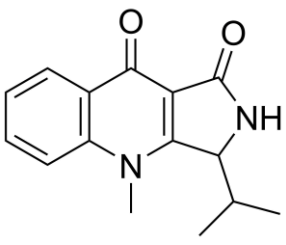
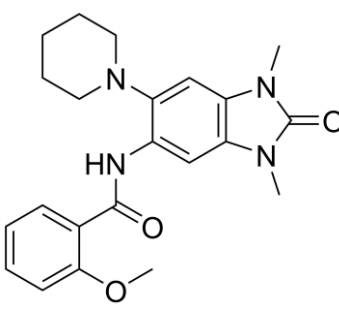
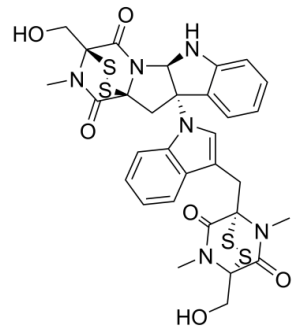
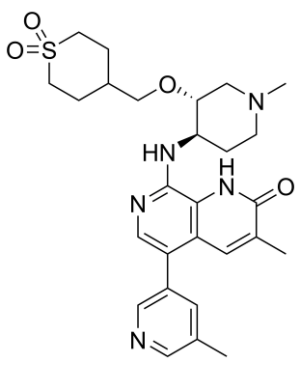
Diana epigenética	BIOFACQUIM	D-DB	BIOFACQUIM	D-DB
BrD1 (ECFP4)	FQNP267	DDB1.007815	FQNP470	DDB1.322690
				
	Media: 0.14		Media: 0.15	
	Similitud: 0.36		Similitud: 0.26	
BrD1 (MACCS keys)	FQNP92	DDB1.095821	FQNP133	DDB1.056049
				
	Media: 0.48		Media: 0.49	
	Similitud: 0.67		Similitud: 0.70	

Tabla 8. Continuación

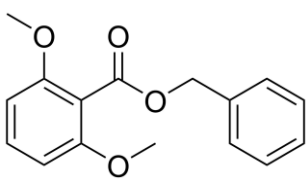
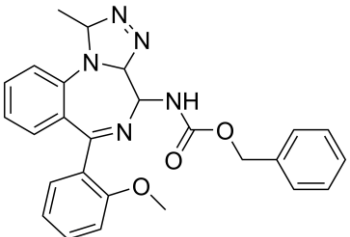
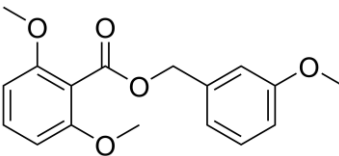
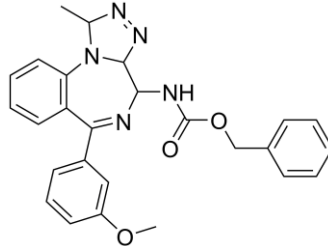
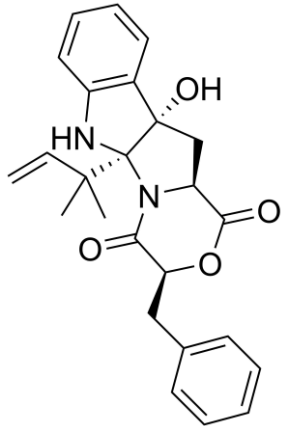
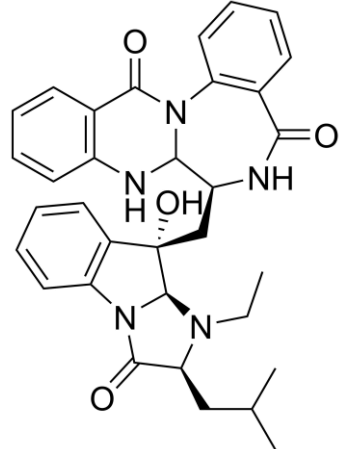
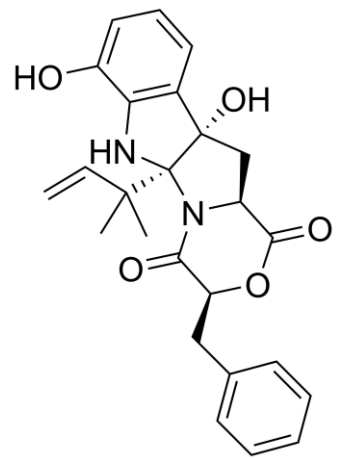
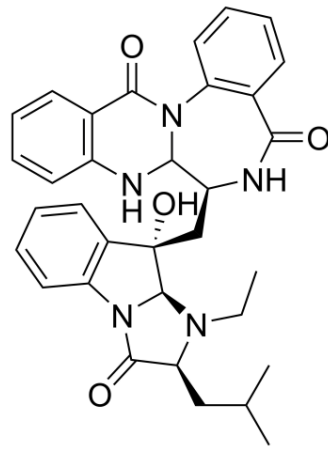
BRD2 (ECFP4)	FQNP439	DDB1.371613	FQNP444	DDB1.218764
				
	Similitud: 0.37		Similitud: 0.34	
BRD2 (MACCS keys)	FQNP46	DDB1.070410	FQNP51	DDB1.070410
				
	Media: 0.48		Media: 0.48	
Similitud: 0.77		Similitud: 0.76		

Tabla 8. Continuación

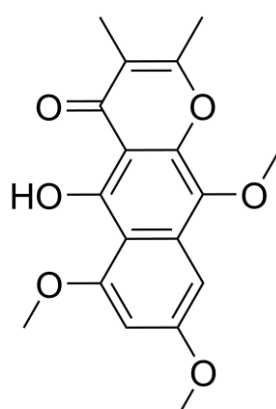
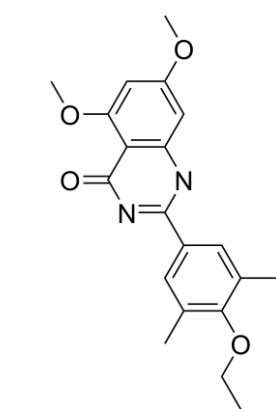
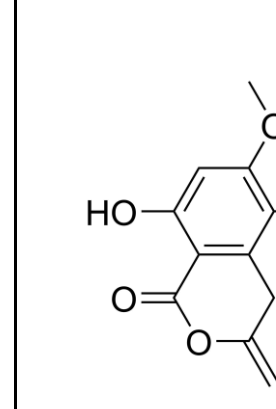
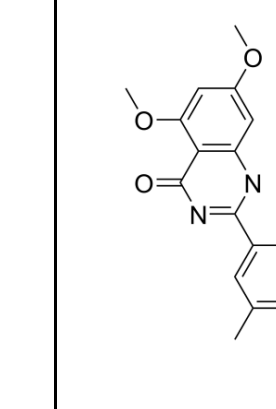
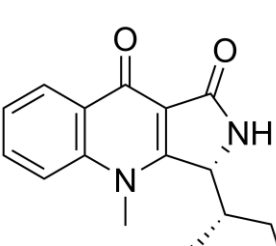
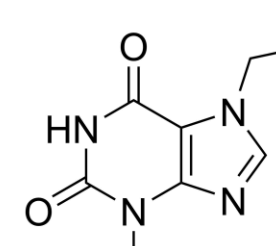
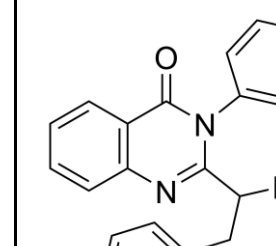
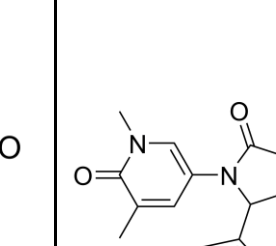
BRD3 (ECFP4)	FQNP340	DDB1.280688	FQNP98	DDB1.280688
				
	Similitud: 0.29		Similitud: 0.28	
BRD3 (MACCS keys)	FQNP90	DDB1.479791	FQNP89	DDB1.379029
	 Media: 0.55		 Media: 0.53	
	Similitud: 0.71		Similitud: 0.70	

Tabla 8. Continuación

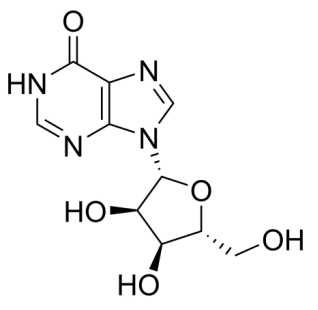
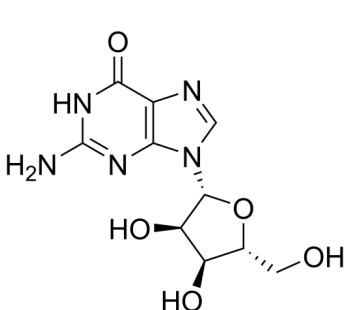
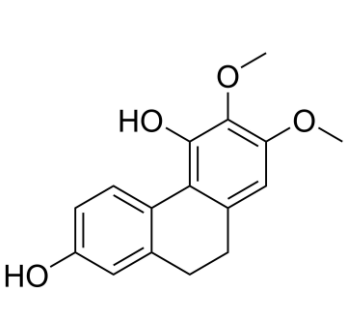
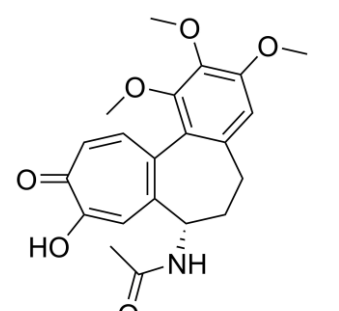
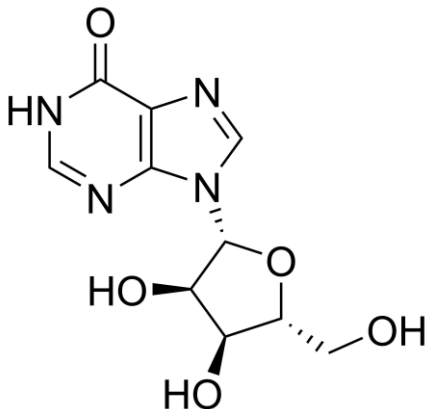
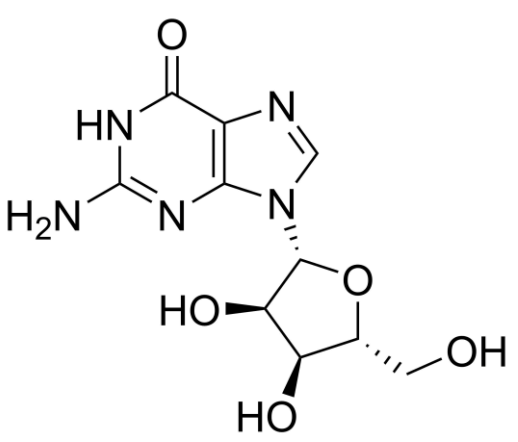
<p>BRD4 (ECFP4)</p>	<p>FQNP323</p> 	<p>DDB1.093590</p> 	<p>FQNP316</p> 	<p>DDB1.349675</p> 
	<p>Similitud: 0.67</p>		<p>Similitud: 0.32</p>	
	<p>FQNP323</p>		<p>DDB1.093590</p>	
<p>BRD4 (MACCS keys)</p>	 <p>Media: 0.52</p>			
	<p>Similitud: 0.93</p>			

Tabla 8. Continuación

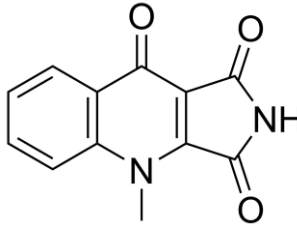
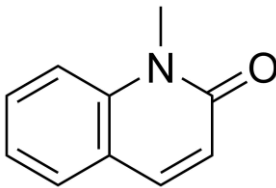
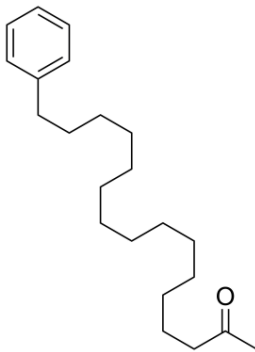
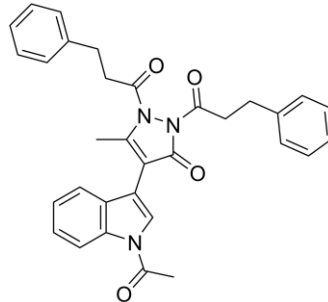
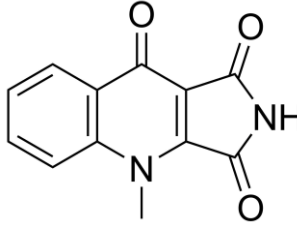
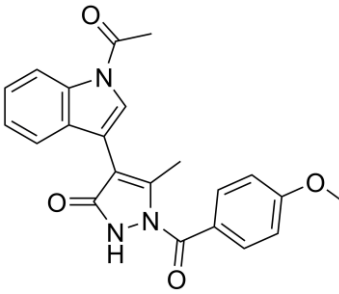
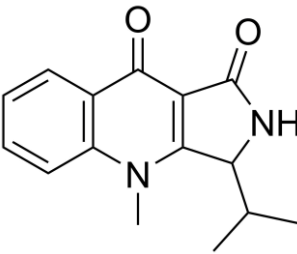
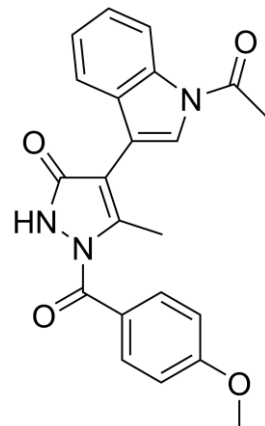
BRD9 (ECFP4)	FQNP93	DDB1.129790	FQNP470	DDB1.440529
				
	Similitud: 0.32		Similitud: 0.31	
BRD9 (MACCS keys)	FQNP93	DDB1.440530	FQNP92	DDB1.440530
				
	Media: 0.56		Media: 0.57	
Similitud: 0.76		Similitud: 0.74		

Tabla 8. Continuación

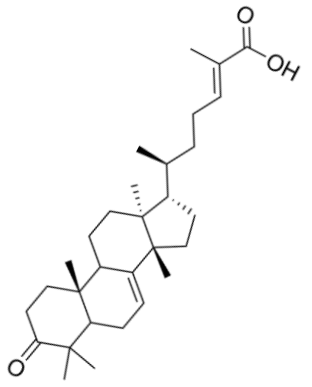
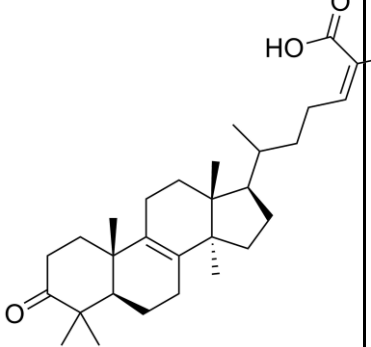
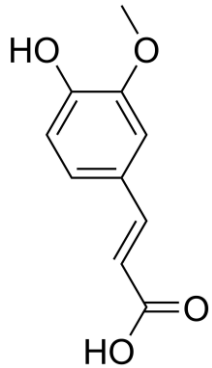
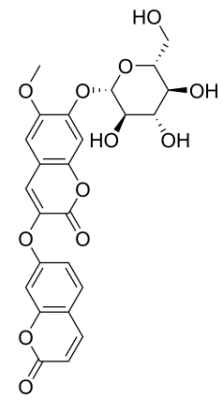
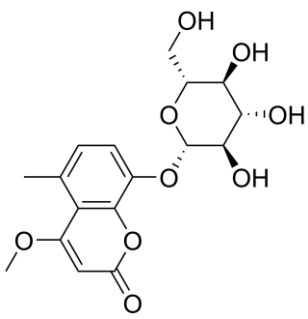
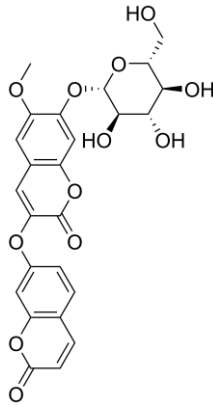
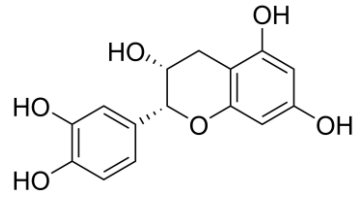
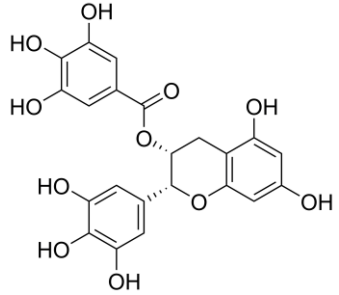
	FQNP152	DDB1.352994	FQNP173	DDB1.157612
DNMT1 (ECFP4)				
	Similitud: 0.53		Similitud: 0.51	
DNMT1 (MACCS keys)	FQNP88	DDB1.157612	FQNP274	
				
Similitud: 0.95		Similitud: 0.92		

Tabla 8. Continuación

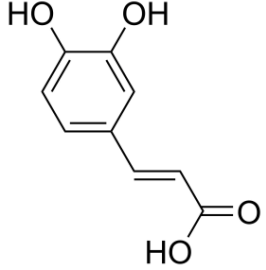
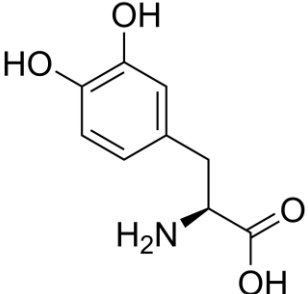
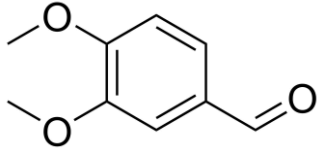
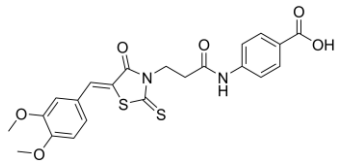
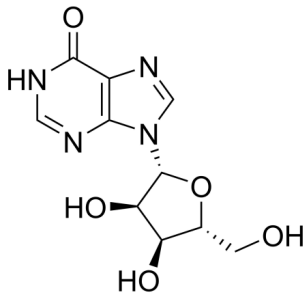
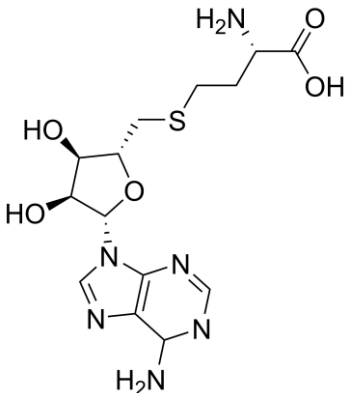
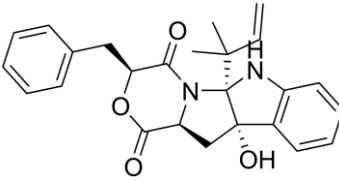
<p>DNMT3A (ECFP4)</p>	<p>FQNP259</p> 	<p>DDB1.000562</p> 	<p>FQNP4</p> <p>Media: 0.15</p> 	<p>DDB1.226610</p> 
	<p>Similitud: 0.35</p>		<p>Similitud: 0.32</p>	
	<p>DNMT3A (MACCS keys)</p>	<p>FQNP323</p>  <p>Media: 0.51</p>	<p>DDB1.193856</p> 	<p>FQNP46</p>  <p>Media: 0.51</p>
<p>Similitud: 0.79</p>		<p>Similitud: 0.69</p>		

Tabla 8. Continuación

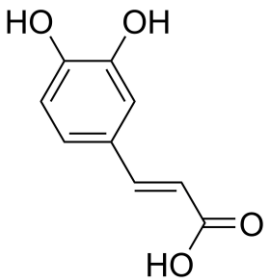
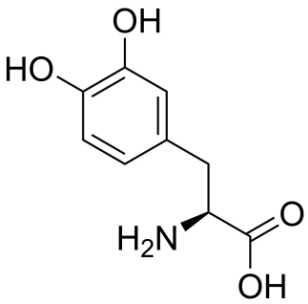
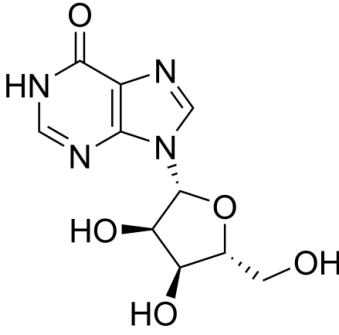
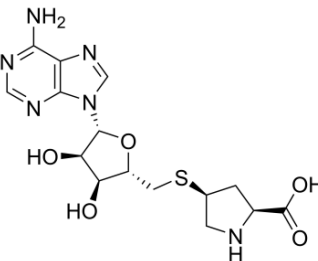
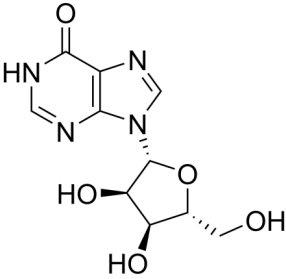
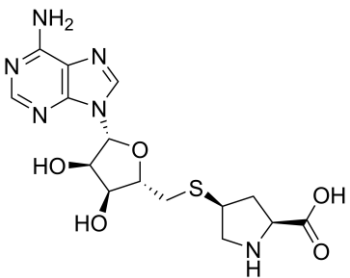
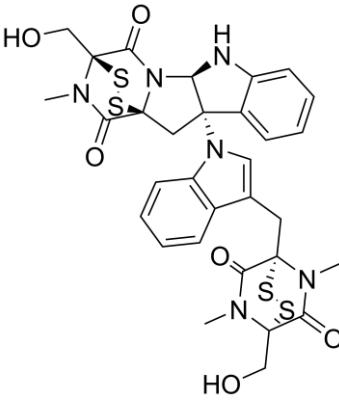
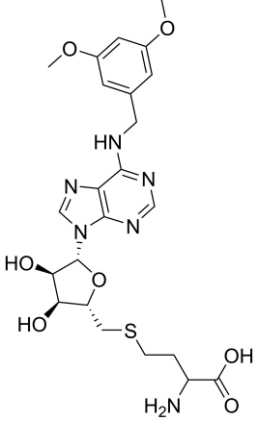
DNMT3B (ECFP4)	FQNP259	DDB1.000562	FQNP323	DDB1.149873
			 Media: 0.23	
	Similitud: 0.35		Similitud: 0.30	
DNMT3B (ECFP4)	FQNP323	DDB1.149873	FQNP133	DDB1.372727
	 Media: 0.69		 Media: 0.60	
	Similitud: 0.82		Similitud: 0.70	

Tabla 8. Continuación

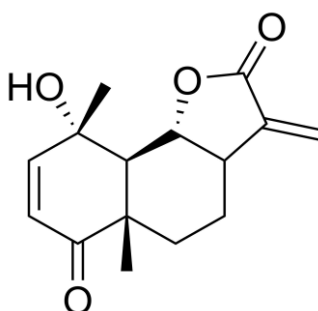
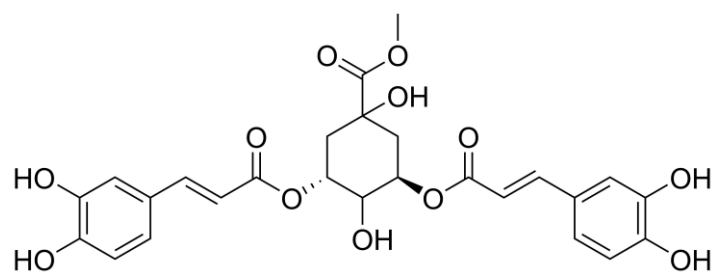
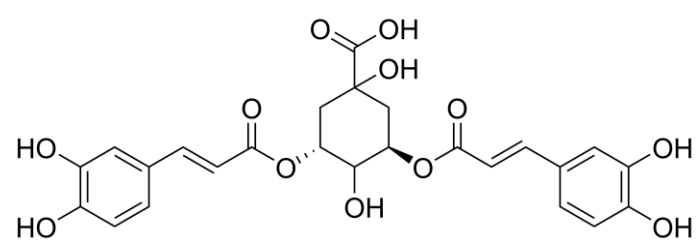
	FQNP116	DDB1.601961
HDAC1, 2, 3 Y 4		
	<p>Similitud ECFP4: 1 Similitud MACCS keys: 1</p>	
	FQNP168	
		
	<p>Similitud ECFP4: 0.76 Similitud MACCS keys: 0.89</p>	

Tabla 8. Continuación

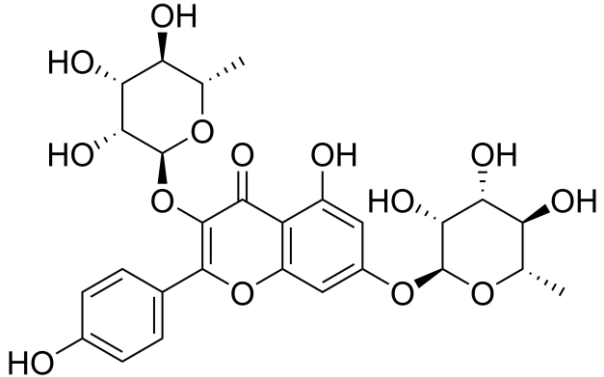
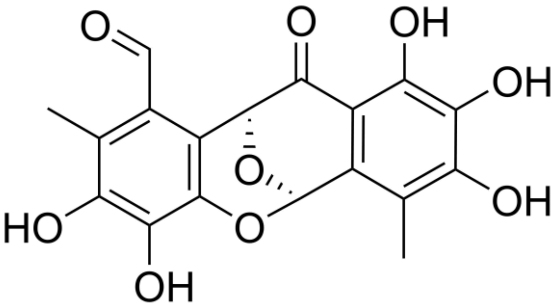
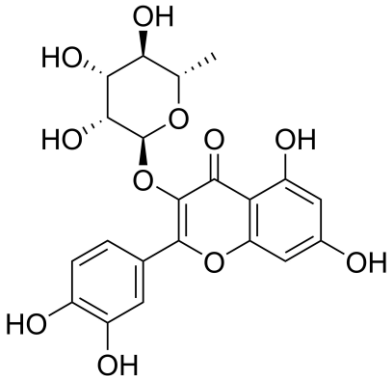
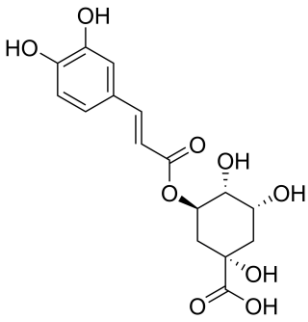
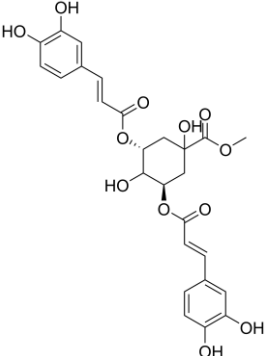
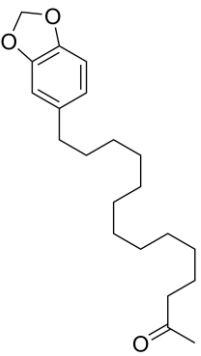
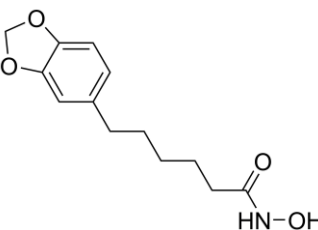
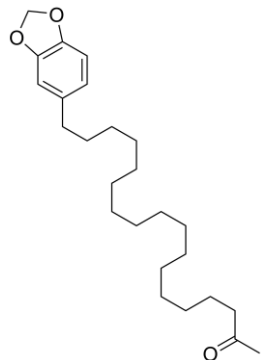
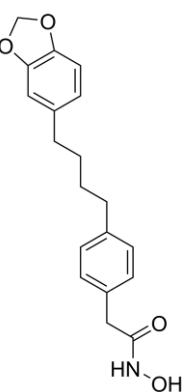
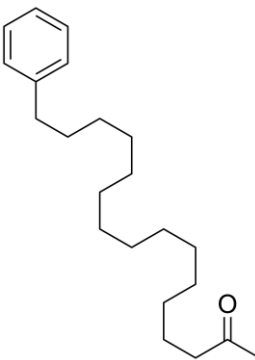
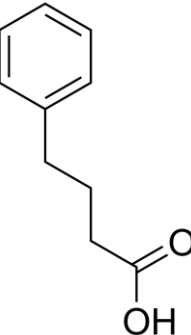
	FQNP139	DDB1.219302	
HDAC 1 Y 2			
	Similitud MACCS keys: 0.94		
	FQNP197		
			
	Similitud MACCS keys: 0.92		

Tabla 8. Continuación

HDAC3	FQNP138	DDB1.601961	FQNP458	DDB1.571314
				
	Similitud ECFP4: 0.70 Similitud MACCS keys: 0.82		Similitud ECFP4: 0.68 Media ECFP4: 0.16	
HDAC4 (ECFP4)	FQNP460	DDB1.401748	FQNP470	DDB1.044336
				
	Media: 0.16		Media: 0.18	
Similitud: 0.57		Similitud: 0.55		

5.2.2 Chemotargets (similitud vs. 4500 dianas)

Se filtraron los resultados obtenidos del programa Chemotargets, seleccionando las interacciones conocidas y predicciones con afinidad > 5 y un valor de confianza > 0.30. No se seleccionaron los compuestos que tenían actividad cualitativa (valor de “A” en el archivo de salida). La Tabla 9 resume las dianas biológicas más comunes y el rango de valores de actividades predichos por Chemotargets.

Tabla 9. Resumen de dianas biológicas más comunes para presentar actividad por los compuestos en BIOFACQUIM usando el programa Chemotargets.

	Conocidos	Predichos
Familias de dianas biológicas más comunes	Enzimas (EC) Cinasas (KC)	Enzimas (EC) Cinasas (KC)
Rango de actividades	5.01-9.77	6.0-11.5

En la Tabla 10 se muestran los compuestos analizados del grupo completo de BIOFACQUIM según el filtro aplicado. Se muestra que al menos la mitad de la base de datos tuvo un valor predicho de actividad. Además, a un 37.1 % de los compuestos se les predice al menos una actividad; 13.48 % de los compuestos tienen una actividad conocida y 10 % a pesar de ya contar con actividad con una diana biológica, son potencialmente activos contra una segunda diana.

Tabla 10. Número de compuestos analizados por Chemotargets.

Compuestos con		Compuestos únicos	
Alguna actividad conocida	57	Con actividad conocida	16
Alguna actividad predicha	198	Con actividad predicha	157
		Compuestos compartidos	41
		Compuestos analizados	214 (50.6 %)

Las Figuras 12 y 13 muestran el porcentaje de compuestos que tienen actividad por familia de proteínas. En global, las cuatro familias de proteínas con más compuestos: EC (enzimas), NR (receptores nucleares), KC (quinasas) y GR (receptores acoplados a proteínas G). El porcentaje se realizó respecto a las 214 moléculas analizadas (Tabla 10). Cabe señalar que en cada columna ningún compuesto se repite, pero pueden repetirse entre las columnas.

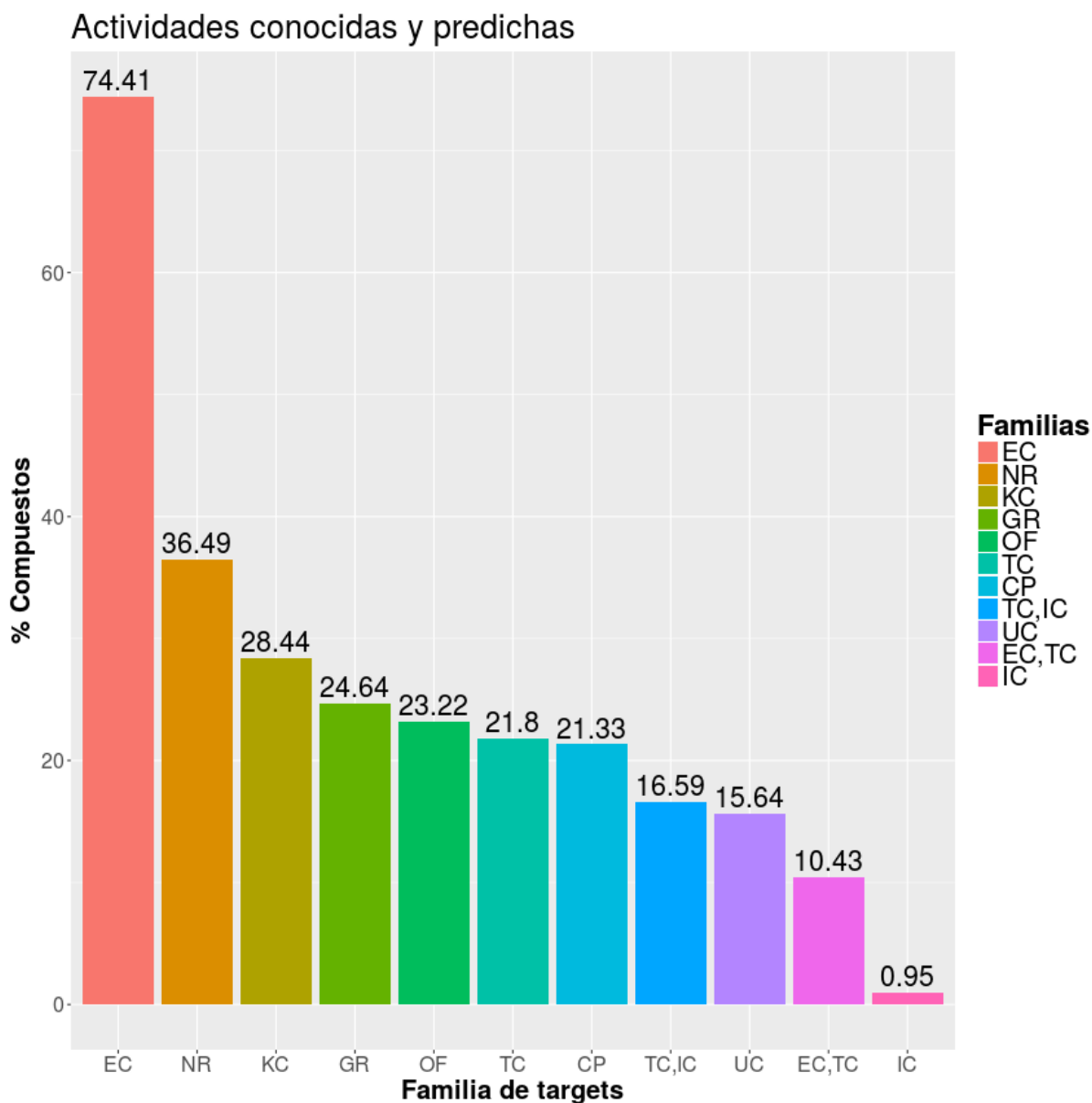


Figura 12. Porcentaje de compuestos con actividad, predicha y conocida, en al menos una diana biológica de cada familia.

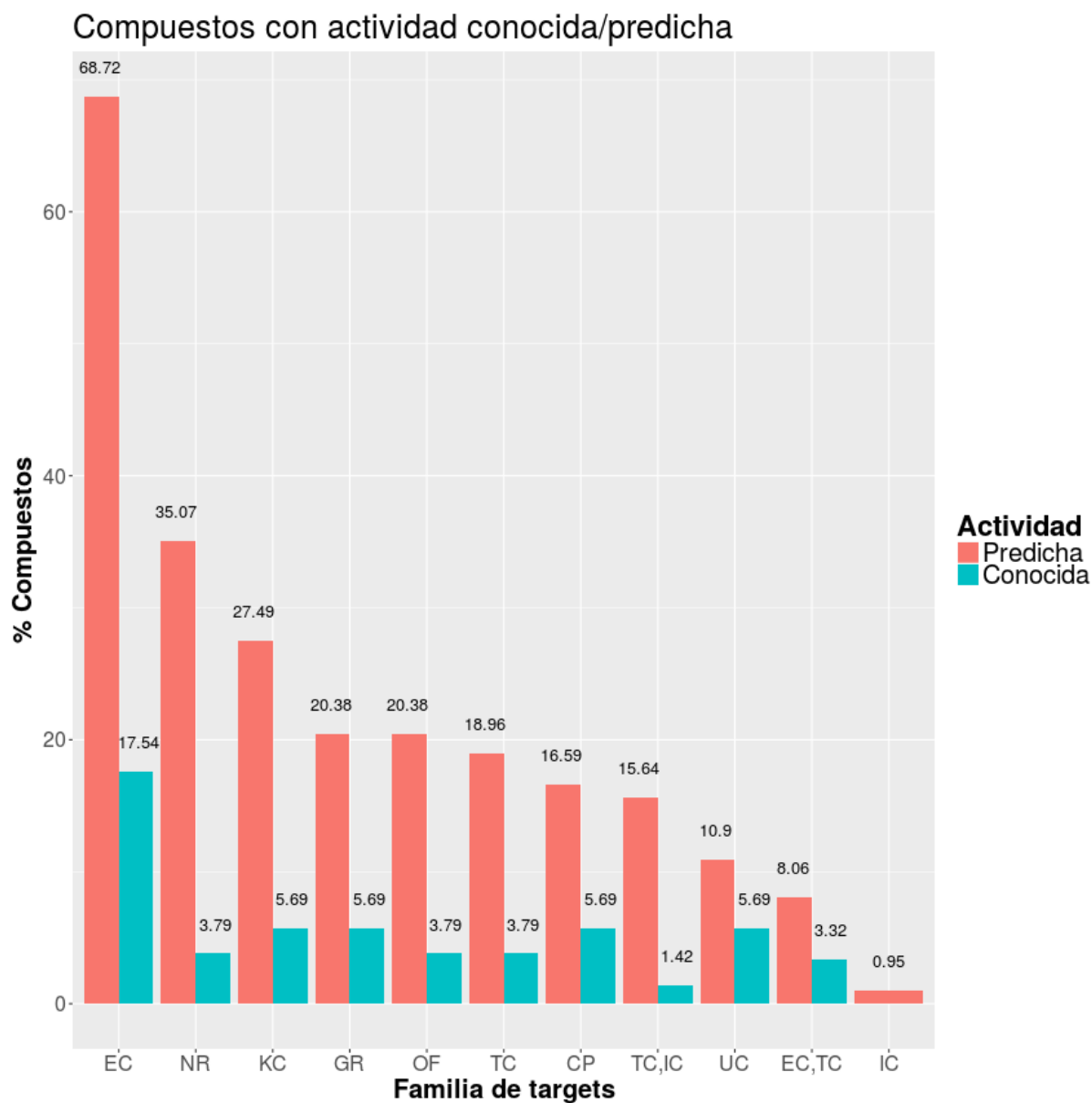


Figura 13. Porcentaje de compuestos con actividad, predicha (rojo) y conocida (azul), en al menos una diana biológica de cada familia.

5.2.2.1 Actividad conocida

Se calculó la actividad máxima, mínima, media y desviación estándar de los compuestos de cada familia de proteínas. Se definieron los criterios indicados en la Tabla 11 de manera heurística, esto con el fin de seleccionar los compuestos con mayor actividad y reducir el número de moléculas por familia.

Tabla 11. Datos estadísticos de cada familia de proteínas con actividad conocida.

Familia *	Max (pAct)	Min (pAct)	Media (pAct)	Desviación estándar (pAct)	Criterio	
EC	8.7	5.01	5.645	0.750	x+2stdv	7.145
KC	7.79	5.05	5.866	0.638	x+1stdv	6.504
CP	8.08	5.1	6.309	0.886	x	6.309
OF	8.09	5.02	5.427	0.626	x	5.427
GR	7.44	5.07	5.765	0.647	x	5.765
UC	6.39	5.08	5.554	0.393	todos	todos
NR	6.18	5.01	5.531	0.429	todos	todos
TC	9.77	5.09	5.879	1.145	todos	todos
EC, TC	6.55	5.05	5.627	0.629	todos	todos
TC, IC	8.02	5.82	6.617	1.219	todos	todos
IC	-	-	-	-	-	-

* EC: enzimas, KC: cinasas, NR receptores nucleares, GR receptores acoplados a proteínas G, CP citocromos, OF otras familias, TC transportadores, IC canales iónicos, UC sin clasificar.

Se aplicó el criterio según el caso de cada diana biológica y se reportan el número de moléculas que se utilizaron para los gráficos (Anexo A2).

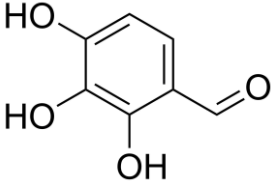
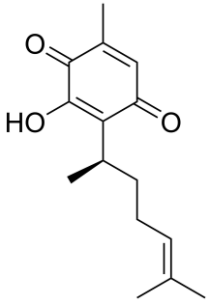
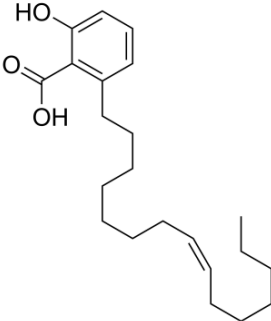
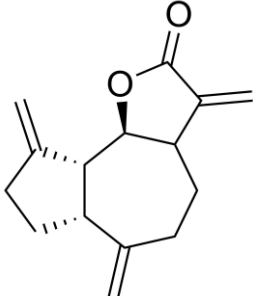
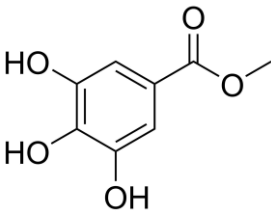
Tabla 12. Número de moléculas de cada familia de proteínas con actividad conocida.

Familia*	Número de moléculas	Número de moléculas aplicado el criterio	Dianas biológicas	Número máximo de moléculas por target (con el criterio)
EC	248	14	9	5
KC	68	11	9	2
CP	50	21	5	8
OF	26	3	3	1
GR	24	19	7	2
UC	16	16	11	3
NR	15	15	9	4
TC	15	15	6	5
EC, TC	7	7	2	5
TC, IC	3	3	1	3
IC	-	-	-	-

* EC: enzimas, KC: cinasas, NR receptores nucleares, GR receptores acoplados a proteínas G, CP citocromos, OF otras familias, TC transportadores, IC canales iónicos, UC sin clasificar.

El número total de actividades conocidas fue 209, después de aplicar el filtro. Este número se tomó como base para calcular la selectividad de los compuestos. La selectividad se eligió con el criterio de si tienen actividad en menos de 5 dianas biológicas se les consideró selectivos y si presentan actividad en más de 20 *targets* son promiscuos [55]. Los resultados de compuestos selectivos y promiscuos representativos están resumidos en las Tablas 13 y 14, respectivamente.

Tabla 13. Compuestos selectivos de BIOFACQUIM con actividad conocida.

ID	Compuesto	Selectividad	Diana molecular
FQNP6		1/209	UC <i>Autoinducer 2-binding periplasmic protein LuxP</i>
FQNP85		1/209	OF <i>Macrophage scavenger receptor types I and II</i>
FQNP149		1/209	EC <i>SUMO-activating enzyme</i>
FQNP379		1/209	OF <i>Transcriptional activator Myb</i>
FQNP260		2/209	EC - <i>FAD-linked sulfhydryl oxidase ALR</i> - <i>Alpha-(1,3)-fucosyltransferase 7</i>

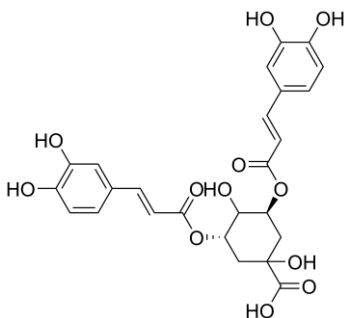
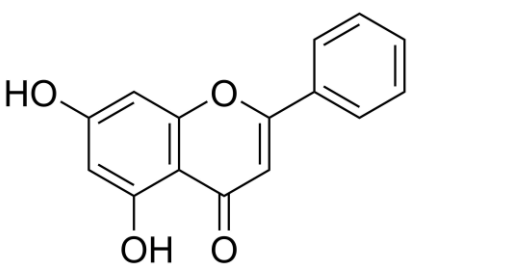
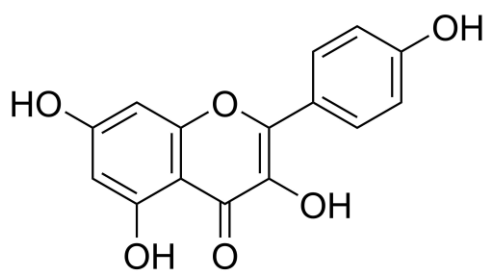
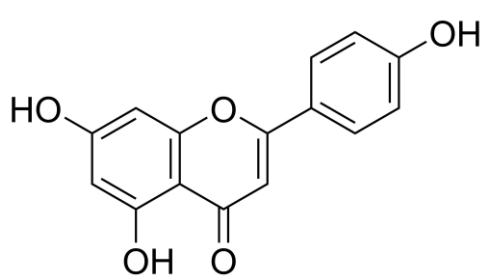
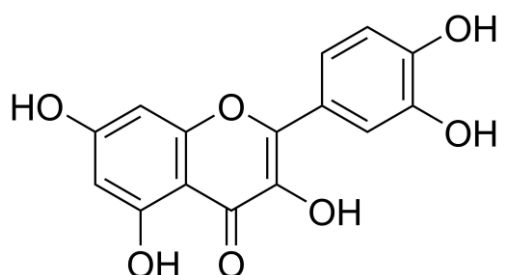
<p>FQNP168</p>		<p>3/209</p>	<p>EC - Aldo-keto reductase family 1 member B10 - Aldose reductase</p> <p>UC HIV-1 integrase</p>
-----------------------	---	--------------	--

Tabla 14. Compuestos promiscuos de BIOFACQUIM con actividad conocida.

ID	Compuesto	Selectividad
<p>FQNP216</p>		<p>30/209</p>
<p>FQNP166</p>		<p>30/209</p>
<p>FQNP317</p>		<p>41/209</p>
<p>FQNP167</p>		<p>76/209</p>

5.2.2.2 Actividad predicha

Se calculó la actividad predicha máxima, mínima, media y desviación estándar de los compuestos de cada familia de proteínas. Se eligieron criterios indicados en la Tabla 15 para la selección de los compuestos con mayor actividad y reducir el número de moléculas por familia.

Tabla 15. Datos estadísticos de cada familia de proteínas con actividad predicha.

Familia*	Max (pAct)	Min (pAct)	Media (pAct)	Desviación estándar (pAct)	Criterio
EC	11	6	6.913	0.833	> 8.579
KC	9.2	6	6.635	0.645	> 7.924
NR	9.7	6	6.872	0.723	> 7.594
GR	9	6	6.838	0.654	> 7.492
CP	8.4	6	6.824	0.576	> 6.824
OF	11.5	6	7.484	1.286	> 7.484
TC, IC	9	6	7.227	0.998	> 7.227
TC	8.9	6	6.696	0.714	> 6.696
UC	8.8	6	6.493	0.627	todos
EC, TC	7.1	6	6.168	0.296	todos
IC	7.2	6.8	7.000	0.283	todos

* EC: enzimas, KC: cinasas, NR receptores nucleares, GR receptores acoplados a proteínas G, CP citocromos, OF otras familias, TC transportadores, IC canales iónicos, UC sin clasificar.

Se aplicó el criterio según el caso de cada diana molecular y se reportan el número de moléculas que se utilizaron para los gráficos (Anexo – A3).

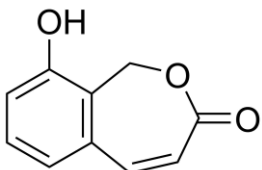
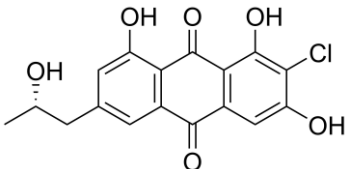
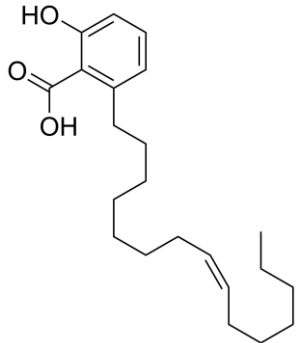
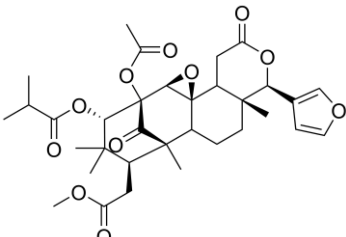
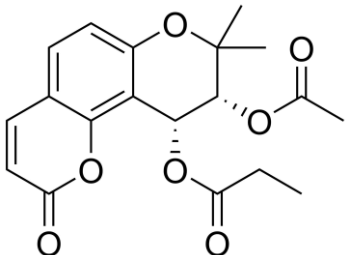
Tabla 16. Número de moléculas de cada familia de proteínas con actividad predicha.

Familia*	Número de moléculas	Criterio	Número de moléculas aplicado el criterio	Dianas biológicas	Número máximo de moléculas por target (con el criterio)
EC	383	x+2stdv	18	10	4
KC	245	x+2stdv	8	2	7
NR	135	x+1stdv	32	10	9
GR	90	x+1stdv	19	12	4
CP	63	x	28	8	14
OF	63	x	30	10	7
TC, IC	48	x	21	9	6
TC	45	x	20	6	8
UC	28	todos	28	12	6
EC, TC	19	todos	19	5	8
IC	2	todos	2	2	1

* EC: enzimas, KC: cinasas, NR receptores nucleares, GR receptores acoplados a proteínas G, CP citocromos, OF otras familias, TC transportadores, IC canales iónicos, UC sin clasificar.

El número total de actividades predichas fueron 324, después de aplicado el filtro. Este número se tomó como base para determinar la selectividad de los compuestos de la misma forma que en la sección 5.2.2.1 [55]. Las Tablas 17 y 18 resumen los resultados de compuestos selectivos y promiscuos representativos, respectivamente. En ambas tablas la columna “Selectividad” cuantifica la proporción del número de dianas con actividad predicha para ese compuesto, entre el número de actividades en total (e.g., 324).

Tabla 17. Compuestos selectivos de BIOFACQUIM con actividad predicha.

ID	Compuesto	Selectividad	Diana biológica
FQNP34		1/324	EC <i>Testosterone 17-beta-dehydrogenase 3</i>
FQNP95		1/324	EC <i>Amine oxidase [flavin-containing] B</i>
FQNP149		1/324	NR <i>Peroxisome proliferator-activated receptor alpha</i>
FQNP236		1/324	TC <i>Solute carrier organic anion transporter family member 1B1</i>
FQNP350		1/324	CP <i>Cytochrome P450 11B2, mitochondrial</i>

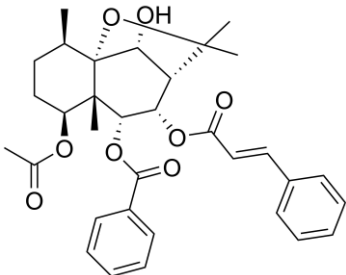
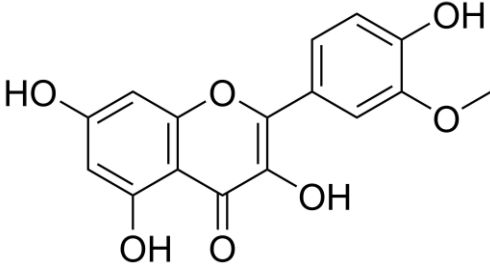
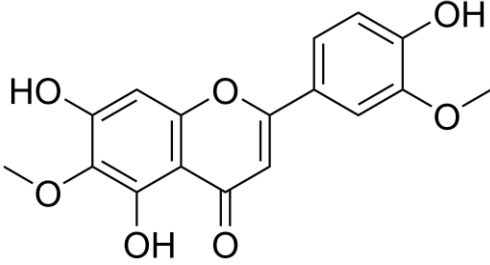
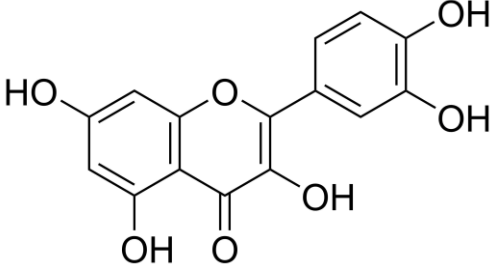
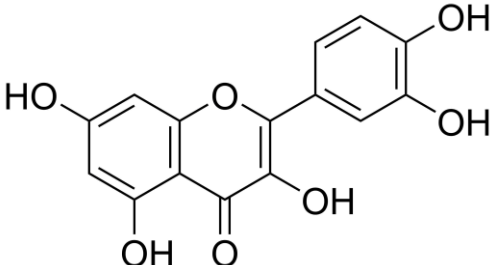
FQNP447		1/324	EC, TC <i>Multidrug resistance protein 1</i>
---------	---	-------	---

Tabla 18. Compuestos promiscuos de BIOFACQUIM con actividad predicha.

ID	Compuesto	Selectividad
FQNP165		26/324
FQNP113		28/324
FQNP167		54/324
FQNP166		59/324

6. CONCLUSIONES

BIOFACQUIM es una base de datos compuesta de productos naturales de México que está siendo construida, curada y mantenida por un grupo académico. La primera y versión actual de BIOFACQUIM descrita en esta tesis tiene 423 compuestos reportados durante los últimos 10 años en la Facultad de Química de la Universidad Nacional Autónoma de México (UNAM). La base de datos contiene el nombre químico, los compuestos en notación SMILES, la referencia (con el nombre de la revista, el año de publicación y número DOI), el reino (*Plantae* o *Fungi*), el género y especie del producto natural. También contiene información de la ubicación geográfica de la recolección. Además, la base de datos incluye la actividad biológica si esta está reportada en la publicación original.

Al igual que en otras bases de datos de productos naturales, BIOFACQUIM se puede usar en la selección virtual para identificar posibles compuestos líder o como puntos de partida para una optimización adicional. La base de datos es de libre acceso a través del sitio web BIOFACQUIM Explorer, versión 1.0 (<https://biofacquim.herokuapp.com>). También se puede acceder libremente como un catálogo dentro de la base de datos de ZINC15 (<http://zinc15.docking.org/catalogs/biofacquimnp>), y además es parte de la iniciativa D-TOOLS (www.difacquim.com/d-tools/) [56].

La evaluación del espacio químico de BIOFACQUIM, en particular su posición relativa con otras bibliotecas de referencia en el espacio químico indicó que los productos naturales en BIOFACQUIM son muy similares a los fármacos basados en cuanto a propiedades fisicoquímicas y huellas digitales moleculares. Por lo tanto, el análisis del espacio químico da soporte al uso de BIOFACQUIM en proyectos de descubrimiento de fármacos.

El análisis de la cobertura y diversidad de BIOFACQUIM en el espacio químico de compuestos activos contra dianas epigenéticas muestra que tienen una amplia cobertura y que se superpone con las regiones en el espacio químico similar a estos.

El análisis por similitud estructural también indicó que hay compuestos en BIOFACQUIM con estructuras químicas muy similares a la de estos compuestos y que pudieran tener actividad contra dianas epigenéticas. La similitud presentada tanto de propiedades fisicoquímicas como estructuralmente con fármacos aprobados y con inhibidores de dianas epigenéticas, sugiere que los compuestos en BIOFACQUIM pueden ser de gran interés farmacéutico.

El perfil multi-diana calculado *in silico* sugirió que hay compuestos en BIOFACQUIM que tienen potencial de ser selectivos contra algunas dianas biológicas de interés terapéutico. Igualmente, se identificaron productos naturales en esta base de datos con el potencial de tener actividad biológica por medio de la interacción con múltiples blancos moleculares. Como se indica en la sección de Perspectivas, el paso siguiente es hacer la selección de compuestos y dianas biológicas a probar experimentalmente.

7. PERSPECTIVAS

- Actualizar la primera versión de BIOFACQUIM. La actualización ya está en desarrollo y se espera liberar la segunda versión en la segunda mitad de 2019.
- Hacer estudios de acoplamiento molecular a BIOFACQUIM con dianas epigenéticas. Los estudios de acoplamiento se realizarán con protocolos validados y establecidos en el grupo de investigación [57, 58].
- Seleccionar compuestos con actividad potencial para su adquisición (por compra, si son de origen comercial, o síntesis) y evaluación biológica en ensayos de inhibición enzimática. Si son de síntesis se hará en colaboraciones ya establecidas con los Dr. Alexandre Gagnon (Universidad de Quebec en Montreal, Canadá) o Dr. Massimo Bertinaria (Universidad de Turín, Italia).
- Cuantificar experimentalmente el perfil multi-diana calculado *in silico* en esta tesis. Las pruebas experimentales se harán mediante una colaboración ya establecida con el Dr. Jordi Mestres (Hospital del Mar y Universitat Pompeu Fabra, Barcelona, España). La selección de compuestos a probar experimentalmente y las dianas biológicas dependerán, entre otros factores, de la disponibilidad de las muestras físicas puras y los costos de los ensayos biológicos.

8. REFERENCIAS

1. Newman DJ, Cragg GM. Natural products as sources of new drugs from 1981 to 2014. *J Nat Prod.* 2016;79: 629–661. doi:10.1021/acs.jnatprod.5b01055
2. A. Hussein R, A. El-Anssary A. Plants secondary metabolites: the key drivers of the pharmacological actions of medicinal plants. In: F. Builders P, editor. *Herbal Medicine.* IntechOpen; 2019. doi:10.5772/intechopen.76139
3. Dewick PM. Secondary metabolism: The building blocks and construction mechanisms. In: John Wiley & Sons Ltd, editor. *Medicinal Natural Products: A Biosynthetic Approach.* Second Edition. 2002. pp. 7–12.
4. Thomford NE, Senthebane DA, Rowe A, Munro D, Seele P, Maroyi A, et al. Natural products for drug discovery in the 21st century: innovations for novel drug discovery. *Int J Mol Sci.* 2018;19. doi:10.3390/ijms19061578
5. Bauer A, Brönstrup M. Industrial natural product chemistry for drug discovery and development. *Nat Prod Rep.* 2014;31: 35–60. doi:10.1039/c3np70058e
6. Alvarez-Ruiz E, Collins AJ, Dann AS, Fosberry AP, Ready SJ, Vázquez Muñiz MJ. *Migalastat Microbiological process.* United States; US20160355856A1, 2018.
7. Pereira DM, Valentão P, Andrade PB. Tuning protein folding in lysosomal storage diseases: The chemistry behind pharmacological chaperones. *Chem Sci.* 2018;9: 1740–1752. doi:10.1039/c7sc04712f
8. Zhanel GG, Lawson CD, Zelenitsky S, Findlay B, Schweizer F, Adam H, et al. Comparison of the next-generation aminoglycoside plazomicin to gentamicin, tobramycin and amikacin. *Expert Rev Anti Infect Ther.* 2012;10: 459–473. doi:10.1586/eri.12.25
9. Cobb R, Boeckh A. Moxidectin: a review of chemistry, pharmacokinetics and use in horses. *Parasit Vectors.* 2009;2 Suppl 2: S5. doi:10.1186/1756-3305-2-S2-S5
10. Guzzo CA, Furtek CI, Porras AG, Chen C, Tipping R, Kleinschmidt CM, et al. Safety, Tolerability, and Pharmacokinetics of Escalating High Doses of Ivermectin in Healthy Adult Subjects. *The Journal of Clinical Pharmacology.* 2002;42: 1122–1133. Available: <https://doi.org/10.1177/009127002237994>

11. Brandt W, Haupt VJ, Wessjohann LA. Chemoinformatic analysis of biologically active macrocycles. *Curr Top Med Chem*. 2010;10: 1361–1379. doi:10.2174/156802610792232060
12. Cuevas C, Francesch A. Development of Yondelis (trabectedin, ET-743). A semisynthetic process solves the supply problem. *Nat Prod Rep*. 2009;26: 322–337. doi:10.1039/b808331m
13. Gajdos C, Elias A. Trabectedin: safety and efficacy in the treatment of advanced sarcoma. *Clin Med Insights Oncol*. 2011;5: 35–43. doi:10.4137/CMO.S4907
14. Miller MA. Chemical database techniques in drug discovery. *Nat Rev Drug Discov*. 2002;1: 220–227. doi:10.1038/nrd745
15. Newman DJ. From natural products to drugs. *Phys Sci Rev*. 2019;4. doi:10.1515/psr-2018-0111
16. Medina-Franco JL. Discovery and Development of Lead Compounds from Natural Sources Using Computational Approaches. Evidence-Based Validation of Herbal Medicine. Elsevier; 2015. pp. 455–475. doi:10.1016/B978-0-12-800874-4.00021-0
17. Gu J, Gui Y, Chen L, Yuan G, Lu H-Z, Xu X. Use of natural products as chemical library for drug discovery and network pharmacology. *PLoS ONE*. 2013;8: e62839. doi:10.1371/journal.pone.0062839
18. Chen CY-C. TCM Database@Taiwan: the world's largest traditional Chinese medicine database for drug screening in silico. *PLoS ONE*. 2011;6: e15939. doi:10.1371/journal.pone.0015939
19. Pilon AC, Valli M, Dametto AC, Pinto MEF, Freire RT, Castro-Gamboa I, et al. NuBBEDB: an updated database to uncover chemical and biological information from Brazilian biodiversity. *Sci Rep*. 2017;7: 7215. doi:10.1038/s41598-017-07451-x
20. Ntie-Kang F, Zofou D, Babiaka SB, Meudom R, Scharfe M, Lifongo LL, et al. AfroDb: a select highly potent and diverse natural product library from African medicinal plants. *PLoS ONE*. 2013;8: e78085. doi:10.1371/journal.pone.0078085
21. Tung C-W. Public databases of plant natural products for computational drug discovery. *Curr Comput Aided Drug Des*. 2014;10: 191–196. doi:10.2174/1573409910666140414145934

22. Nguyen-Vo T-H, Le T, Pham D, Nguyen T, Le P, Nguyen A, et al. VIETHERB: A database for vietnamese herbal species. *J Chem Inf Model.* 2019;59: 1–9. doi:10.1021/acs.jcim.8b00399
23. Martinez-Mayorga K, Marmolejo-Valencia AF, Cortes-Guzman F, García-Ramos JC, Sánchez-Flores EI, Barroso-Flores J, et al. Toxicity Assessment of Structurally Relevant Natural Products from Mexican Plants with Antinociceptive Activity. *J Mex Chem Soc.* 2017;61. doi:10.29356/jmcs.v61i3.344
24. Leach AR, Gillet VJ. An introduction to chemoinformatics. Dordrecht: Springer Netherlands; 2007. doi:10.1007/978-1-4020-6291-9
25. Brown FK. Chemoinformatics: What is it and How does it Impact Drug Discovery. Elsevier; 1998. pp. 375–384. doi:10.1016/S0065-7743(08)61100-8
26. Hann M, Green R. Chemoinformatics — a new name for an old problem? *Curr Opin Chem Biol.* 1999;3: 379–383. doi:10.1016/S1367-5931(99)80057-X
27. Bajorath J. Understanding chemoinformatics: a unifying approach. *Drug Discovery Today.* 2004;9: 13–14. doi:10.1016/S1359-6446(04)02916-2
28. Jesús Naveja J, Saldívar-González FI, Sánchez-Cruz N, Medina-Franco JL. Cheminformatics approaches to study drug polypharmacology. Totowa, NJ: Humana Press; 2018. doi:10.1007/7653_2018_6
29. Weber WW. Epigenetics. *Comprehensive medicinal chemistry II.* Elsevier; 2007. pp. 251–278. doi:10.1016/B0-08-045044-X/00007-9
30. Tough DF, Tak PP, Tarakhovsky A, Prinjha RK. Epigenetic drug discovery: breaking through the immune barrier. *Nat Rev Drug Discov.* 2016;15: 835–853. doi:10.1038/nrd.2016.185
31. Medina-Franco JL. Epi-informatics: Discovery and development of small molecule epigenetic drugs and probes. Amsterdam: Elsevier/AP;
32. Castillo-Aguilera O, Depreux P, Halby L, Arimondo PB, Goossens L. DNA methylation targeting: the DNMT/HMT crosstalk challenge. *Biomolecules.* 2017;7. doi:10.3390/biom7010003

33. A. Salvador L, Luesch H. Discovery and mechanism of natural products as modulators of histone acetylation. *Curr Drug Targets*. 2012;13: 1029–1047. doi:10.2174/138945012802008973
34. Meng F, Wang C, Wan W, Lu W, Lu W, Luo C. Discovery and Development of Small Molecules Targeting Epigenetic Enzymes with Computational Methods. In: Medina-Franco JL, editor. *Epi-Informatics: Discovery and Development of Small Molecule Epigenetic Drugs and Probes*. London: Elsevier Inc.; 2016. pp. 75–112. doi:10.1016/B978-0-12-802808-7.00004-6
35. Zwergel C, Valente S, Mai A. DNA Methyltransferases Inhibitors from Natural Sources. *Curr Top Med Chem*. 2016;16: 680–696. doi:10.2174/1568026615666150825141505
36. Lascano S, Lopez M, Arimondo PB. Natural products and chemical biology tools: Alternatives to target epigenetic mechanisms in cancers. *Chem Rec*. 2018;18: 1854–1876. doi:10.1002/tcr.201800133
37. Chemical Computing Group INC. Molecular Operating Environment (MOE). Montreal, Quebec, Canada; disponible en: www.chemcomp.com/
38. Saldívar-González FI, Medina Franco JL, Lira Rocha A, Hernández Luis F. *Manual de Quimioinformática*. UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO; 2017.
39. Saldívar-González FI, Valli M, Andricopulo AD, da Silva Bolzani V, Medina-Franco JL. Chemical space and diversity of the nubbe database: A chemoinformatic characterization. *J Chem Inf Model*. 2019;59: 74–85. doi:10.1021/acs.jcim.8b00619
40. Sander T, Freyss J, von Korff M, Rufener C. DataWarrior: An open-source program for chemistry aware data visualization and analysis. *J Chem Inf Model*. 2015;55: 460–473. doi:10.1021/ci500588j
41. Lever J, Krzywinski M, Altman N. Points of Significance: Principal component analysis. *Nat Methods*. 2017;14: 641–642. doi:10.1038/nmeth.4346
42. Berthold MR, Cebron N, Dill F, Gabriel TR, Kötter T, Meinel T, et al. KNIME - the Konstanz information miner. *SIGKDD Explor Newsl*. 2009;11: 26. doi:10.1145/1656274.1656280

43. Bemis GW, Murcko MA. The properties of known drugs. 1. Molecular frameworks. *J Med Chem.* 1996;39: 2887–2893. doi:10.1021/jm9602928
44. González-Medina M, Medina-Franco JL. Platform for unified molecular analysis: PUMA. *J Chem Inf Model.* 2017;57: 1735–1740. doi:10.1021/acs.jcim.7b00253
45. Naveja JJ, Vogt M, Stumpfe D, Medina-Franco JL, Bajorath J. Systematic extraction of analogue series from large compound collections using a new computational compound–core relationship method. *ACS Omega.* 2019;4: 1027–1032. doi:10.1021/acsomega.8b03390
46. Osolodkin DI, Radchenko EV, Orlov AA, Voronkov AE, Palyulin VA, Zefirov NS. Progress in visual representations of chemical space. *Expert Opin Drug Discov.* 2015;10: 959–973. doi:10.1517/17460441.2015.1060216
47. van der Maaten L, Hinton G. Visualizing Data using t-SNE. *Journal of Machine Learning Research.* 2008;9: 2579–2605.
48. Sheridan RP, Kearsley SK. Why do we need so many chemical similarity search methods? *Drug Discov Today.* 2002;7: 903–911. doi:10.1016/S1359-6446(02)02411-X
49. Naveja JJ, Rico-Hidalgo MP, Medina-Franco JL. Analysis of a large food chemical database: chemical space, diversity, and complexity [version 1; peer review: 2 approved, 1 approved with reservations]. *F1000Res.* 2018;7: 993. doi:10.12688/f1000research.15440.1
50. Medina-Franco JL, Martínez-Mayorga K, Bender A, Scior T. Scaffold diversity analysis of compound data sets using an entropy-based measure. *QSAR Comb Sci.* 2009;28: 1551–1560. doi:10.1002/qsar.200960069
51. Naveja J, Gonzalez-Medina M, Ravindranath AC, Campillos M, Medina-Franco JL. D-DATABASE: An integrated target-molecule database ready for cheminformatic analysis. 2018; submitted.
52. Chemotargets SL. Chemotargets CLARITY. 2017. Disponible en: www.chemotargets.com
53. Ellis CR, Racz R, Kruhlak NL, Kim MT, Hawkins EG, Strauss DG, et al. Assessing the structural and pharmacological similarity of newly identified drugs of abuse to controlled substances using public health assessment via structural evaluation. *Clin Pharmacol Ther.* 2019; doi:10.1002/cpt.1418

54. González-Medina M, Prieto-Martínez FD, Owen JR, Medina-Franco JL. Consensus Diversity Plots: a global diversity analysis of chemical libraries. *J Cheminform.* 2016;8: 63. doi:10.1186/s13321-016-0176-9
55. Nonell-Canals A, Mestres J. In silico target profiling of one billion molecules. *Mol Inform.* 2011;30: 405–409. doi:10.1002/minf.201100018
56. Naveja JJ, Oviedo-Osornio CI, Trujillo-Minero NN, Medina-Franco JL. Chemoinformatics: a perspective from an academic setting in Latin America. *Mol Divers.* 2018;22: 247–258. doi:10.1007/s11030-017-9802-3
57. López-López E, Prieto-Martínez FD, Medina-Franco JL. Activity landscape and molecular modeling to explore the SAR of dual epigenetic inhibitors: A focus on G9a and DNMT1. *Molecules.* 2018;23. doi:10.3390/molecules23123282
58. Garella D, Atlante S, Borretto E, Cocco M, Giorgis M, Costale A, et al. Design and synthesis of N-benzoyl amino acid derivatives as DNA methylation inhibitors. *Chem Biol Drug Des.* 2016;88: 664–676. doi:10.1111/cbdd.12794

ANEXOS

A1. Ejemplo de contenido de la base de datos BIOFACQUIM.

ID	Name	SMILES	Year	Kingdom	Genus	Specie	DOI
FQNP394	Orcinol	Oc1cc(O)cc(C)c1	2000	Fungus	Parmotrema	Parmotrema_tinctorum	10.1021/np0001326
FQNP341	Rubrofusarin_B	O(C)c1c2c(c(O)c3C(=O)C-	2000	Fungus	Guanomyces	Guanomyces_polythrix	10.1021/np990534h
FQNP349	Cyclolaudenyl_Acetate	O=C(O[C@@H]1C(C)(C)C	2001	Plant	Tillandsia	Tillandsia_fasciculata	10.1021/np0100744
FQNP288	Xanthorrhizol	Oc1c(C)ccc([C@@H](CC/C	2001	Plant	Iostephane	Iostephane_heterophylla	10.1080/10575630108041265
FQNP418	p-hydroxybenzoic_acid	O=C(O)Cc1ccc(O)cc1	2001	Fungus	Guanomyces	Guanomyces_polythrix	10.1016/S0031-9422(01)00278-3
FQNP269	Friedelin	O=C1[C@H](C)[C@]2(C)[C	2002	Plant	Hippocratea	Hippocratea_excelsa	10.1016/S0378-8741(01)00414-7
FQNP356	(+)-ciskhellactone	O=C1Oc2c3[C@@H](O)[C	2002	Plant	Prionosciadium	Prionosciadium_watsoni	10.1021/np010448t
FQNP281	Nidemin	[C@H](CCC(C(=C)C)(C)C	2002	Plant	Scaphyglottis	Scaphyglottis_livida	10.1080/10575630290019967
FQNP400	Gymnopusin	O(C)c1c(OC)c2c3c(c(OC)c	2002	Plant	Maxillaria	Maxillaria_densa	10.1016/S0031-9422(02)00220-0
FQNP302	Rubrofusarin_B	O(C)c1c2c(O)c3C(=O)C=C	2003	Fungus	Guanomyces	Guanomyces_polythrix	10.1021/jf030115x
FQNP388	2'-2"-dimethoxysesamin	O(C)c1c([C@H]2OC[C@@	2003	Plant	Leucophyllum	Leucophyllum_ambiguum	10.1021/np020346i
FQNP439	benzyl_2_6-dimethoxybenzoat	O=C(OCc1ccccc1)c1c(OC	2005	Plant	Brickellia	Brickellia_veronicifolia	10.1055/s-2005-864097

ID	Journal	Site	State	IC50_1	Activity_1	Bioactivity
FQNP394	Journal_of_Natural_Products	Los_Tuxtlas	Veracruz	1.00E-03	Seeding_Growth_A_hypochondriacus	Phytotoxic
FQNP341	Journal_of_Natural_Products	Tepoztlan	Morelos	1.30E-05	Seeding_Growth_A_hypochondriacus	Phytotoxic
FQNP349	Journal_of_Natural_Products	Quintana_Roo	El_Eden			
FQNP288	Natural_Products_Letters	Ciudad_de_Mexico	Mercado_de_Sonora			Antimicrobial
FQNP418	Phytochemistry	Morelos	Tepozotlan	1.70E-05	Phytogrowth_inhibition_of_A_hypochondriac	Phytotoxic
FQNP269	Journal_of_Ethnopharmacology	Guerrero	Costa_Grande			Gastroprotective
FQNP356	Journal_of_Natural_Products	San_Luis_Potosi	Sierra_Alvarez	3.61E+04	Phytogrowth_inhibition_of_A_hypochondriac	Phytotoxic
FQNP281	Natural_Products_Letters	Veracruz	Catemaco			
FQNP400	Phytochemistry	Veracruz	Catemaco	1.00E-04	Phytogrowth_inhibition_of_L_pausicostata	Phytotoxic
FQNP302	Journal_of_Agricultural_and_Food_Chemistry	Morelos	Ocotlitlan	4.70E-06	CaM_Dependent_PDE	CaM_Inhibitor
FQNP388	Journal_of_Natural_Products	San_Luis_Potosi	San_Luis_Potosi	5.60E-06	Calmodulin_inhibition	CaM_Inhibitor
FQNP439	Planta_Medica	Estado_de_Mexico	Aculco	1.84E-06	Inhibition_of_the_spontaneous_contractions	Relaxant_action

* Base de datos completa , disponible en: <https://biofacquim.herokuapp.com>

A2. Resultados de Chemotargets para compuestos con actividad conocida.

Se presenta para todas las familias:

- I. Gráfico de número de moléculas por diana biológica.
- II. Gráfico de actividad por compuesto en cada diana biológica.

A.1.1 Citocromo (CP)

A.1.2 Enzimas (EC)

A.1.3 Enzimas y Transportadores (EC, TC)

A.1.4 Receptores acoplados a proteínas G (GR)

A.1.5 Cinasas (KC)

A.1.6 Receptores nucleares (NR)

A.1.7 Otras familias (OF)

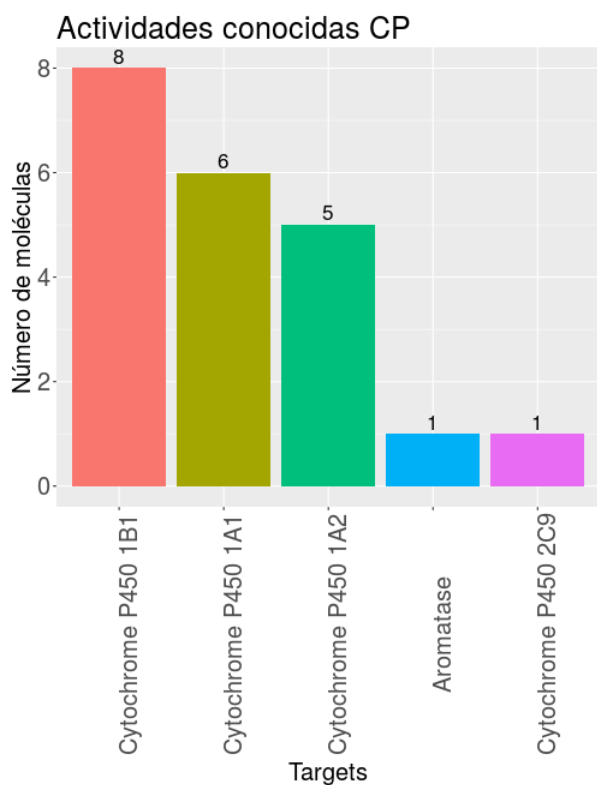
A.1.8 Transportadores (TC)

A.1.9 Transportadores y canales iónicos (TC, IC)

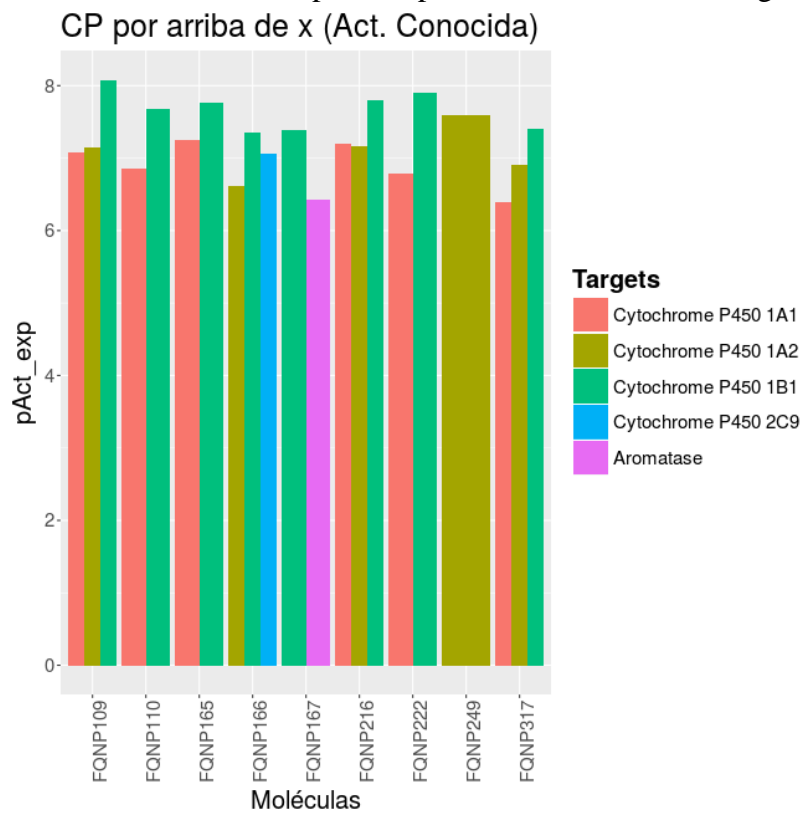
A.1.10 Sin clasificar (UC)

A.2.1 Citocromo (CP)

I. Gráfico de número de moléculas por diana biológica de la familia CP.

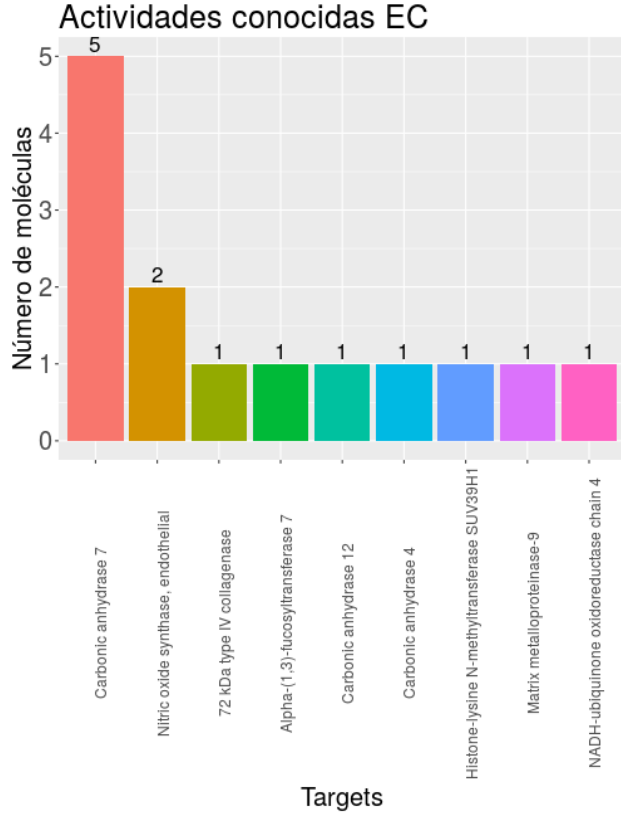


II. Gráfico de actividad por compuesto en cada diana biológica de la familia CP.

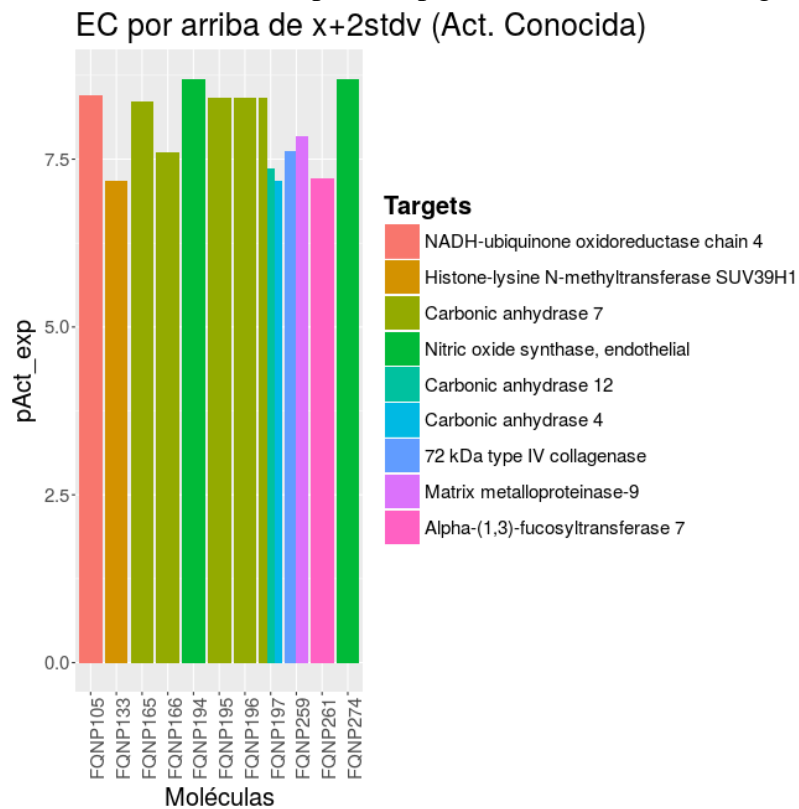


A.2.2 Enzimas (EC)

I. Gráfico de número de moléculas por diana biológica de la familia EC.

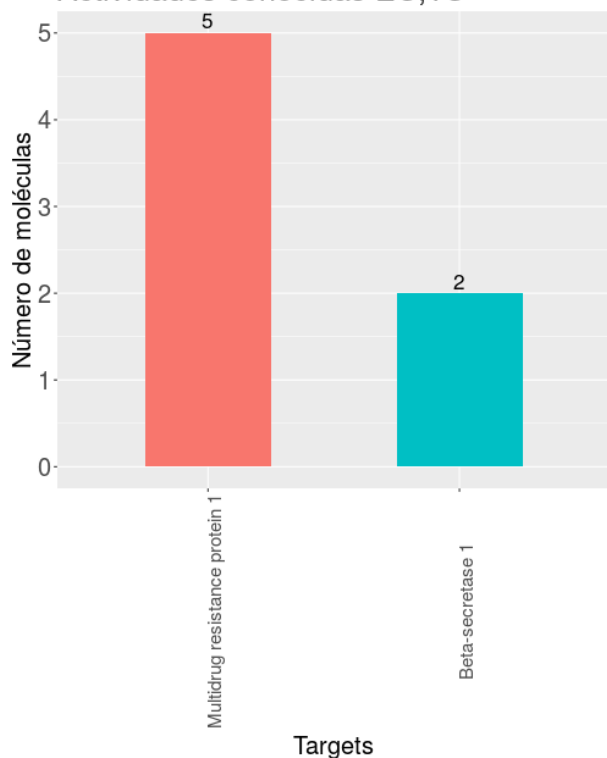


II. Gráfico de actividad por compuesto en cada diana biológica de la familia EC.

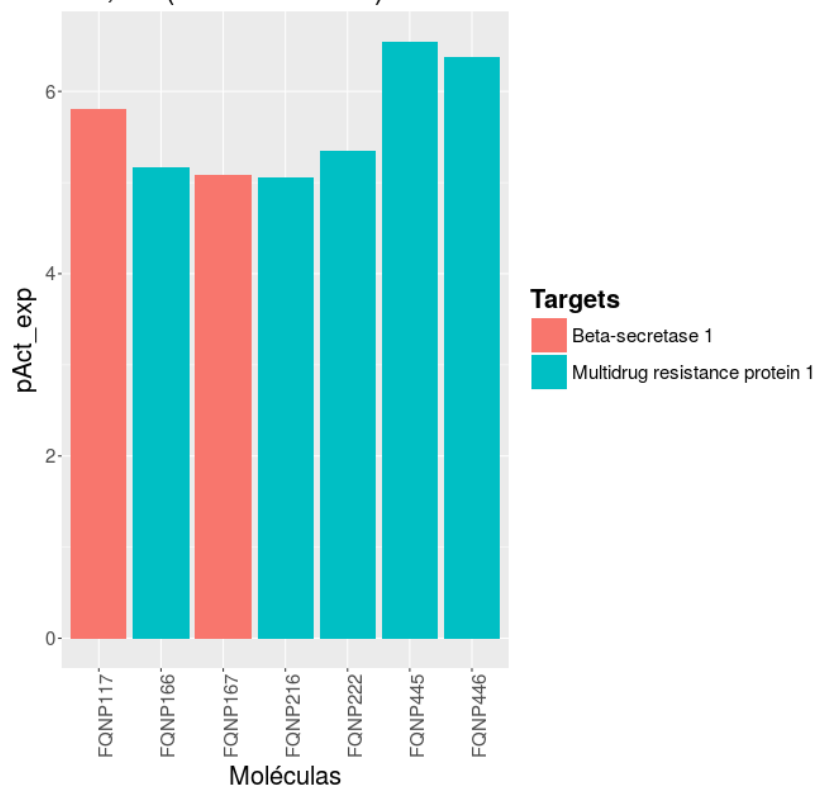


A.2.3 Enzimas y Transportadores (EC, TC)

I. Gráfico de número de moléculas por diana biológica de la familia EC, TC.
Actividades conocidas EC,TC

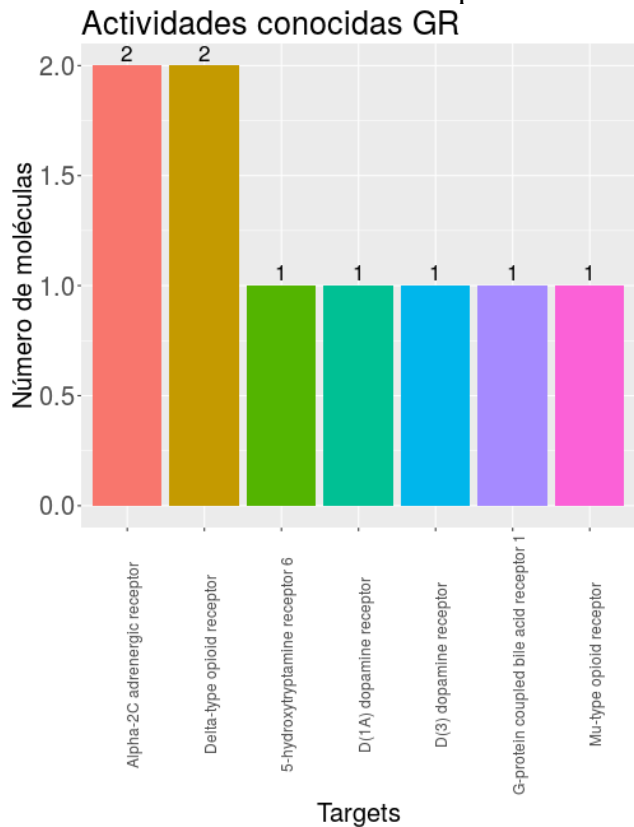


II. Gráfico de actividad por compuesto en cada diana biológica de la familia EC, TC.
EC,TC (Act. Conocida)



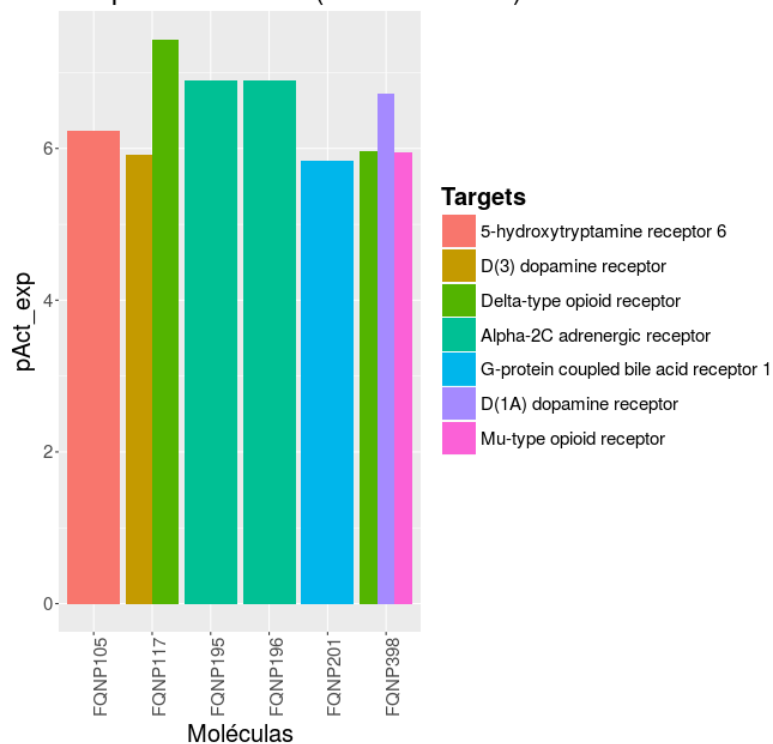
A.2.4 Receptores acoplados a proteínas G (GR)

I. Gráfico de número de moléculas por diana biológica de la familia GR.



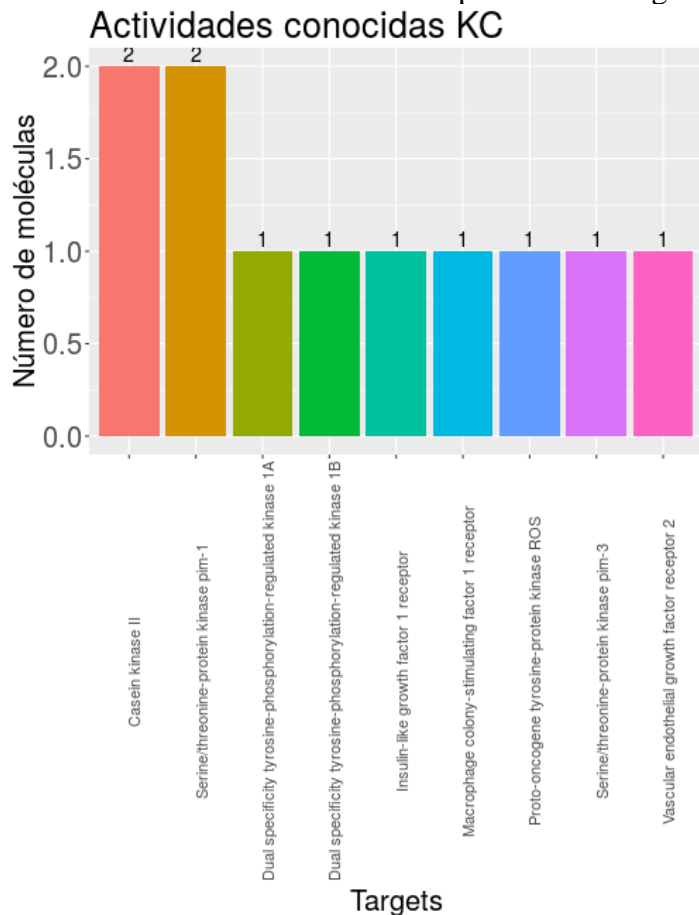
II. Gráfico de actividad por compuesto en cada diana biológica de la familia GR.

GR por arriba de x (Act. Conocida)

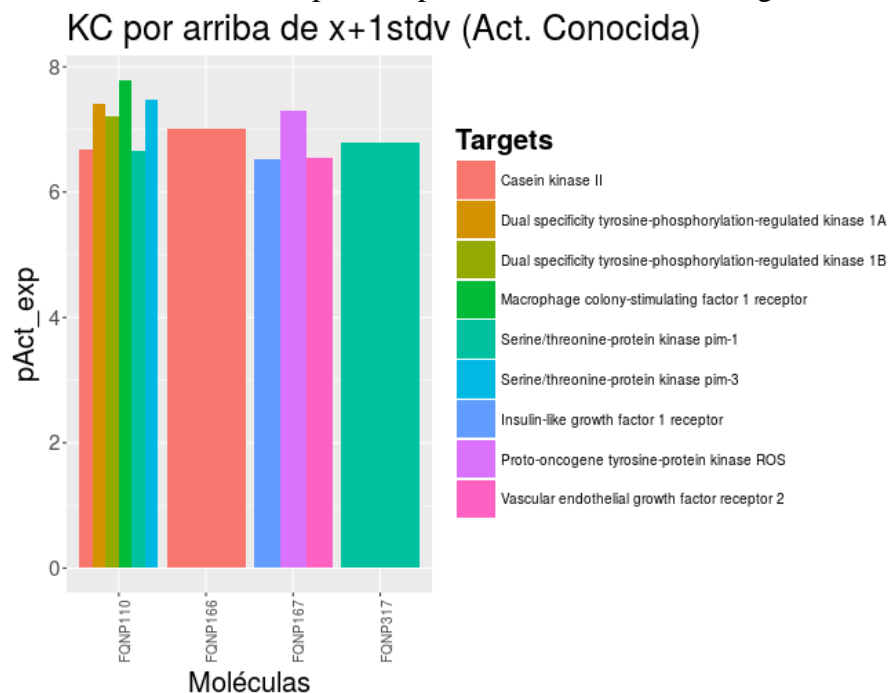


A.2.5 Cinasas (KC)

I. Gráfico de número de moléculas por diana biológica de la familia KC.

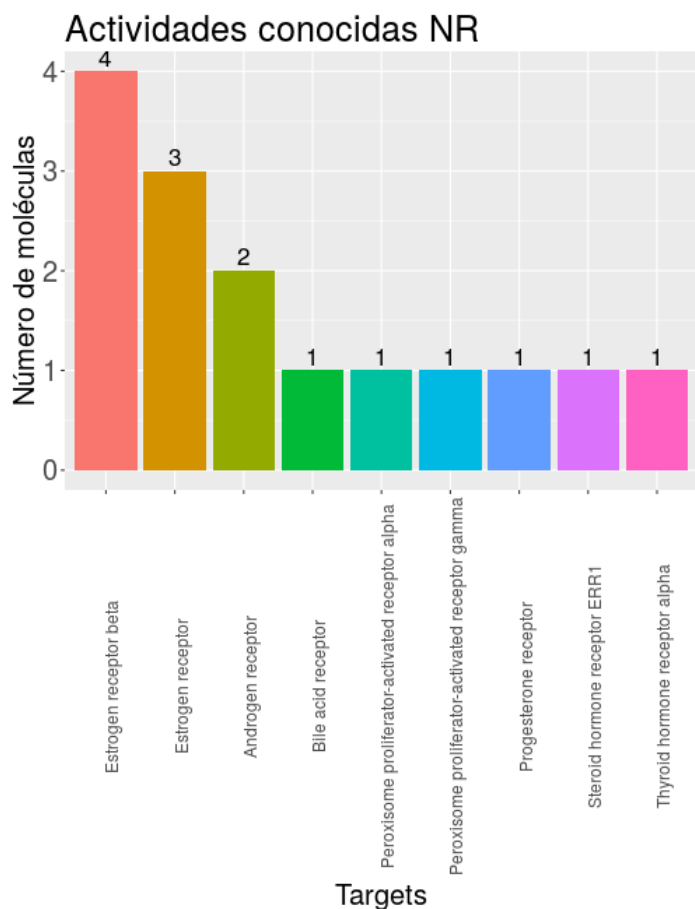


II. Gráfico de actividad por compuesto en cada diana biológica de la familia KC.

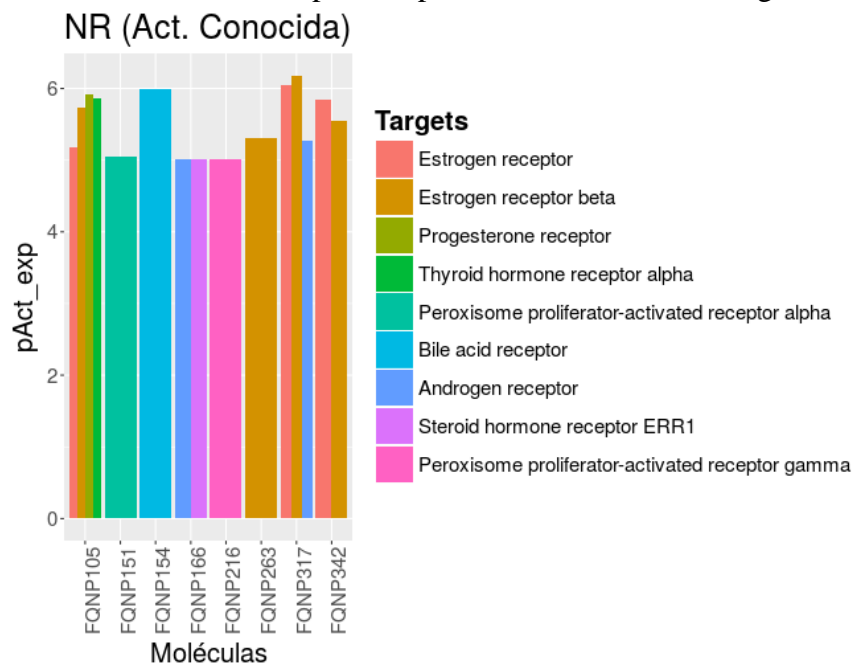


A.2.6 Receptores nucleares (NR)

I. Gráfico de número de moléculas por diana biológica de la familia NR.

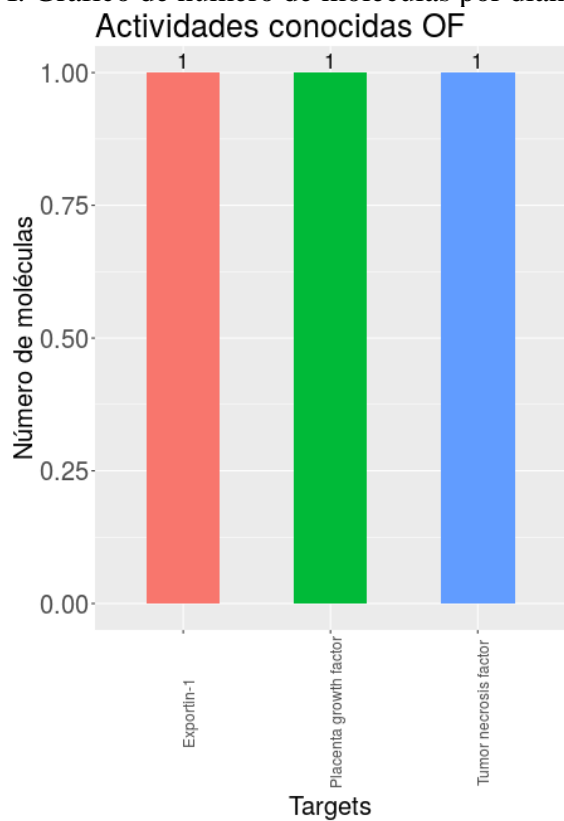


II. Gráfico de actividad por compuesto en cada diana biológica de la familia NR.

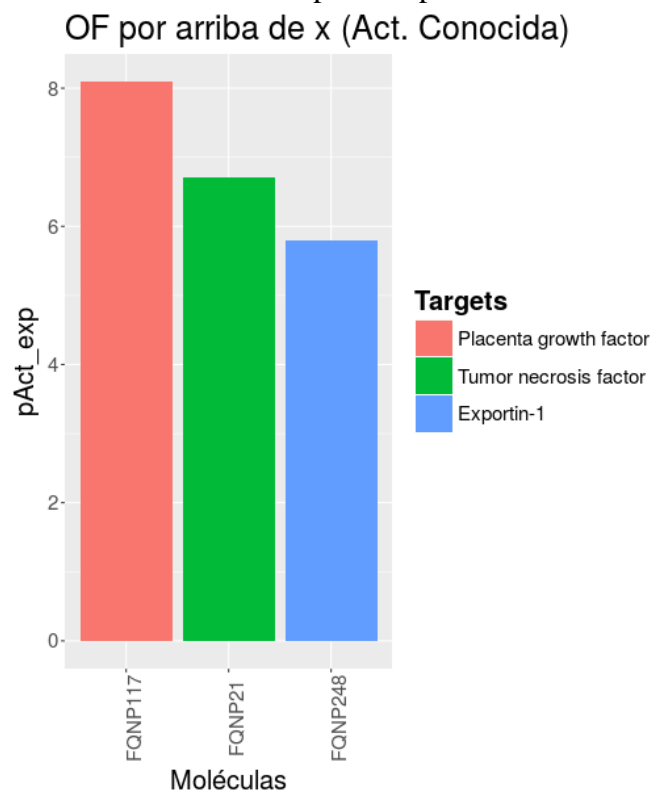


A.2.7 Otras familias (OF)

I. Gráfico de número de moléculas por diana biológica de la familia OF.

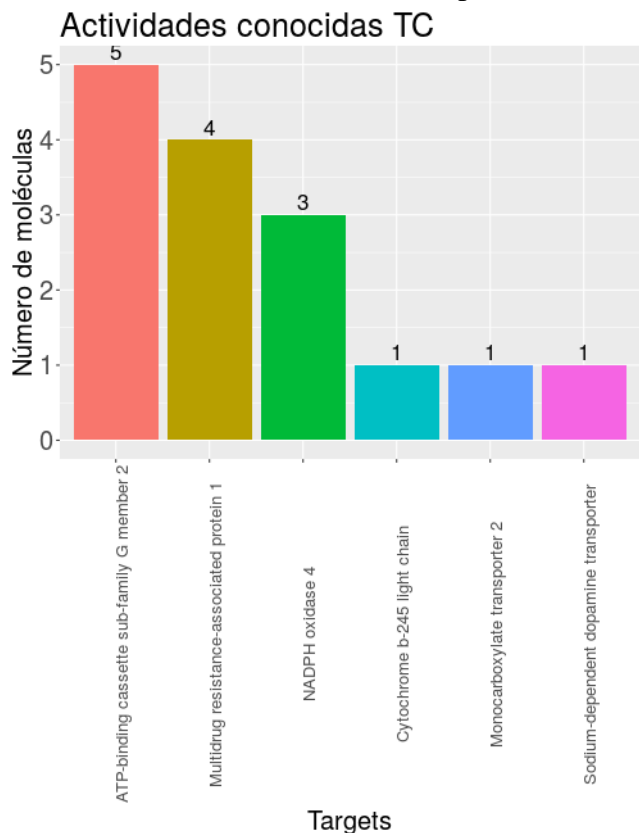


II. Gráfico de actividad por compuesto en cada diana biológica de la familia OF.

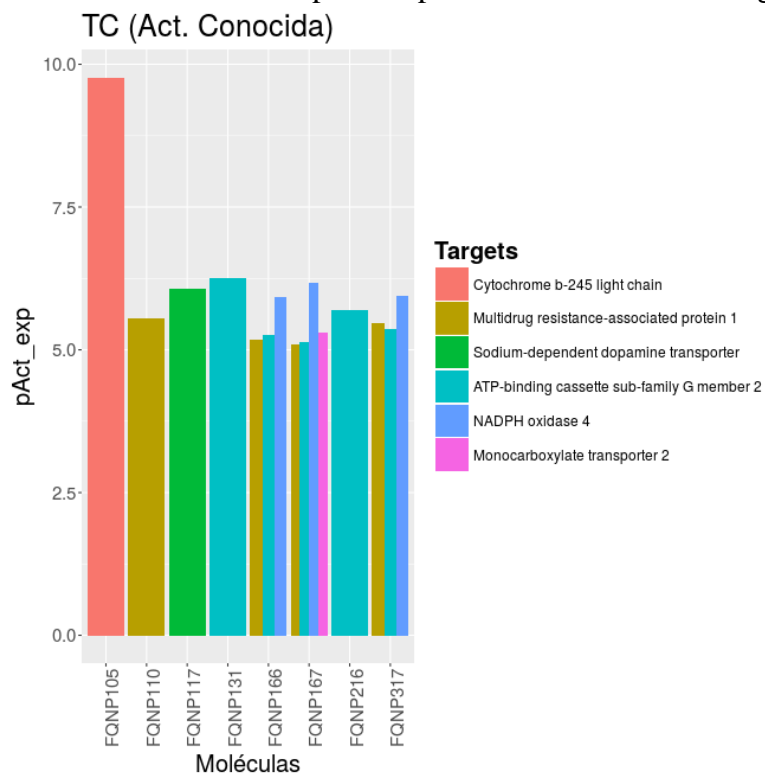


A.2.8 Transportadores (TC)

I. Gráfico de número de moléculas por diana biológica de la familia TC.

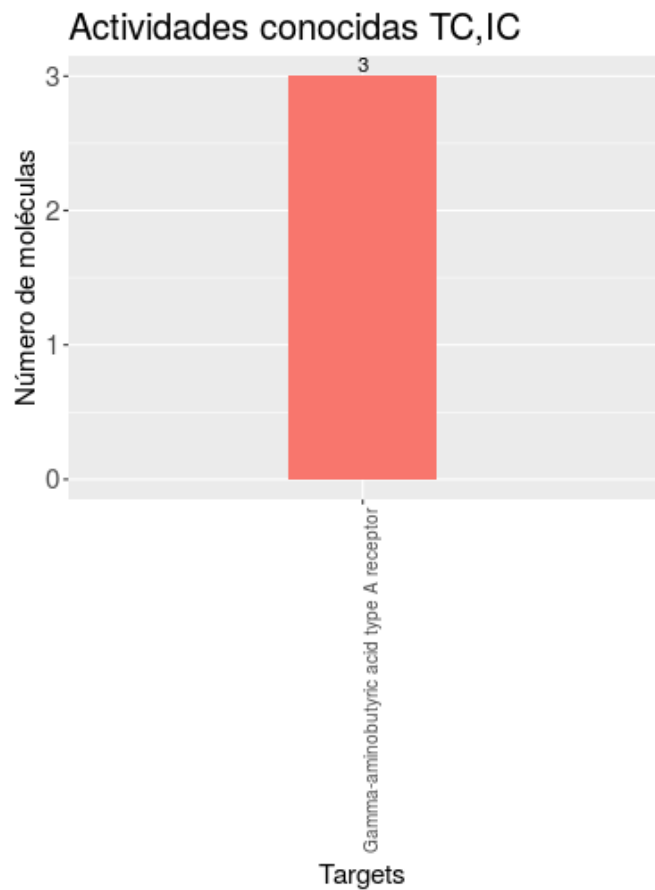


II. Gráfico de actividad por compuesto en cada diana biológica de la familia TC.

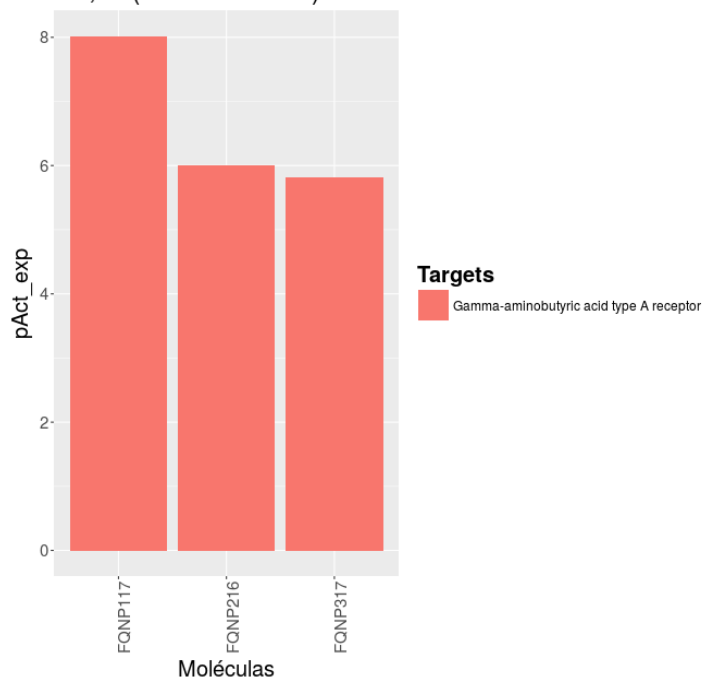


A.2.9 Transportadores y canales iónicos (TC, IC)

I. Gráfico de número de moléculas por diana biológica de la familia TC, IC.

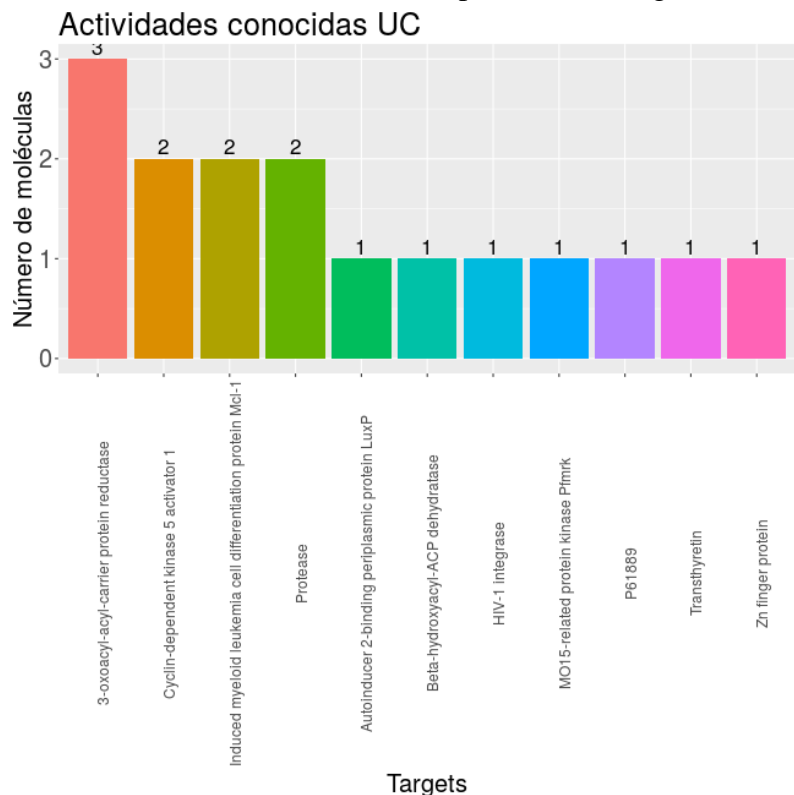


II. Gráfico de actividad por compuesto en cada diana biológica de la familia TC, IC.
TC,IC (Act. Conocida)

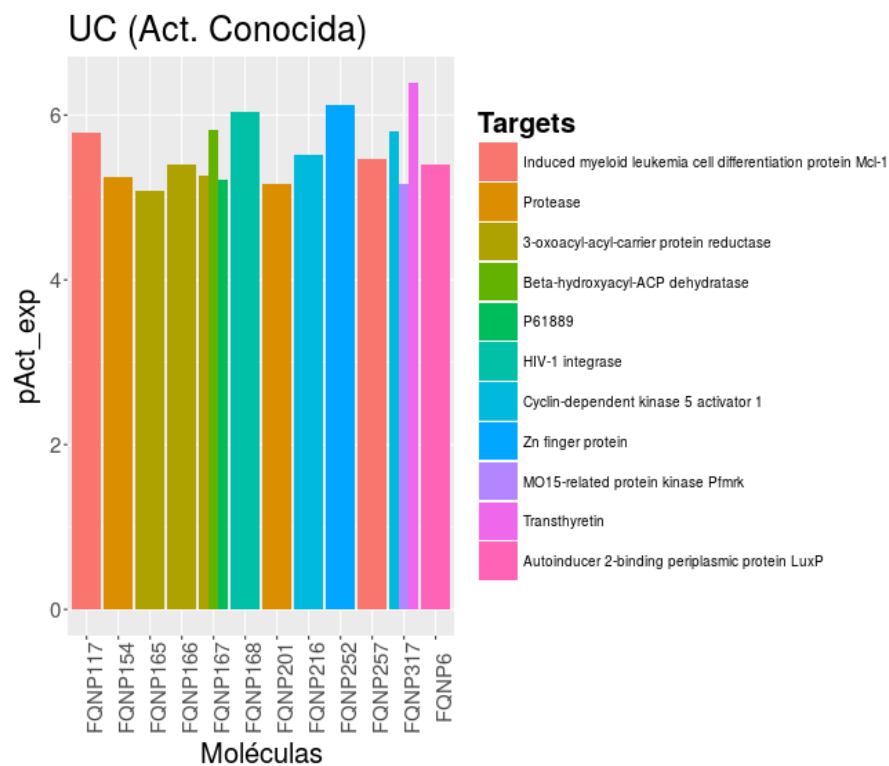


A.2.10 Sin clasificar (UC)

I. Gráfico de número de moléculas por diana biológica de la familia UC.



II. Gráfico de actividad por compuesto en cada diana biológica de la familia UC.



A3. Resultados de Chemotargets para compuestos con actividad predicha.

Se presenta para todas las familias:

- I. Gráfico de número de moléculas por diana biológica.
- II. Gráfico de actividad por compuesto en cada diana biológica.

A.2.1 Citocromo (CP)

A.2.2 Enzimas (EC)

A.2.3 Enzimas y Transportadores (EC, TC)

A.2.4 Receptores acoplados a proteínas G (GR)

A.2.5 Cinasas (KC)

A.2.6 Receptores nucleares (NR)

A.2.7 Otras familias (OF)

A.2.8 Transportadores (TC)

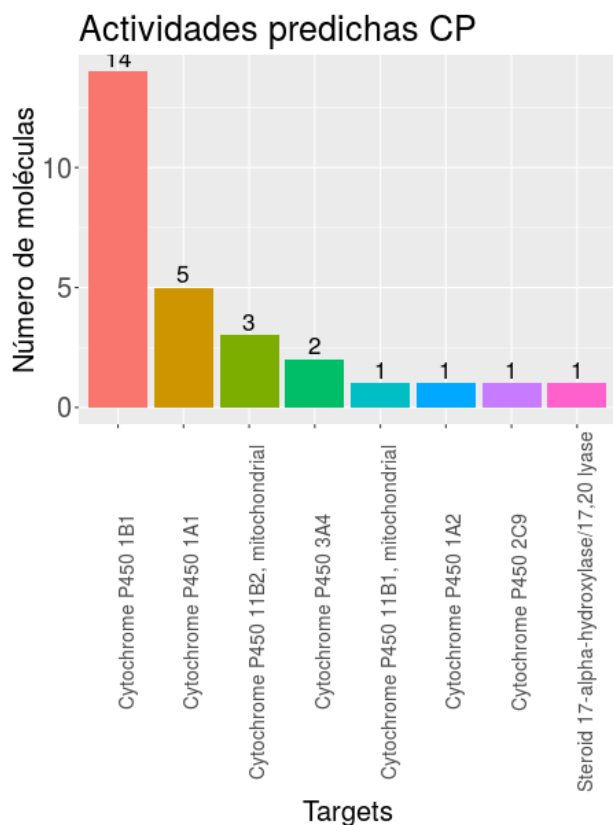
A.2.9 Transportadores y canales iónicos (TC, IC)

A.2.10 Sin clasificar (UC)

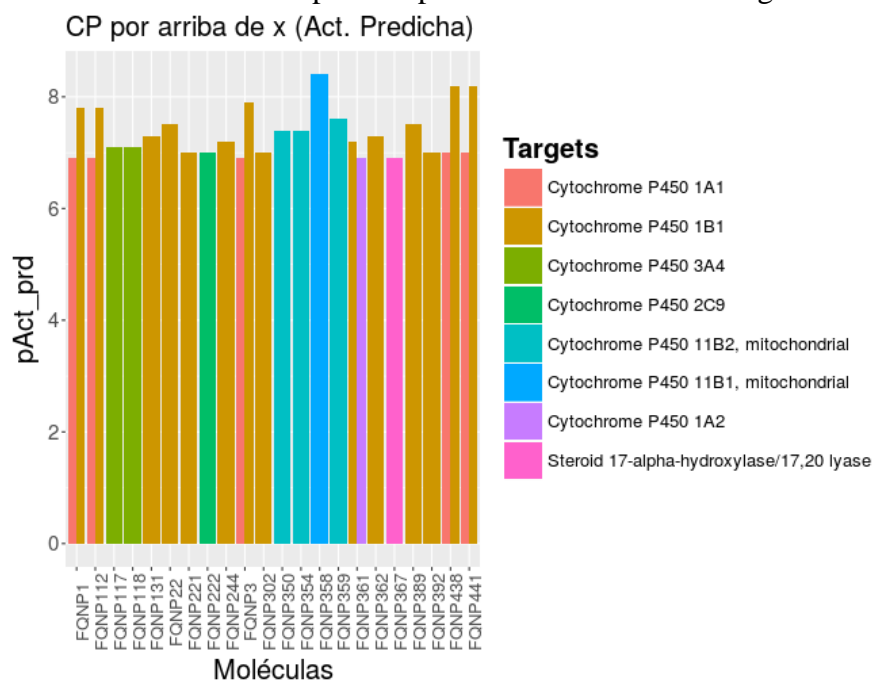
A.2.11 Canales iónicos (IC)

A.3.1 Citocromo (CP)

I. Gráfico de número de moléculas por diana biológica de la familia CP.

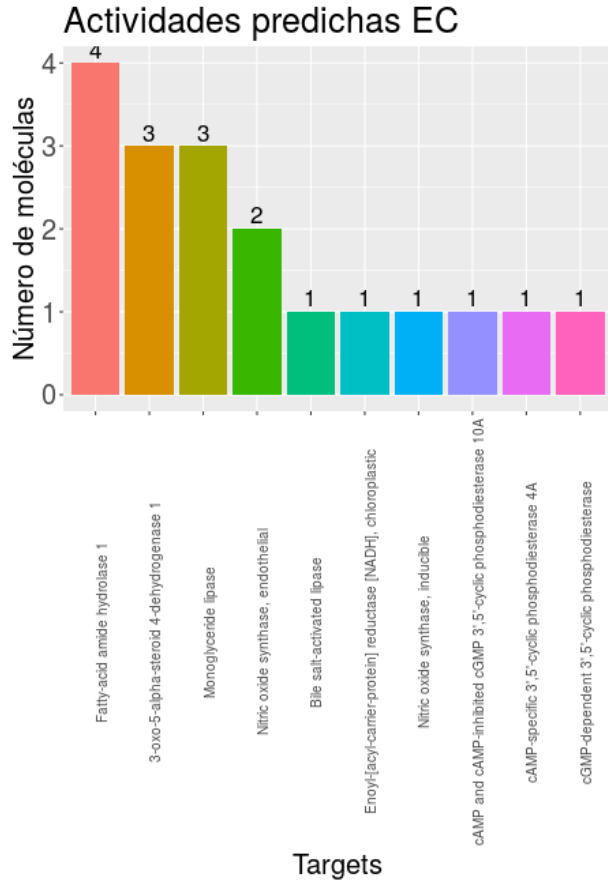


II. Gráfico de actividad por compuesto en cada diana biológica de la familia CP.

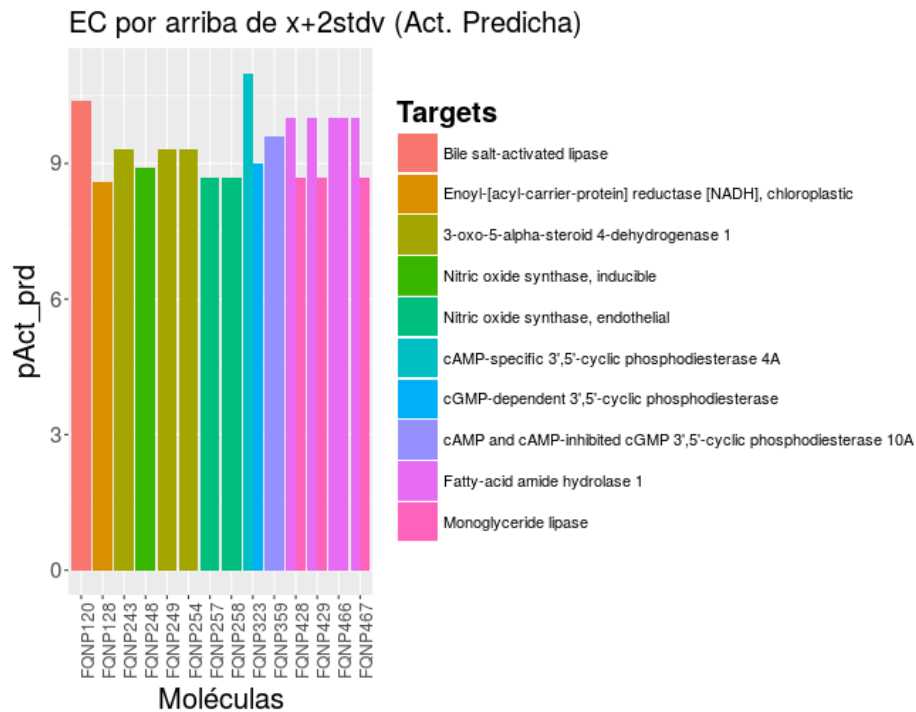


A.3.2 Enzimas (EC)

I. Gráfico de número de moléculas por diana biológica de la familia EC.

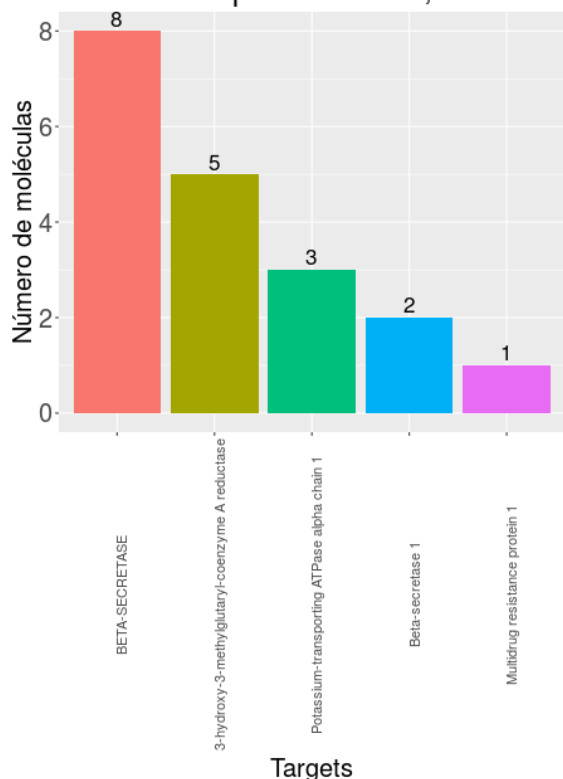


II. Gráfico de actividad por compuesto en cada diana biológica de la familia EC.

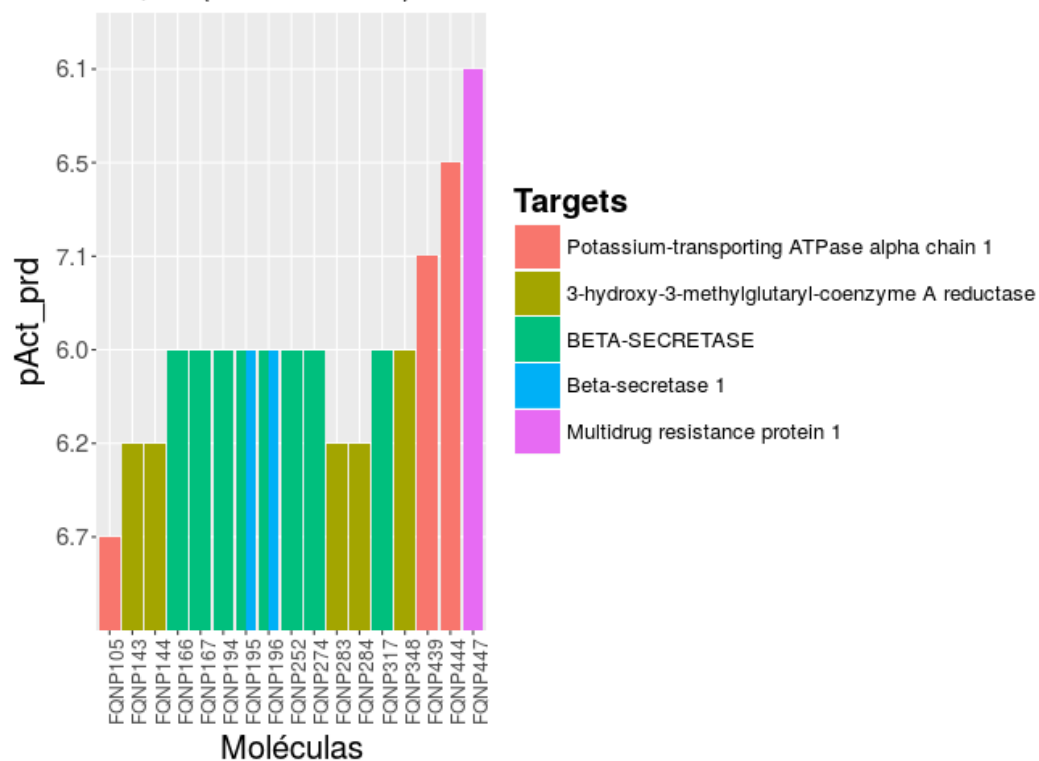


A.3.3 Enzimas y Transportadores (EC, TC)

I. Gráfico de número de moléculas por diana biológica de la familia EC, TC. Actividades predichas EC, TC

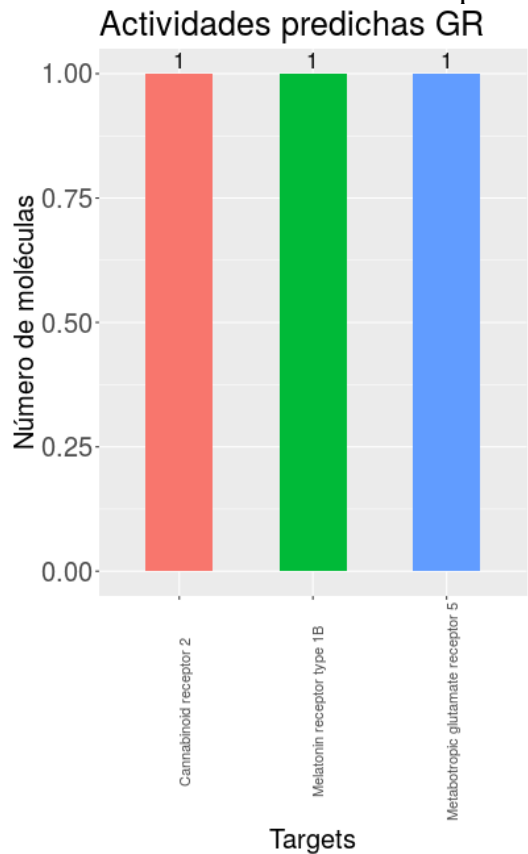


II. Gráfico de actividad por compuesto en cada diana biológica de la familia EC, TC. EC,TC (Act. Predicha)

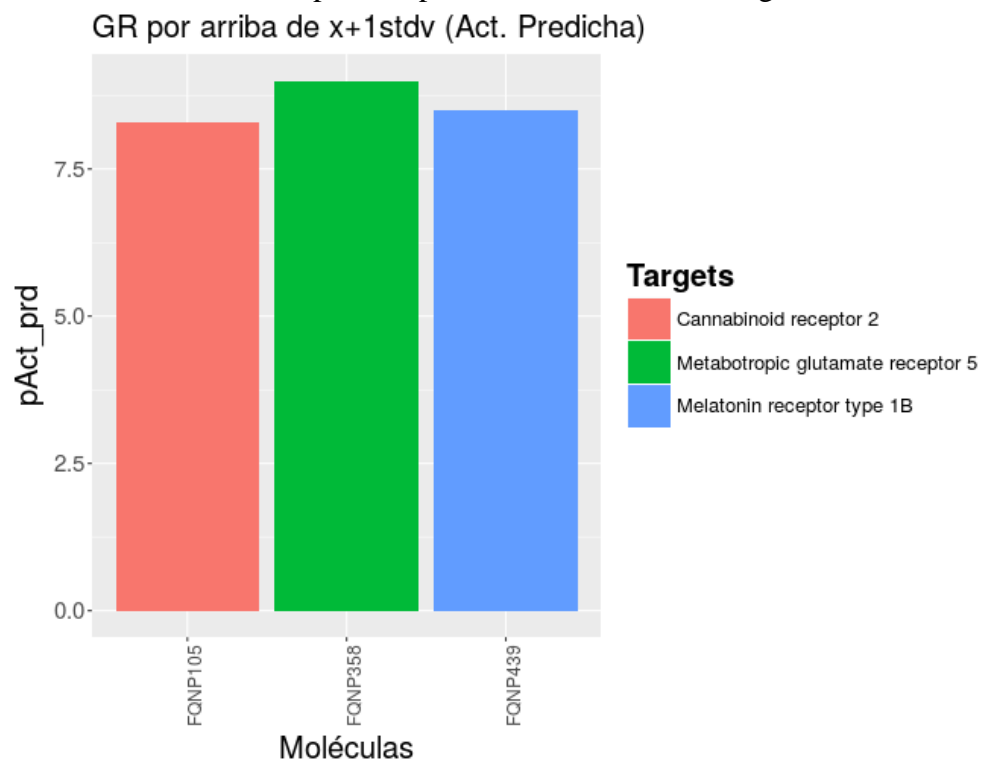


A.3.4 Receptores acoplados a proteínas G (GR)

I. Gráfico de número de moléculas por diana biológica de la familia GR.

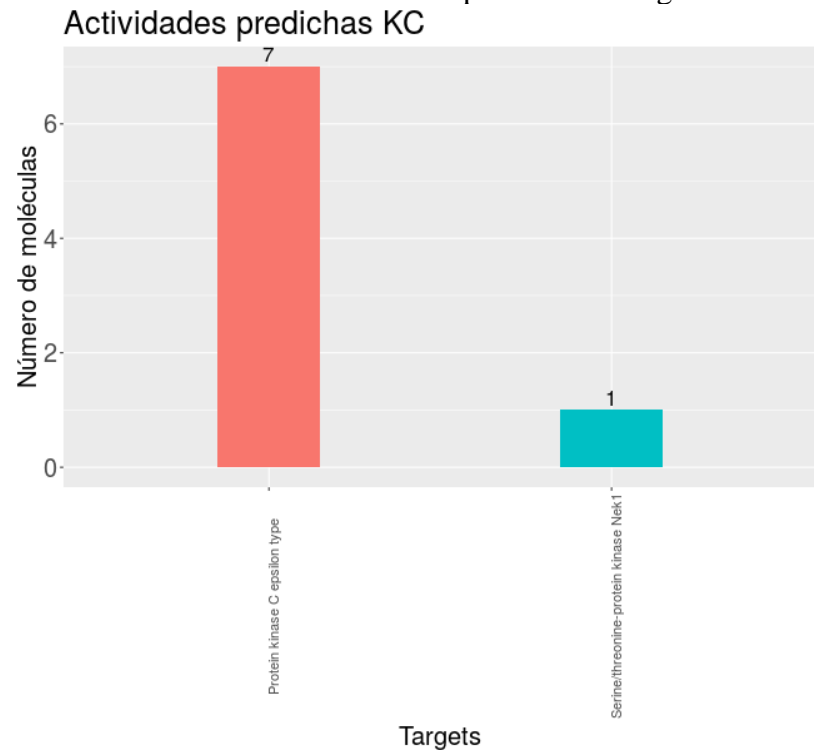


II. Gráfico de actividad por compuesto en cada diana biológica de la familia GR.

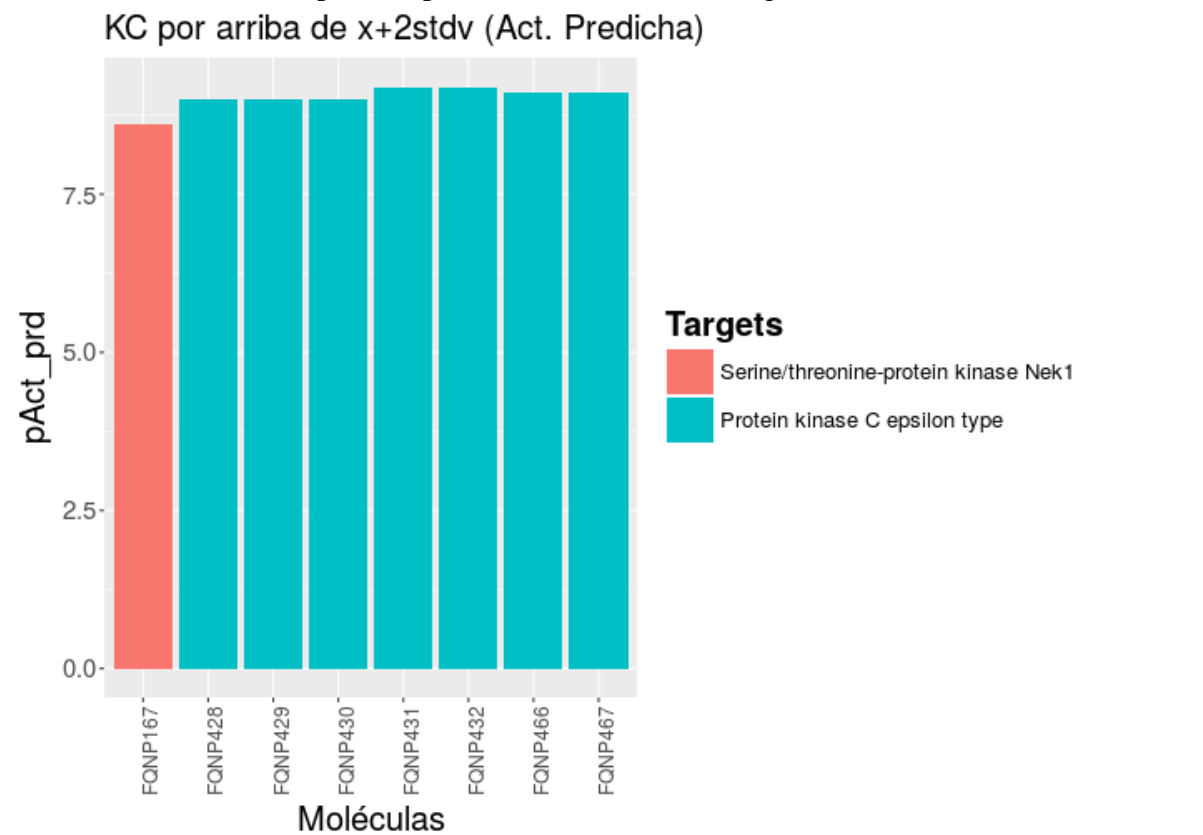


A.3.5 Cinasas (KC)

I. Gráfico de número de moléculas por diana biológica de la familia KC.

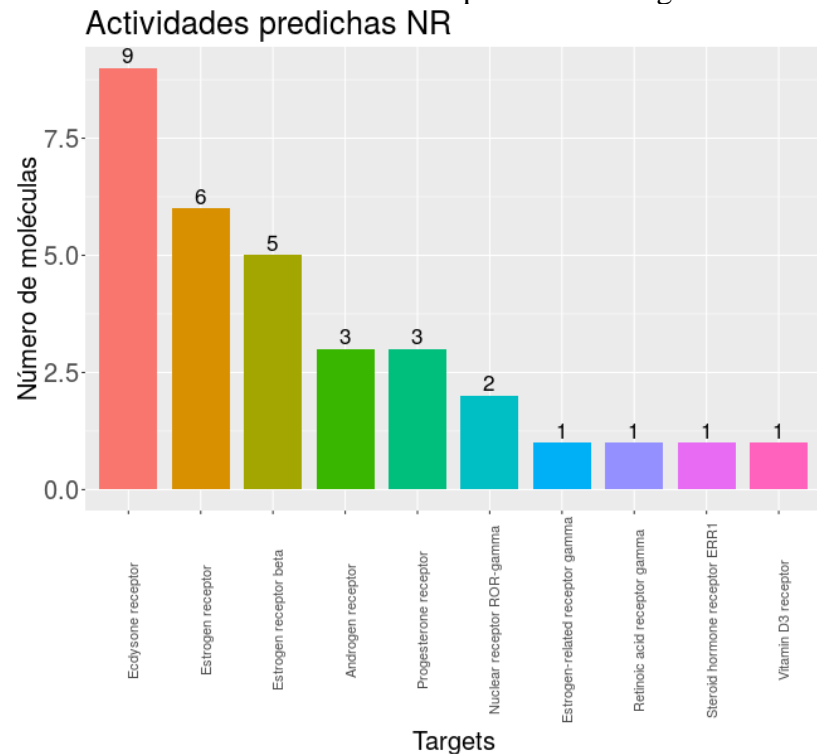


II. Gráfico de actividad por compuesto en cada diana biológica de la familia KC.

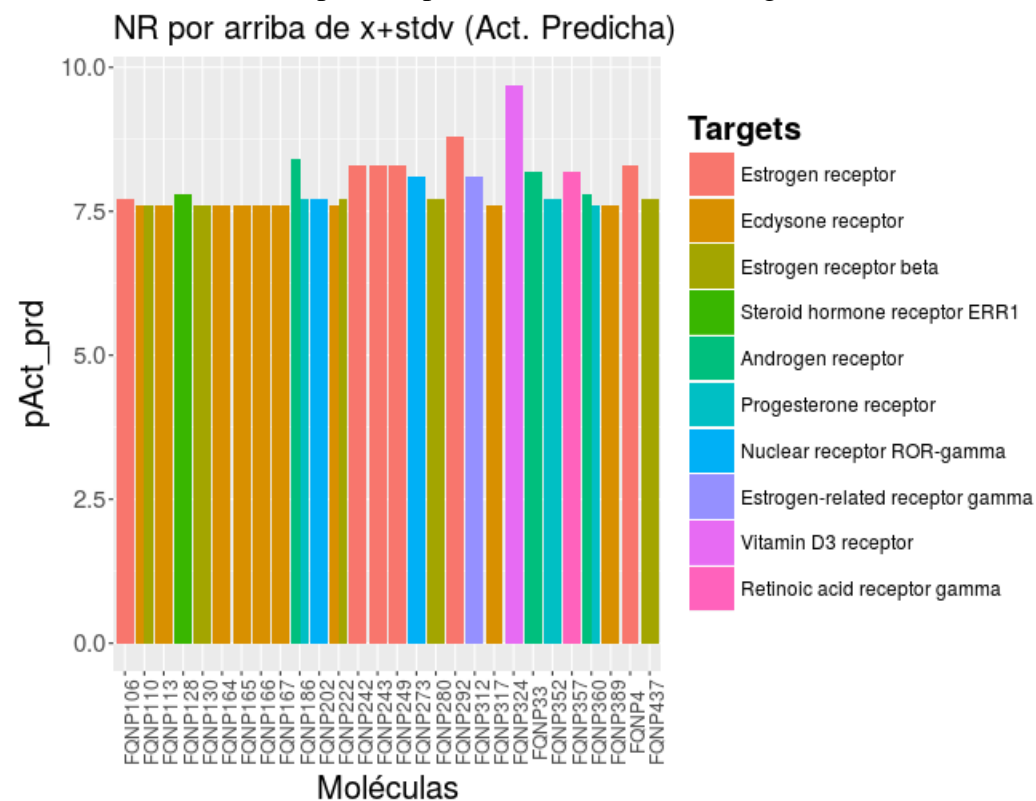


A.3.6 Receptores nucleares (NR)

I. Gráfico de número de moléculas por diana biológica de la familia NR.

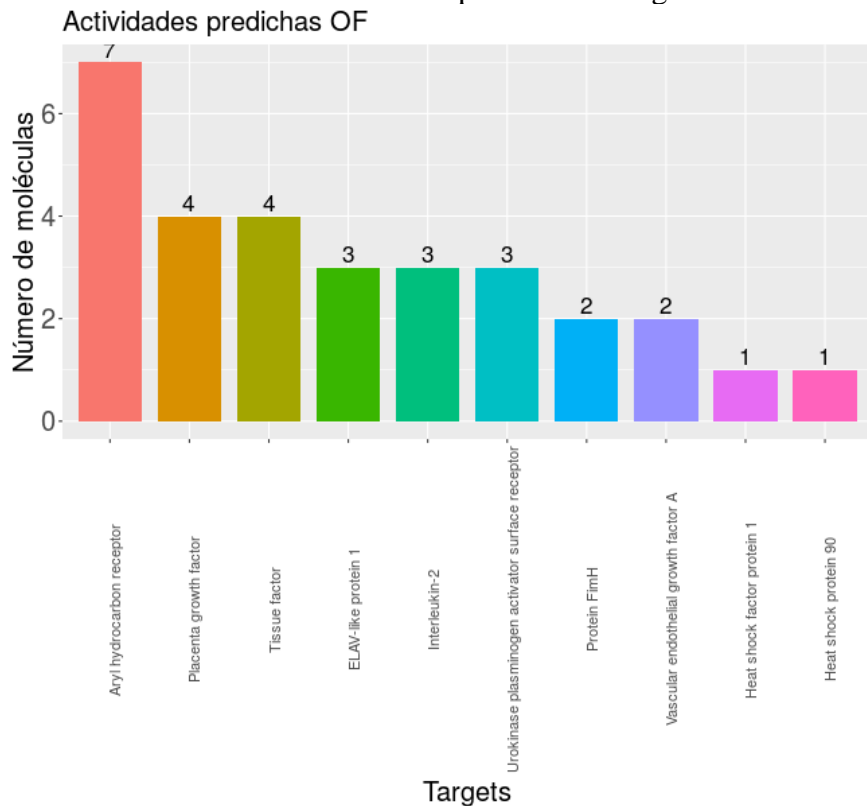


II. Gráfico de actividad por compuesto en cada diana biológica de la familia NR.

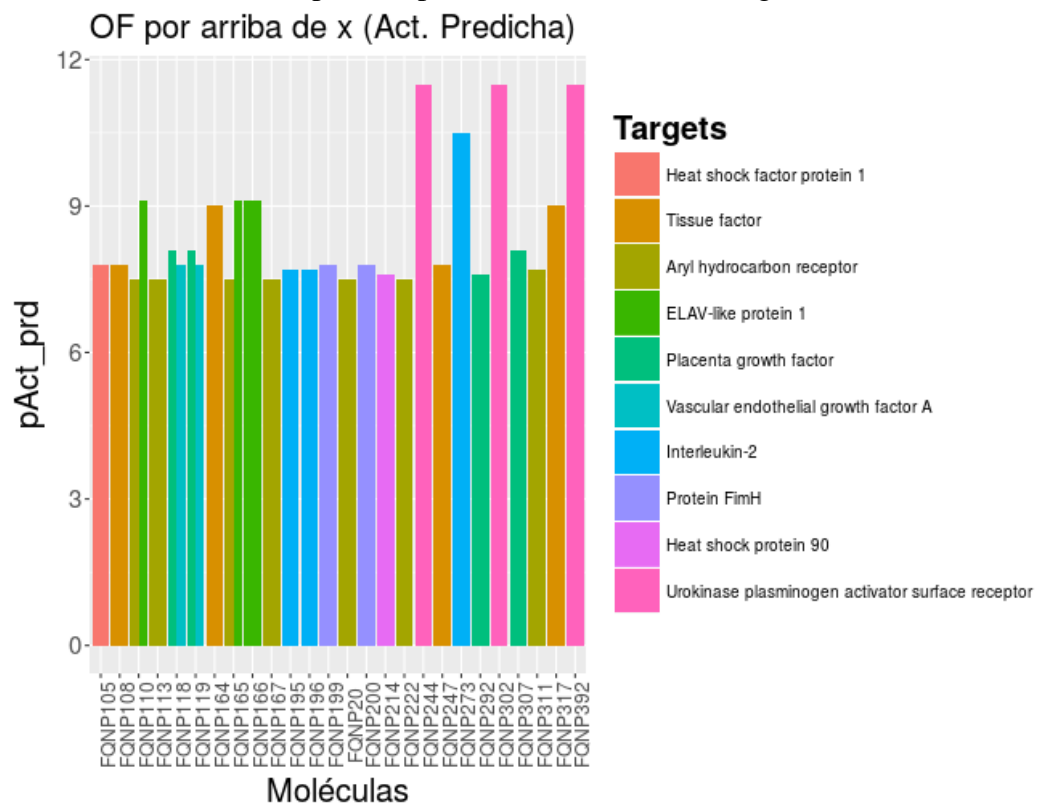


A.3.7 Otras familias (OF)

I. Gráfico de número de moléculas por diana biológica de la familia OF.

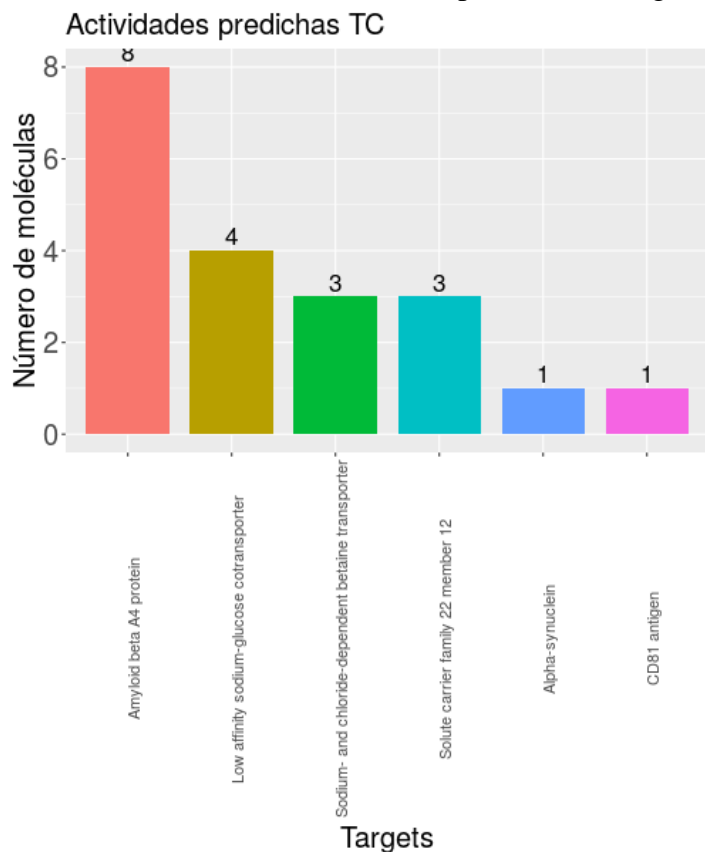


II. Gráfico de actividad por compuesto en cada diana biológica de la familia OF.

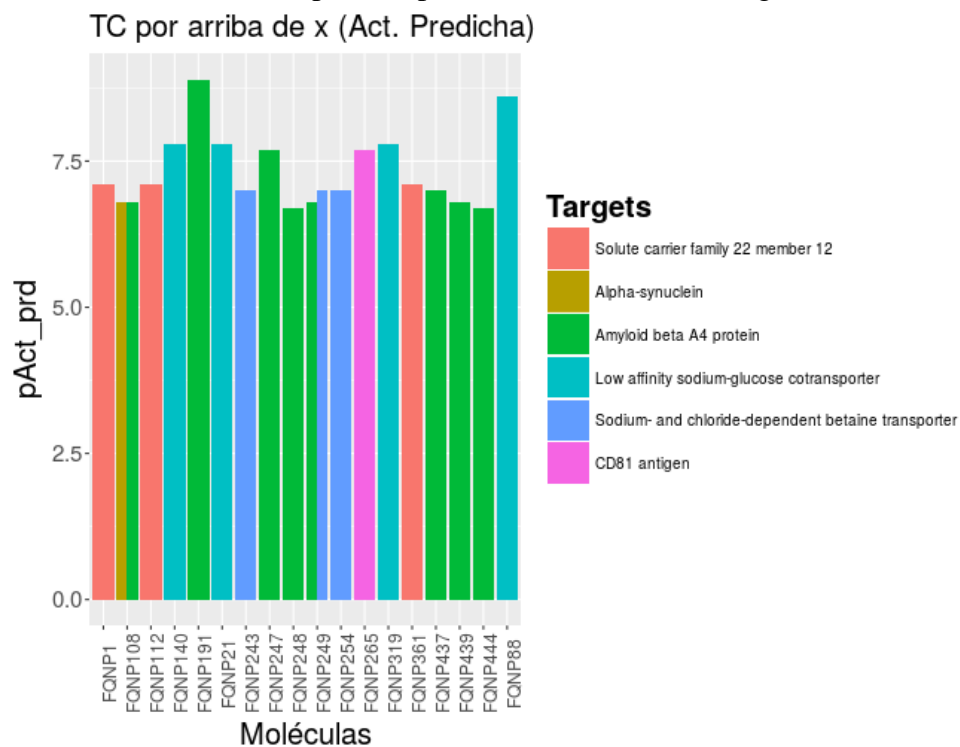


A.3.8 Transportadores (TC)

I. Gráfico de número de moléculas por diana biológica de la familia TC.

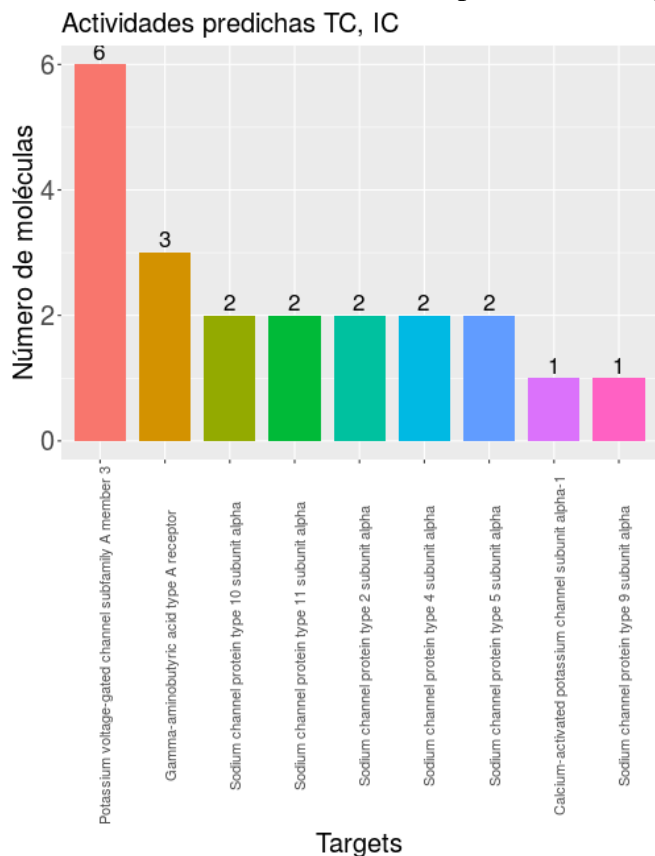


II. Gráfico de actividad por compuesto en cada diana biológica de la familia TC.

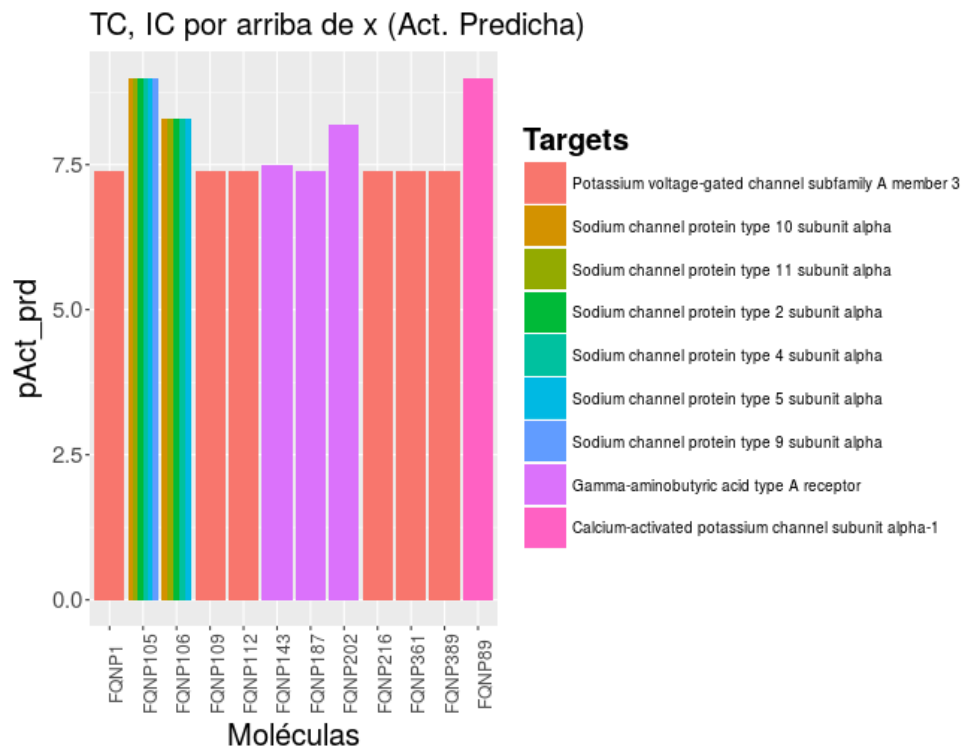


A.3.9 Transportadores y canales iónicos (TC, IC)

I. Gráfico de número de moléculas por diana biológica de la familia TC, IC.

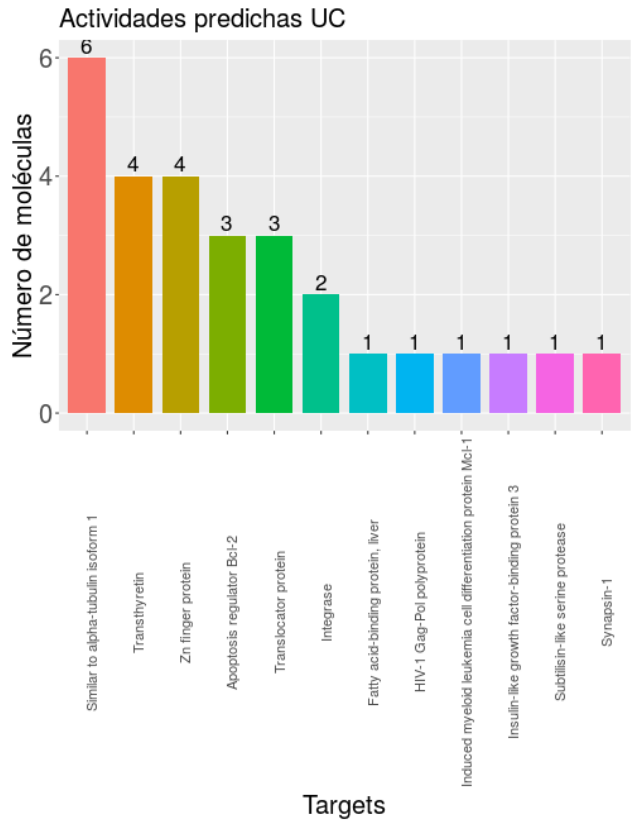


II. Gráfico de actividad por compuesto en cada diana biológica de la familia TC, IC.

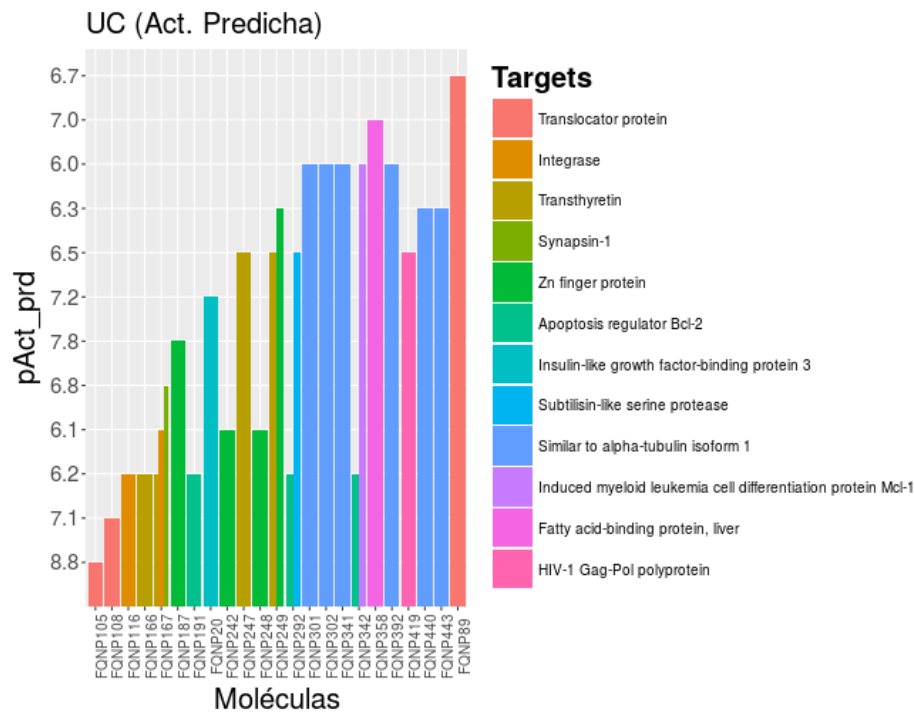


A.3.10 Sin clasificar (UC)

I. Gráfico de número de moléculas por diana biológica de la familia UC.

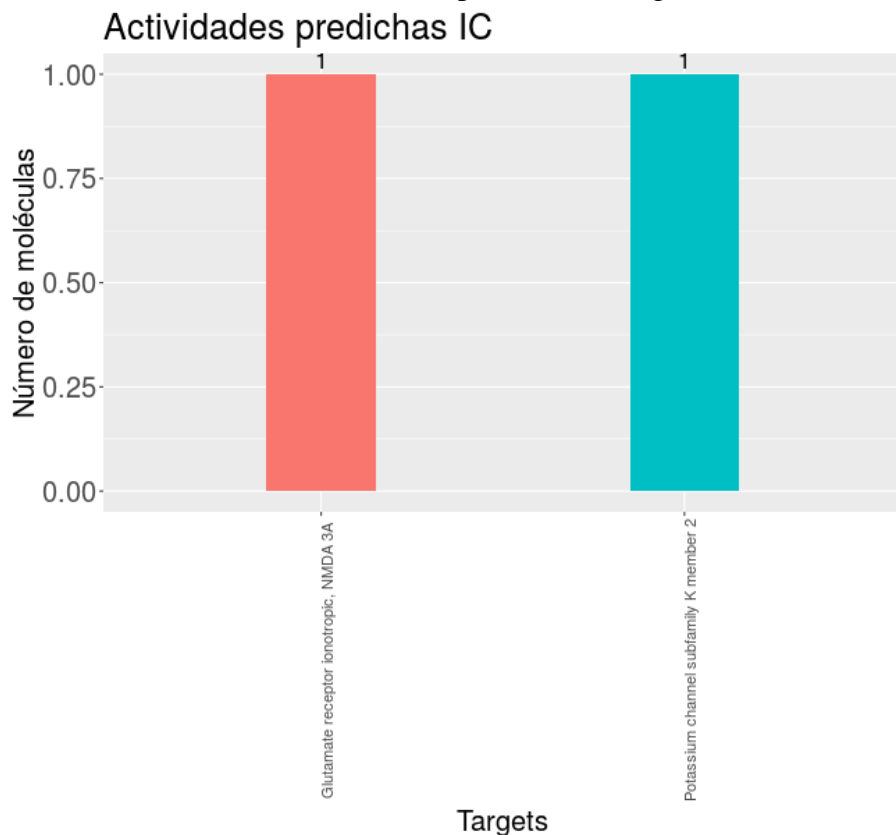


II. Gráfico de actividad por compuesto en cada diana biológica de la familia UC.

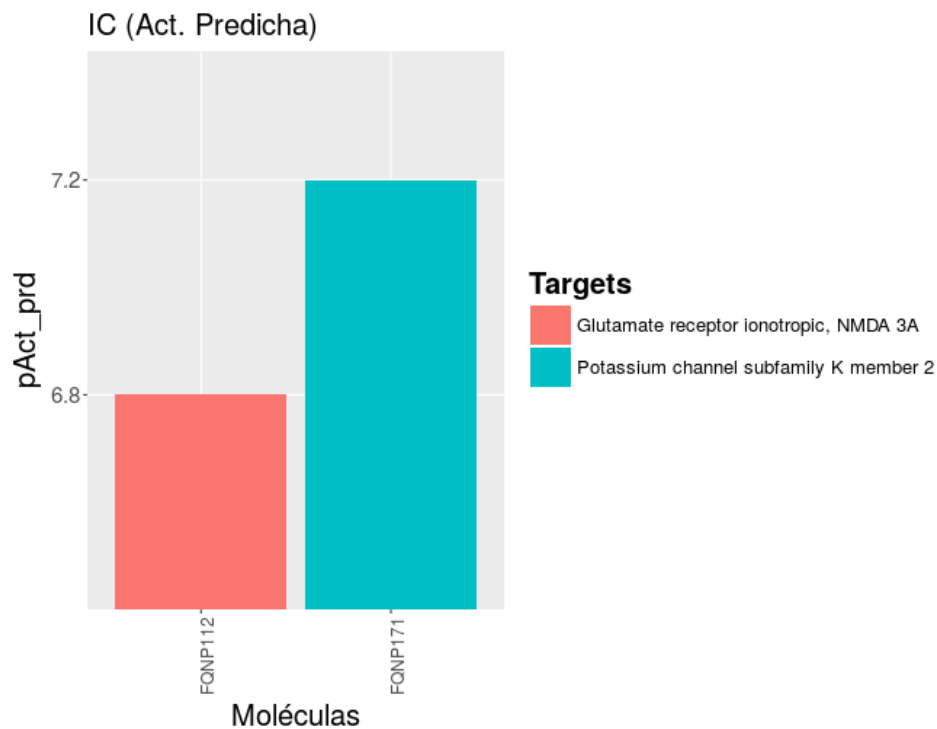


A.3.11 Canales iónicos (IC)

I. Gráfico de número de moléculas por diana biológica de la familia IC.



II. Gráfico de actividad por compuesto en cada diana biológica de la familia IC.



A.4 Artículos publicados

A.4.1 Pilón-Jiménez BA, Saldívar-González FI, Díaz-Eufracio BI, Medina-Franco JL. BIOFACQUIM: A Mexican compound database of natural products. *BIOMOLECULES* **2019** 9:31



A.4.2 Saldívar-González FI, Gómez-García A, Chávez-Ponce de León D.E, Sánchez-Cruz N, Ruiz-Rios J, Pilón-Jiménez BA, Medina-Franco J.L. 2018. Inhibitors of DNA methyltransferases from natural sources: A computational perspective. *FRONTIERS IN PHARMACOLOGY* **2018** 9:1144.

A.4.3. Naveja JJ., Pilón-Jiménez BA, Bajorath J, Medina-Franco JL. A general approach for retrosynthetic molecular core analysis. *JOURNAL OF CHEMINFORMATICS* (**2019**), en revisión de pares.

A.4.4 Saldívar-González FI, Pilón-Jiménez BA, and Medina-Franco JL. 2018. Chemical space of naturally occurring compounds *PHYSICAL SCIENCES REVIEW* **2019**, 4: 20180103.

Article

BIOFACQUIM: A Mexican Compound Database of Natural Products

B. Angélica Pilon-Jiménez , Fernanda I. Saldívar-González, Bárbara I. Díaz-Eufracio and José L. Medina-Franco * 

Department of Pharmacy, National Autonomous University of Mexico, Mexico City 04510, Mexico; angiepilon96@gmail.com (B.A.P.-J.); felilang12@gmail.com (F.I.S.-G.); debi_1223@hotmail.com (B.I.D.-E.)

* Correspondence: medinajl@unam.mx; Tel.: +5255-5622-3899

Received: 29 November 2018; Accepted: 15 January 2019; Published: 17 January 2019



Abstract: Compound databases of natural products have a major impact on drug discovery projects and other areas of research. The number of databases in the public domain with compounds with natural origins is increasing. Several countries, Brazil, France, Panama and, recently, Vietnam, have initiatives in place to construct and maintain compound databases that are representative of their diversity. In this proof-of-concept study, we discuss the first version of BIOFACQUIM, a novel compound database with natural products isolated and characterized in Mexico. We discuss its construction, curation, and a complete chemoinformatic characterization of the content and coverage in chemical space. The profile of physicochemical properties, scaffold content, and diversity, as well as structural diversity based on molecular fingerprints is reported. BIOFACQUIM is available for free.

Keywords: chemical space; chemical data set; chemoinformatics; consensus diversity plot; drug discovery; molecular diversity; visualization

1. Introduction

The significance of compound databases in drug discovery projects is continuously increasing. In fact, compound databases and chemical data sets are a centerpiece in pharmaceutical companies and other academic and government research centers [1]. In addition to their role in compound databases, natural products have been a major resource in drug discovery [2,3]. As reviewed elsewhere, there are several drugs recently approved for clinical use that are natural products or synthetic analogues of hit compounds initially identified from natural sources. A notable example is the fungi metabolite migalastat (Galafold®), approved in 2018 for the treatment of the Fabry disease [4]. Not unsurprisingly, natural product-based drug discovery is being coupled with other major drug discovery strategies such as high-throughput screening and virtual screening. Natural products are again gaining attention in the scientific community to address novel and/or difficult molecular endpoints, for instance, epigenetic targets [5,6].

Several compound databases of natural products have been constructed, curated and often maintained by academic and other not-for-profit research groups. Notable examples are the Universal Natural Product Database (UNPD) [7] and the Traditional Chinese Medicine (TCM) Database@Taiwan [8]. Of note, UNPD is no longer available online but represents the efforts of an academic group to assemble a large natural product database. Reference [4] confirms that there are other compound databases that collect natural products from specific geographical areas and countries, such as NuBBE_{DB} for natural products from Brazil [9] VIETHERB: A Database for Vietnamese Herbal Species was recently released to the public [10]. Other databases of natural products are discussed elsewhere [11–13]. Despite the fact that Mexico also has high levels of biodiversity, there are limited

efforts to assemble a compound database of natural products. One example is UNIQUIM, recently reviewed by Medina-Franco [11].

The objective of this work is to introduce BIOFACQUIM as one of the first compound databases of natural products isolated and characterized in Mexico. In this proof-of-concept study, we discuss the assembly of the first version of this chemical data set along with a chemoinformatic characterization of molecular diversity, scaffold content and coverage in chemical space. The compound database is freely available via the web-interface BIOFACQUIM Explorer (<https://biofacquim.herokuapp.com/>), and is part of an initial effort towards building, updating and maintaining a compound database representative of the biodiversity of Mexico. Compounds in BIOFACQUIM are also available from ZINC15 at <http://zinc15.docking.org/catalogs/biofacquimnp/>

2. Materials and Methods

2.1. BIOFACQUIM Database

The database of natural products was assembled from a literature search. For the construction of the first version of BIOFACQUIM, the Scopus database (www.scopus.com) was searched using the keywords “natural products” and “School of Chemistry of the National Autonomous University of Mexico (FQ, UNAM)”. This search led to a list of scientific papers and researchers that work with natural products. The eight journals that had contributed the most thus far were selected: *Journal of Ethnopharmacology*, *Natural Products Research*, *Journal of Agricultural and Food Chemistry*, *Journal of Natural Products*, *Planta Medica*, *Phytochemistry*, *Natural Product Letters*, and *Molecules*. As part of the search, three filters were used for the selection of the articles in each journal. The first filter was the search by institution (FQ, UNAM), the second was the search by publication year (2000–2018), and the last was the detailed analysis of the articles to identify if the procedure for the isolation, purification and characterization of the compounds from natural products was present. We want to emphasize that this is the first version of BIOFACQUIM; future versions will have natural products from more years, more peer-reviewed journals and more institutions, to achieve a database representative of the biodiversity of Mexico.

With the module ‘Wash’, from the molecular operating environment (MOE) program version 2018 [14], the database was curated. This was done to normalize and collect the most relevant information from the molecules. The data curation involved the elimination of salts, the adjustment of the protonation states, the optimization of the geometry by energy minimization and the elimination of the duplicated molecules. The default settings of the ‘Wash’ module were used.

2.2. Reference Data Sets

In order to characterize the diversity of BIOFACQUIM and to explore its coverage in chemical space, seven compound databases of broad interest in drug discovery were used as references. The structure files used in this work were taken from previous comparisons and chemoinformatic analyses of natural products [15]. The structures of the reference compounds were curated using the same procedure described to prepare BIOFACQUIM. Table 1 summarizes the reference databases and the number of compounds. Of note, the reference collections include seven data sets of natural products.

Table 1. Reference databases [15] compared for BIOFACQUIM.

Database	Size ^a
Approved drugs	1806
Cyanobacteria metabolites	473
Fungi metabolites	206
Marine	6253
MEGx	4103

Table 1. Cont.

Database	Size ^a
Semi-synthetics (NATx)	26,318
NuBBE _{DB}	2214

^a Number unique compounds after data curation.

2.3. Molecular Properties of Pharmaceutical Relevance

The curated BIOFACQUIM database was characterized by calculating six physicochemical properties of therapeutic interest, namely: molecular weight (MW), octanol/water partition coefficient (SlogP), topological surface area (TPSA), number of rotatable bonds (RB), number of H-bond donor atoms (HBD) and number of H-bond acceptor atoms (HBA). The statistical analysis was done, with the program DataWarrior [16], by calculating the mean, median and standard deviation of the calculated properties. Based on these statistics BIOFACQUIM was further compared with other natural products databases (NuBBE_{DB}, cyanobacteria, fungi, marine, and MEGx), approved drugs, and semisynthetic compounds (NATx) (Table 1).

2.4. Scaffold Content

Scaffold content analysis enabled us to identify the most frequent scaffolds in compound data sets and, in this work, to compare the scaffolds containing approved drugs with those containing natural products. The scaffold content analyses also enabled us to identify potential novel scaffolds. The most frequent core molecular scaffolds of BIOFACQUIM were computed using the definition described by Bemis and Murcko [17], in which the core scaffold is obtained by systematically removing the side chains of the compounds. The most frequent scaffolds in BIOFACQUIM were compared with data from the literature (vide infra).

2.5. Visual Representation of Chemical Space

In order to generate a visual representation of the chemical space of BIOFACQUIM, two visualization methods were used: principal component analysis (PCA) and *t*-distributed stochastic neighbor embedding (*t*-SNE). PCA reduces data dimensions by geometrically projecting them onto lower dimensions called principal components (PCs). The first PC is chosen to minimize the total distance between the data and its projection on the PC and to maximize the variance of the projected points.

t-SNE is a nonlinear dimension reduction in which Gaussian probability distributions over high-dimensional space are constructed and used to optimize a Student *t*-distribution in low-dimensional space. The low-dimensional space maintains the pairwise similarity to the high-dimensional space, leading to a clustering on the embedding space without any significant loss of structural information. Further details of each visualization method of the chemical space are discussed elsewhere [18,19]. In this work, for *t*-SNE, subsets of compounds were retrieved from large reference data sets (Table 1), namely: 40 % of the Marine, MEGx, and NuBBE_{DB} data sets (2501, 1641, and 886 compounds, respectively). For NATx and approved drugs, 1000 molecules were used. For cyanobacteria metabolites and fungi data sets the entire databases were employed (473 and 206 compounds, respectively).

2.6. Global Diversity: Consensus Diversity Analysis

Since the chemical diversity strongly depends on the structure representation, it is practical to consider multiple representations for a complete, global assessment. To this end, consensus diversity (CD) plots have been proposed as simple two-dimensional graphs that enable the comparison of the diversity of compound data sets using four sets of structural representations [20]; these are typically

the molecular fingerprints, scaffolds, molecular properties, and number of compounds. CD plots have been used to compare the diversity of natural products and other compound data sets [21]. Briefly, in a typical CD plot the scaffold and fingerprint diversity are represented along the y - and x -axes, respectively. The diversity based on whole molecular properties of pharmaceutical interest is represented by a continuous color scale and the number of compounds is mapped into the plot using different size data points. Further details are provided elsewhere [20]. To generate the CD plot of this work, for the y -axis we used the area under the cyclic system recovery curve [22]. For the x -axis, we employed the median of the fingerprint-based diversity computed with MACCS keys (166-bits) and the Tanimoto coefficient. Both are established and are representative metrics of the scaffold and fingerprint-based diversity. Subsets of the compounds were retrieved from large reference data sets (Table 1), considering the size of the databases. For NATx, Marine, MEGx, NuBBE_{DB} and approved drugs, 2000, 1500, 1000, 800 and 700 molecules, respectively, were used. For cyanobacteria metabolites and fungi data sets, the entire databases were employed (473 and 206 compounds, respectively).

3. Results and Discussion

First, we present the results of the construction of the first proof-of-concept version of the BIOFACQUIM database followed by a first chemoinformatic characterization in terms of physicochemical properties, scaffold content, diversity and coverage in chemical space.

3.1. BIOFACQUIM Database

As described in the Materials and Methods section, after the first survey in Scopus with the researchers of the FQ, UNAM, three filters were applied to the eight selected journals. Each of the 92 scientific papers selected was analyzed individually to extract information about the natural products. Of note, in this manuscript we disclose the first version of BIOFACQUIM as a proof-of-concept collection in which current content may be biased by the type of compounds published by a research group (e.g., based on their expertise and/or the analytical techniques available to their groups) and the type of compounds and characteristics accepted for publication by a given journal (e.g., compounds with the biological activity of compounds with drug-like features). It is anticipated that these biases will be reduced as the content of BIOFACQUIM is updated in future releases, by increasing the number of research groups, number of journals and number of years covered (cf. the Conclusions section).

The current version of BIOFACQUIM contains the following information: identification number (ID), compound name, simplified molecular input line entry system (SMILES), reference (with the name of the journal, digital object identifier (DOI) number and publication year), kingdom (Plantae or Fungi), genus, species, and geographical location of the collection of the natural product. In addition, the biological activity, if it was reported in the publication, has been included. The current and first version of BIOFACQUIM has 423 compounds. It should be noted that 316 compounds were isolated from 49 different plant genera, 98 were isolated from 19 genera of fungi, and nine compounds were isolated from Mexican propolis (a sticky dark-colored hive product collected by bees from living plant sources). Figure 1 shows the distribution of compounds per year reported since the year 2000, as contained in the first version of the chemical data set. The compounds in the database that were published in 2018 are not included in Figure 1.

Figure 2 shows the chemical structures of representative compounds from the first version of BIOFACQUIM (discussed further below).

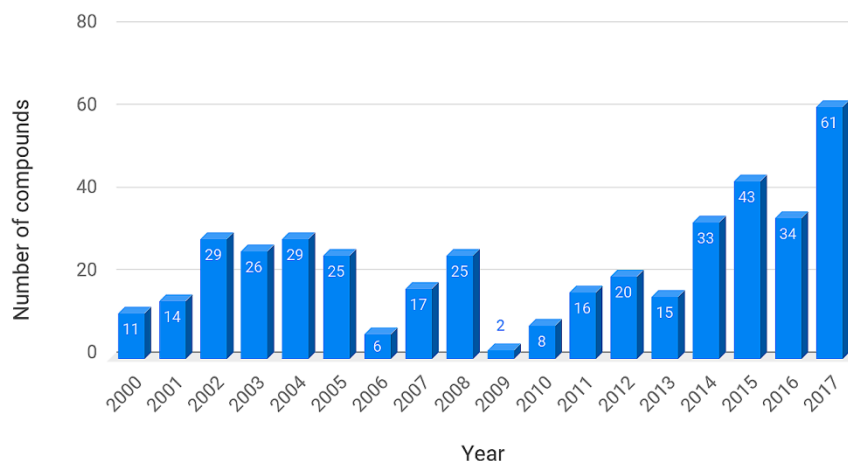


Figure 1. Distribution of compounds reported from 2000 to 2017, as contained in the first version of BIOFACQUIM. Compounds published in 2018 are not shown in this graph.

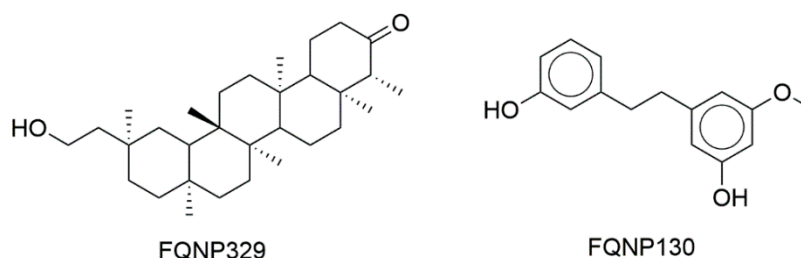


Figure 2. Select compounds contained in BIOFACQUIM.

3.2. Molecular Properties

Figure 3 shows box plots of the distribution of the six calculated physicochemical properties (vide supra) calculated for BIOFACQUIM. For comparative purposes, the box plots also include the distribution of the same properties of the seven reference data sets that were retrieved from the literature [15]. The corresponding violin plots are shown in the Supplementary Figure S3. The three main molecular properties, size, flexibility, and molecular polarity, are described by MW, RB, and SlogP, TPSA, HBA, and HBD, respectively. In these plots, the boxes enclose the data points with values within the first and third quartile; the line that divides the box denotes the median of the distributions. The lines above and below indicate the upper and lower adjacent values. The red asterisks indicate the data points with values beyond the upper and lower adjacent values. Summary statistics are presented at the bottom of the box plots. The figure also includes a table below each box plot with the maximum, median, mean, standard deviation and minimum values for each property and each data set.

According to Figure 3 (and the violin plots in the Supplementary Material), based on the mean of RB, BIOFACQUIM compounds have comparable flexibility to approved drugs. The figure also shows that, except for cyanobacteria metabolites, all databases have a median of up to five rotatable bonds (including approved drugs). The median and mean MW of BIOFACQUIM are 340.5 and 412 g/mol, respectively. Notably, BIOFACQUIM and NuBBEDB have the most similar MW profile compared to drugs. BIOFACQUIM has a median of 4 HBA, the same as that of the NuBBEDB and Marine data sets. Furthermore, BIOFACQUIM has a very similar profile of HBA compared to MEGx. Comparing HBD, BIOFACQUIM, NuBBEDB, NATx, and cyanobacteria have the same median values, with similar profiles to approved drugs and higher standard deviations than approved drugs. Regarding TPSA, the compounds in BIOFACQUIM are those that share the closest values to the approved drugs. It should be noted that the cyanobacteria metabolite set has the largest distribution and the highest mean values of TPSA, being the double of the mean of the approved drugs. The distribution of the SlogP values indicates that, overall, natural products are slightly more hydrophobic than approved drugs.

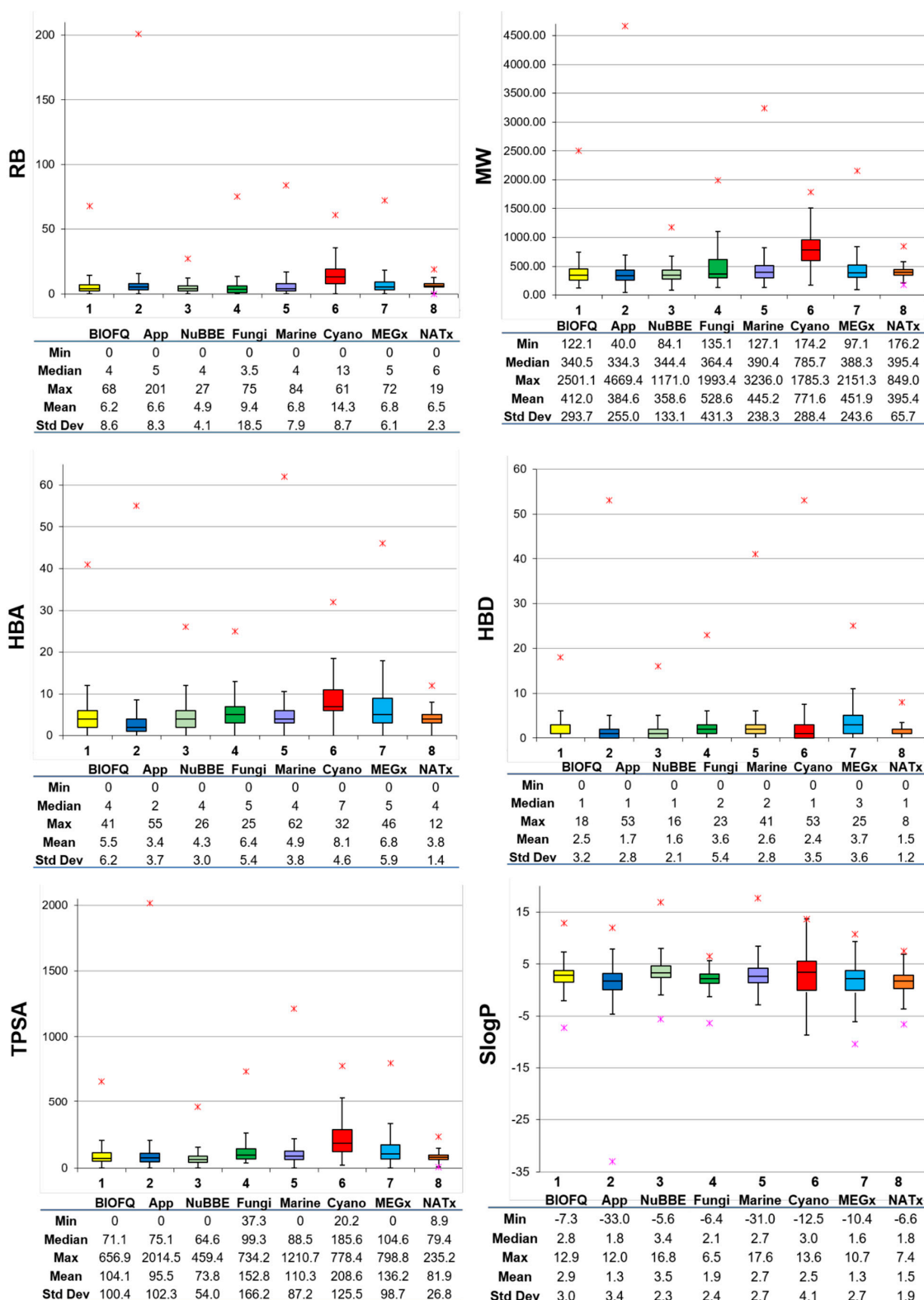


Figure 3. Box plots for the physicochemical properties of BIOFACQUIM (BIOFQ) and reference data sets (Table 1). The boxes enclose data points with values within the first and third quartile. The red asterisks indicate outliers. Summary statistics are included below each plot. RB: number of rotatable bonds; MW: molecular weight; HBA: number of H-bond acceptor atoms; HBD: number of H-bond donor atoms; TPSA: topological surface area; SlogP: octanol/water partition coefficient.

Taking together the results of the general profile of the properties, it can be concluded that the current version of BIOFACQUIM is, in general, most similar to the NuBBEDB and Fungi data sets. This outcome is in agreement with the findings that, while assembling BIOFACQUIM and analyzing the source papers in detail, the compounds were mostly isolated from plants and fungi.

3.3. Scaffold Content

Figure 4 shows the 27 most populated molecular scaffolds in BIOFACQUIM that included half (50.6 %) of the 423 compounds making up the database. Aside from benzene which is also frequent in several other compound databases [21], the second most frequent scaffold was a flavan-related scaffold (5 %), followed by 1,3-benzodioxole and dibenzyl core scaffolds (2.4 %). Interestingly, the last three frequent scaffolds in BIOFACQUIM are not the most frequent in other databases of natural products [15].

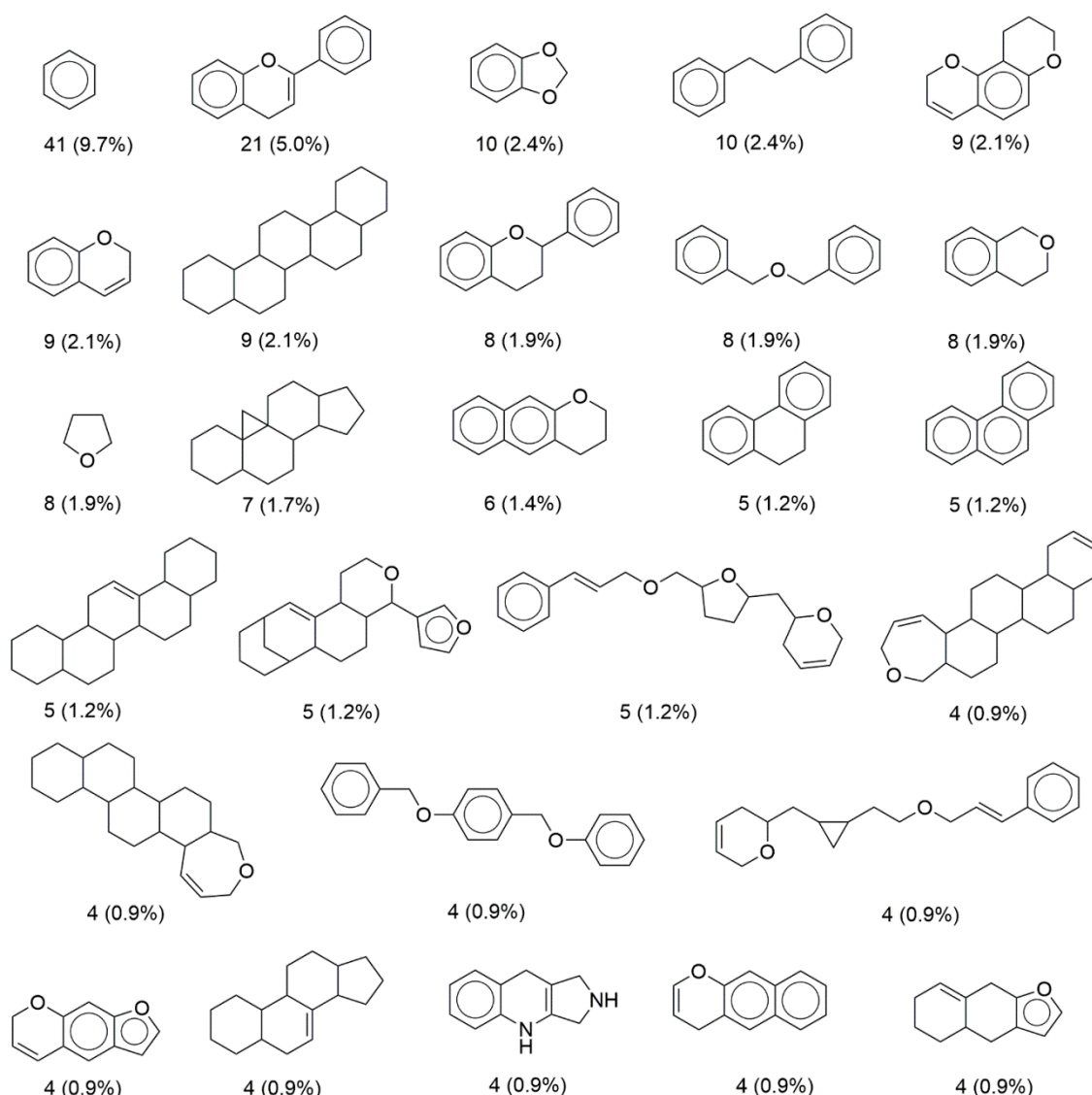


Figure 4. Most frequent scaffolds in BIOFACQUIM. The frequency and percentage are shown. The 27 scaffolds shown in the figure contain half of the total compounds in the database (50.6%).

3.4. Chemical Space

As explained in the Materials and Methods section, a visual analysis of the chemical space of BIOFACQUIM was done with two visualization methods, PCA and *t*-SNE. The visual representation

with PCA was based on the physicochemical properties while the visualization with *t*-SNE was based on the molecular topological fingerprints.

3.4.1. Visual Representation Based on Properties

Using the program KNIME [23], we did a visual comparison of the chemical space of BIOFACQUIM and the reference databases. We used the “Normalizer” node in KNIME which gives a linear transformation of all values, the minimum and maximum of each database. Then, PCA was applied to reduce the dimensionality of the six calculated physicochemical properties and to compare BIOFACQUIM with the reference collections (vide supra, Table 1).

Figure 5 shows a visual representation of the property-based chemical space. Table S1 in the Supplementary Material summarizes the corresponding loadings and eigenvalues for the first three PCs. The first two PCs capture 84% of the variance while the first three recover 92% of the variance. Table S1 shows that for the first PC, the larger loadings corresponded to SlogP, followed by RB, whereas for the second PC the largest loading corresponded to HBD.

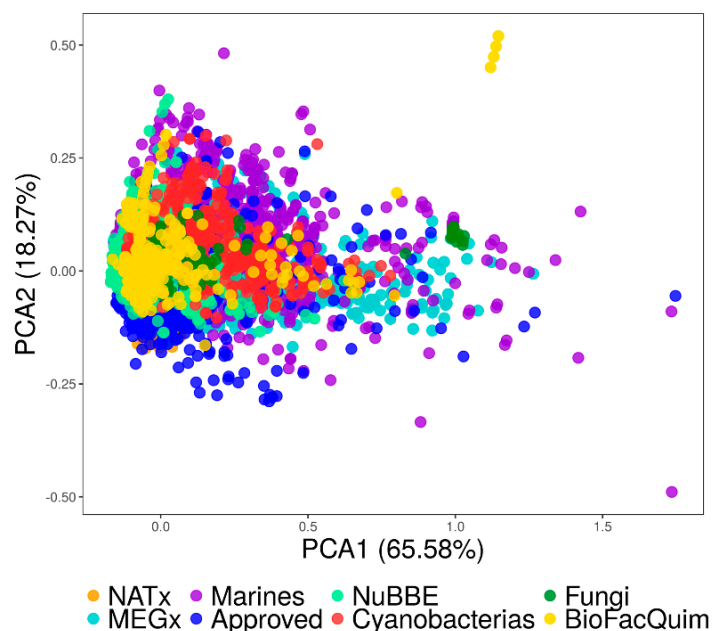


Figure 5. Visual representation of the chemical space based on the physicochemical properties of eight data sets. BIOFACQUIM (423 compounds, yellow); fungi metabolites (206 compounds, green); cyanobacteria metabolites (473 compounds, red); NuBBE_{DB} (2214 compounds, light green); NATx (26318 compounds, orange); MEGx (4103 compounds, blue); marine metabolites (6253 compounds, lilac); US Food and Drug Administration (FDA)-approved drugs (1806 compounds, dark blue).

The visual representation of the chemical space in Figure 5 indicates that some of the natural product compounds occupy the same space as the already approved drugs. It also shows that there are molecules in BIOFACQUIM and the Marine set that cover neglected regions of the currently drug-like chemical space. Finally, Figure 5 suggests that BIOFACQUIM shares the chemical space of almost all Fungi and NuBBE_{DB}.

3.4.2. Visual Representation Based on Molecular Fingerprints

Figure 6 shows a visual representation of the chemical space of the current version of BIOFACQUIM based on topological fingerprints using *t*-SNE (see Materials and Methods). Figure 6a compares BIOFACQUIM with all other reference data sets. Figure 6b shows a comparison of BIOFACQUIM with approved drugs. Figure 6a shows three main groups or clusters in which all the databases have compounds. The clusters indicate that the visualization method and the fingerprints

can distinguish three major core structures that would have detailed variations in the structure. Figure 6b indicates that there are compounds in BIOFACQUIM with high structural similarity to approved drugs. Notable examples are the compounds FQNP329 (chemical structure in Figure 2), similar to ethinylestradiol (App_75), and FQNP130, similar to choline (App_878). Other comparisons with *t*-SNE are shown in Figure S3 in the Supplementary Material.

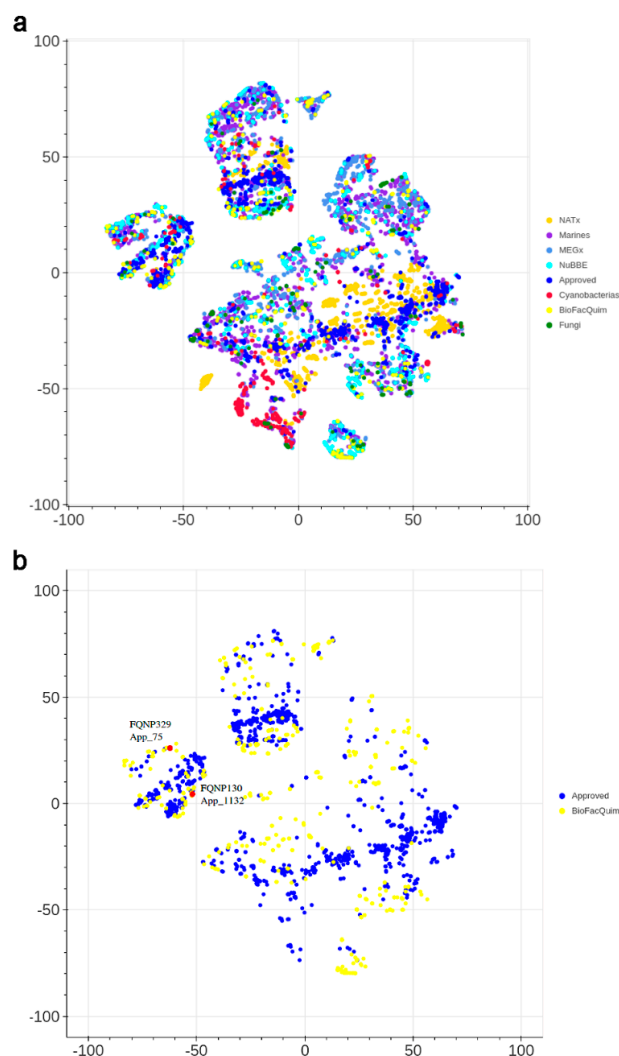


Figure 6. Visual representation of the chemical space of BIOFACQUIM compared with: (a) all reference data sets; and (b) approved drugs. The visualization was generated using *t*-distributed stochastic neighbor embedding (*t*-SNE) based on topological fingerprints. The red dots indicate the position of two representative compounds of BIOFACQUIM that are very similar to approved drugs.

Based on the assessment of the chemical space, in particular the position of BIOFACQUIM relative to other reference libraries in chemical space, it can be concluded that the compounds in BIOFACQUIM are very similar to drugs, based on their physicochemical properties (PCA) and structural fingerprints (*t*-SNE). Therefore, the chemical space analysis further supports the use of BIOFACQUIM in drug discovery projects.

3.5. Global Diversity: Consensus Diversity Analysis

As elaborated in the Materials and Methods section, a CD plot was used to compare the diversity of BIOFACQUIM with the diversity of the reference data sets, based on molecular fingerprints, scaffolds, and whole (physicochemical) properties. Figure 7 shows the CD plot, representing the

MACCS keys/Tanimoto similarity on the x -axis. Here, lower values indicate larger fingerprint-based diversity (further details of the fingerprint-based diversity assessment are presented in Figure S1 in the Supplementary Material). The y -axis of the CD plot represents the scaffold diversity where lower values (the area under the scaffold recovery curve—see Table S2 in the Supplementary Material) indicate higher scaffold diversity. The property-based diversity of BIOFACQUIM and each database was calculated as the Euclidean distance of the scaled properties. The values are represented on the color CD plot with data points on a continuous color scale. The darker color represents lower diversity while lighter colors represent higher diversity. Finally, the relative size of the databases is represented with different point sizes, where smaller data points indicate data sets with less number of molecules. The CD plot in Figure 7 shows that BIOFACQUIM and Cyanobacteria are found in the area representing low diversity of both scaffold and fingerprints. This may be attributed to the fact that this is the first version of the database. Regarding the diversity, based on physicochemical properties, the cyanobacteria metabolites were observed to have more diversity (e.g., lighter blue data point in Figure 7) than BIOFACQUIM. This is consistent with the analysis of the box plots discussed in Section 3.2. Figure 7 also indicates that approved drugs have high scaffold and fingerprint diversity that is consistent with previous reports [20,21].

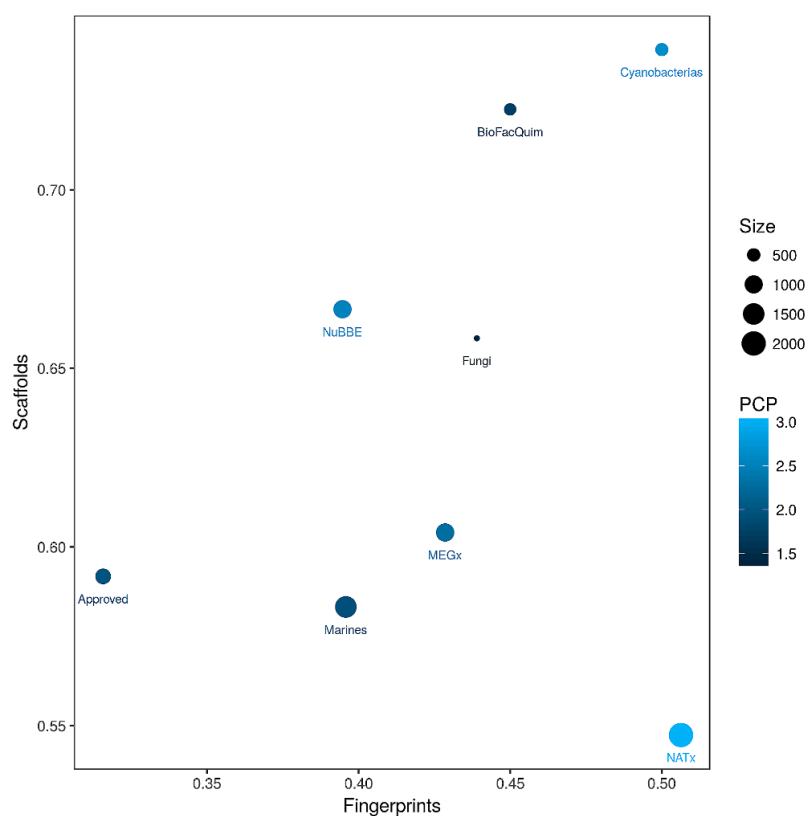


Figure 7. Consensus Diversity Plot comparing the global diversity of BIOFACQUIM with other natural product databases. The structural diversity (fingerprint diversity), calculated with the median Tanimoto coefficient of MACCS keys fingerprints, is plotted on the x -axis. The scaffold diversity of each database was defined as the area under the curve (AUC) of the respective scaffold recovery curves and is represented on the y -axis. The diversity, based on physicochemical properties (PCP), was calculated with the Euclidean distance of six scaled properties (SlogP, TPSA, MW, RB, HBD and HBA) and is shown on a color scale. The distance is represented with a continuous color scale from light blue (more diverse) to dark blue (less diverse). The relative size of the data set is represented with the size of the data point, smaller data points indicate compound data sets with fewer molecules.

4. Conclusions

BIOFACQUIM is a compound database of natural products from Mexico being constructed, curated and maintained by an academic group. The first and current version of BIOFACQUIM includes 423 compounds reported over the past 10 years at the School of Chemistry of the National Autonomous University of Mexico (UNAM). The compound database contains the chemical name, SMILES notation, reference (with name of the journal, year of publication and DOI number), kingdom (Plantae or Fungi), genus and species of the natural product, and geographical location of the collection. In addition, the biological activity, if it was reported in the publication, was included. The chemoinformatic characterization and analysis of the coverage and diversity of BIOFACQUIM in chemical space suggest broad coverage, overlapping with regions in the drug-like chemical space. The analysis also indicated that there are compounds in BIOFACQUIM with chemical structures very similar to drugs approved for clinical use that could, based on the similarity principle, be of pharmaceutical interest. Similar to other natural product databases, BIOFACQUIM can be used, via virtual screening, to identify potential lead compounds or starting points for additional optimization. The database is freely accessible through the website BIOFACQUIM Explorer, version 1.0 (<https://biofacquim.herokuapp.com>) and is part of the initiative D-TOOLS, described in detail elsewhere [24]. Compounds in BIOFACQUIM are also available from ZINC15 at <http://zinc15.docking.org/catalogs/biofacquimnp/>

One of the major objectives of this work, currently in progress, is to augment the size of BIOFACQUIM by expanding the search to other universities and research centers in Mexico, increasing the number of years and the number of scientific international peer-reviewed journals covered (with DOI number available). A second major objective of this work is to continue improving and maintaining the web-based interface BIOFACQUIM Explorer following general guidelines for the development and maintenance of public biological databases [25].

Supplementary Materials: The following are available online at <http://www.mdpi.com/2218-273X/9/1/31/s1>. Table S1. Loadings for the first three principal components of the property space of eight databases. Table S2. Statistics of the cyclic system recovery curves for BIOFACQUIM and the reference data sets. Figure S1. Distribution of the pairwise similarity values calculated for BIOFACQUIM and the reference data sets computed with MACCS keys (166-bits) and the Tanimoto coefficient. Figure S2. Visual representation of the chemical space of BIOFACQUIM generated with *t*-SNE. Figure S3. Violin plots for the physicochemical properties of BIOFACQUIM and reference data sets.

Author Contributions: Conceptualization, B.A.P.-J., F.I.S.-G., and J.L.M.-F.; methodology, B.A.P.-J., F.I.S.-G., and B.I.D.-E.; formal analysis, B.A.P.-J. and B.I.D.-E.; writing and editing, B.A.P.-J. and J.L.M.-F.; funding acquisition, J.L.M.-F.

Funding: This research was supported by the Programa de Apoyo a la Investigación y el Posgrado (PAIP) grant 5000-9163, Facultad de Química, UNAM, and project PAPIME (DGAPA, UNAM) PE200118.

Acknowledgments: B.A.P.-J. is grateful for the support given by the subprogram 127 “Basic Training in Research” of the School of Chemistry, UNAM. F.I.S.-G. and B.I.D.-E. are thankful to Consejo Nacional de Ciencia y Tecnología, Mexico (CONACyT) for scholarships, numbers 629458 and 620289, respectively. Discussions with Oscar Palomino-Hernández to implement *t*-SNE are acknowledged. We also thank John Irwin and Khanh Tang for adding BIOFACQUIM in the database ZINC15.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Miller, M.A. Chemical database techniques in drug discovery. *Nat. Rev. Drug Discov.* **2002**, *1*, 220–227. [[CrossRef](#)] [[PubMed](#)]
2. Newman, D.J. From natural products to drugs. *Phys. Sci. Rev.* **2018**. [[CrossRef](#)]
3. Newman, D.J.; Cragg, G.M. Natural products as sources of new drugs from 1981 to 2014. *J. Nat. Prod.* **2016**, *79*, 629–661. [[CrossRef](#)] [[PubMed](#)]
4. Saldívar-González, F.I.; Pilon-Jiménez, B.A.; Medina-Franco, J.L. Chemical space of naturally occurring compounds. *Phys. Sci. Rev.* **2018**. [[CrossRef](#)]

5. Saldívar-González, F.I.; Gómez-García, A.; Chávez-Ponce de León, D.E.; Sánchez-Cruz, N.; Ruiz-Rios, J.; Pilon-Jiménez, B.A.; Medina-Franco, J.L. Inhibitors of DNA methyltransferases from natural sources: A computational perspective. *Front. Pharmacol.* **2018**, *9*, 1144. [[CrossRef](#)] [[PubMed](#)]
6. Thomford, N.; Senthebane, D.; Rowe, A.; Munro, D.; Seele, P.; Maroyi, A.; Dzobo, K. Natural products for drug discovery in the 21st century: Innovations for novel drug discovery. *Int. J. Mol. Sci.* **2018**, *19*, 1578. [[CrossRef](#)] [[PubMed](#)]
7. Gu, J.; Gui, Y.; Chen, L.; Yuan, G.; Lu, H.-Z.; Xu, X. Use of natural products as chemical library for drug discovery and network pharmacology. *PLoS ONE* **2013**, *8*, e62839. [[CrossRef](#)]
8. Chen, C.Y.-C. TCM database@Taiwan: The world's largest traditional chinese medicine database for drug screening in silico. *PLoS ONE* **2011**, *6*, e15939. [[CrossRef](#)]
9. Pilon, A.C.; Valli, M.; Dametto, A.C.; Pinto, M.E.F.; Freire, R.T.; Castro-Gamboa, I.; Andricopulo, A.D.; Bolzani, V.S. NuBBE_{DB}: An updated database to uncover chemical and biological information from brazilian biodiversity. *Sci Rep* **2017**, *7*, 7215. [[CrossRef](#)]
10. Nguyen-Vo, T.-H.; Le, T.Q.M.; Pham, D.T.; Nguyen, T.D.; Le, P.H.; Nguyen, A.D.T.; Nguyen, T.D.; Nguyen, T.-N.N.; Nguyen, V.A.; Do, H.T.; et al. VIETHERB: A database for vietnamese herbal species. *J. Chem. Inf. Model.* **2018**. [[CrossRef](#)]
11. Medina-Franco, J.L. Discovery and development of lead compounds from natural sources using computational approaches. In *Evidence-Based Validation of Herbal Medicine*; Mukherjee, P., Ed.; Elsevier: Amsterdam, The Netherlands, 2015; pp. 455–475.
12. Tung, C.-W. Public databases of plant natural products for computational drug discovery. *Curr. Comput. Aided Drug Des.* **2014**, *10*, 191–196. [[CrossRef](#)] [[PubMed](#)]
13. Chen, Y.; Garcia de Lomana, M.; Friedrich, N.-O.; Kirchmair, J. Characterization of the chemical space of known and readily obtainable natural products. *J. Chem. Inf. Model.* **2018**, *58*, 1518–1532. [[CrossRef](#)] [[PubMed](#)]
14. *Molecular Operating Environment (MOE)*, version 2018.08; Chemical Computing Group Inc.: Montreal, QC, Canada, 2018; Available online: <http://www.chemcomp.com> (accessed on 28 November 2018).
15. Saldívar-González, F.I.; Valli, M.; Andricopulo, A.D.; da Silva Bolzani, V.; Medina-Franco, J.L. Chemical diversity of NuBBE database: A chemoinformatic characterization. *J. Chem. Inf. Model.* **2019**. [[CrossRef](#)]
16. Sander, T.; Freyss, J.; von Korff, M.; Rufener, C. Datawarrior: An open-source program for chemistry aware data visualization and analysis. *J. Chem. Inf. Model.* **2015**, *55*, 460–473. [[CrossRef](#)] [[PubMed](#)]
17. Bemis, G.W.; Murcko, M.A. The properties of known drugs. 1. Molecular frameworks. *J. Med. Chem.* **1996**, *39*, 2887–2893. [[CrossRef](#)]
18. Van der Maaten, L.; Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.
19. 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. *Exp. Opin. Drug Discov.* **2015**, *10*, 959–973. [[CrossRef](#)] [[PubMed](#)]
20. González-Medina, M.; Prieto-Martínez, F.D.; Medina-Franco, J.L. Consensus diversity plots: A global diversity analysis of chemical libraries. *J. Cheminf.* **2016**, *8*, 63. [[CrossRef](#)]
21. Naveja, J.; Rico-Hidalgo, M.; Medina-Franco, J. Analysis of a large food chemical database: Chemical space, diversity, and complexity. *F1000Research* **2018**, *7*, 993. [[CrossRef](#)]
22. Medina-Franco, J.L.; Martínez-Mayorga, K.; Bender, A.; Scior, T. Scaffold diversity analysis of compound data sets using an entropy-based measure. *QSAR Comb. Sci.* **2009**, *28*, 1551–1560. [[CrossRef](#)]
23. Berthold, M.R.; Cebron, N.; Dill, F.; Gabriel, T.R.; Kötter, T.; Meinl, T.; Ohl, P.; Sieb, C.; Thiel, K.; Wiswedel, B. Knime: The konstanz information miner. In *Data analysis, machine learning and applications: Proceedings of the 31st Annual Conference of the Gesellschaft für Klassifikation e.V., Albert-Ludwigs-Universität Freiburg, Freiburg im Breisgau, Germany, 7–9 March 2007*; Preisach, C., Burkhardt, H., Schmidt-Thieme, L., Decker, R., Eds.; Springer: Berlin/Heidelberg, Germany, 2008; pp. 319–326.
24. Naveja, J.J.; Oviedo-Osornio, C.I.; Trujillo-Minero, N.N.; Medina-Franco, J.L. Chemoinformatics: A perspective from an academic setting in Latin America. *Mol. Divers.* **2018**, *22*, 247–258. [[CrossRef](#)] [[PubMed](#)]
25. Helmy, M.; Crits-Christoph, A.; Bader, G.D. Ten simple rules for developing public biological databases. *PLoS Comput. Biol.* **2016**, *12*, e1005128. [[CrossRef](#)] [[PubMed](#)]





Inhibitors of DNA Methyltransferases From Natural Sources: A Computational Perspective

*Fernanda I. Saldívar-González, Alejandro Gómez-García, David E. Chávez-Ponce de León, Norberto Sánchez-Cruz, Javier Ruiz-Ríos, B. Angélica Pilon-Jiménez and José L. Medina-Franco**

Department of Pharmacy, School of Chemistry, National Autonomous University of Mexico, Mexico City, Mexico

OPEN ACCESS

Edited by:

Shibashish Giri,
University of Leipzig, Germany

Reviewed by:

Jaigopal Sharma,
Delhi Technological University, India
Sujata Mohanty,
All India Institute of Medical Sciences,
India
Rup Lal,
University of Delhi, India

*Correspondence:

José L. Medina-Franco
medinajl@unam.mx;
jose.medina.franco@gmail.com

Specialty section:

This article was submitted to
Pharmacogenetics
and Pharmacogenomics,
a section of the journal
Frontiers in Pharmacology

Received: 08 August 2018

Accepted: 21 September 2018

Published: 10 October 2018

Citation:

Saldívar-González FI,
Gómez-García A,
Chávez-Ponce de León DE,
Sánchez-Cruz N, Ruiz-Ríos J,
Pilon-Jiménez BA and
Medina-Franco JL (2018) Inhibitors
of DNA Methyltransferases From
Natural Sources: A Computational
Perspective.
Front. Pharmacol. 9:1144.
doi: 10.3389/fphar.2018.01144

Naturally occurring small molecules include a large variety of natural products from different sources that have confirmed activity against epigenetic targets. In this work we review chemoinformatic, molecular modeling, and other computational approaches that have been used to uncover natural products as inhibitors of DNA methyltransferases, a major family of epigenetic targets with therapeutic interest. Examples of computational approaches surveyed in this work are docking, similarity-based virtual screening, and pharmacophore modeling. It is also discussed the chemoinformatic-guided exploration of the chemical space of naturally occurring compounds as epigenetic modulators which may have significant implications in epigenetic drug discovery and nutriepigenetics.

Keywords: chemical space, chemoinformatics, databases, DNMT inhibitors, drug discovery, molecular modeling, similarity searching, virtual screening

SECTION 1: INTRODUCTION

Epigenetics has been defined as a change in phenotype without an underlying change in genotype (Berger et al., 2009). In the 1940s Waddington suggested the term “epigenetics” trying to describe “the interactions of genes with their environment, which brings the phenotype into being” (Waddington, 2012). Alterations in epigenetic modifications have been related to several diseases including cancer, diabetes, neurodegenerative disorders, and immune-mediated diseases (Dueñas-González et al., 2016; Tough et al., 2016; Hwang et al., 2017; Lu et al., 2018). Moreover, epigenetic targets are also attractive for the treatment of antiparasitic infections (Sacconay et al., 2014).

In epigenetic drug discovery, epigenetic targets have been classified into three main groups (Ganesan, 2018). “Writers” are enzymes that catalyze the addition of a functional group to a protein or nucleic acid; “readers” are macromolecules that function as recognition units that can distinguish a native macromolecule vs. the modified one; and “erasers” that are enzymes that aid in the removal of chemical modifications introduced by the writers. Thus far, several targets from these three major families have reached different stages of drug discovery, ranging from lead discovery, preclinical development, clinical trials and approval. Currently, there are seven compounds approved for clinical use (Ganesan, 2018).

DNA methyltransferases (DNMTs) are a family of “writer” enzymes responsible for DNA methylation that is the addition of a methyl group to the carbon atom number five (C5) of cytosine. As surveyed in this work, since DNA methylation has an essential role for cell differentiation and

development, alterations in the function of DNMTs have been associated with cancer (Castillo-Aguilera et al., 2017) and other diseases (Lyko, 2017).

Several natural products have been identified as inhibitors of epigenetic targets including DNMTs. Most of these compounds have been uncovered fortuitously. However, there are recent efforts to screen systematically natural products as DNMT inhibitors. The vastness of the chemical space of natural products led to the hypothesis that many more active compounds could potentially be identified. Indeed, it has been estimated that more than 95% of the biodiversity in nature remains to be explored to identify potential bioactive molecules (Ho et al., 2018).

The aim of this work is to discuss a broad range of computational methods to identify novel inhibitors of DNMTs from natural products. The manuscript also discusses the chemical space of natural products as inhibitors of DNMTs. The manuscript is organized into nine sections. After this introduction, Section 2 reviews briefly the structure of DNMTs including different isoforms. The next section covers major aspects of the function of DNMTs including the mechanism of methylation. Section 4 reviews currently known inhibitors of DNMTs from natural sources including food chemicals. Section 5 discusses the epigenetic relevant chemical space of natural products comparing the chemical space of DNMT inhibitors from natural sources vs. other compounds. The next section reviews computational strategies that are used to identify natural compounds as potential epi-hits or epi-leads targeting DNMTs. Sections 7 and 8 presents Summary conclusions and Perspectives, respectively.

SECTION 2: STRUCTURE OF DNMTs

The human genome encodes DNMT1, DNMT2, DNMT3A, DNMT3B, and DNMT3L. While DNMT1, DNMT3A, and DNMT3B have catalytic activity, DNMT2 and DNMT3L do not (Lyko, 2017). DNMT1 is a maintenance methyltransferase, responsible for duplicating the pattern of DNA methylation during replication. DNMT1 is essential for proper mammalian development and it has been proposed as the most interesting target for experimental cancer therapies (Dueñas-González et al., 2016). DNMT3A and DNMT3B are *de novo* methyltransferases. Human DNMT1 has 1616 amino acids whose structure can be divided into an N-terminal regulatory domain and a C-terminal catalytic domain (Jeltsch, 2002; Jurkowska et al., 2011). The N-terminal domain contains a replication foci-targeting domain, a DNA-binding CXXC domain, and a pair of bromo-adjacent homology domains. The C-terminal catalytic domain has 10 amino acid motifs. The cofactor and substrate binding sites in the C-terminal catalytic domain are comprised of motif I and X and motif IV, VI, and VIII, respectively (Lan et al., 2010). The target recognition domain which is maintained by motif IX and involved in DNA recognition, is not conserved between the DNMT family. **Figure 1A** shows a three-dimensional (3D) model of

a DNMT1 (PDB ID: 4WXX) (Zhang et al., 2015). **Figure 1B** shows a schematic diagram of human DNMT1, 2, 3A, 3B, and L.

Section 2.1: Isoforms

Two isoforms of DNMT3A have been identified, DNMT3A1 and DNMT3A2. At the N-terminal domain both isoforms have a PWWP (Pro-Trp-Trp-Pro) and an ADD (ATRX-DNMT3-DNMT3L) domains (Jurkowska et al., 2011). The C-terminal domain is identical in the two isoforms (Choi et al., 2011).

There are more than 30 isoforms of DNMT3B, however, only DNMT3B1 and DNMT3B2 are catalytically active (Ostler et al., 2007). Similar to DNMT3A, DNMT3B1, and DNMT3B2 have a PWWP and ADD domains at the N-terminal region (Lyko, 2017). The rest of the isoforms are not catalytically active. Some of these such as DNMT3B3, DNMT3B4, and DNMT3B7 are overexpressed in many tumor cell lines (Gordon et al., 2013). Δ DNMT3B has seven isoforms and lacks 200 amino acids from the N-terminal region of DNMT3B (Wang et al., 2006). Δ DNMT3B1–4 possess catalytic activity whereas Δ DNMT3B5–7 lacks the catalytic domain (Wang et al., 2006). Δ DNMT3B is mainly expressed in non-small cell lung cancer (Wang et al., 2006; Ostler et al., 2007). **Figure 1C** shows the identity matrix of 14 DNMTs isoforms. The identity matrix indicates that the amino acid sequence at the catalytic site of DNMT3A1 and DNMT3A2 isoforms is identical. In the same manner, the amino acid sequence at the C-terminal domain of the catalytically active isoforms DNMT3B1, DNMT3B2, and Δ DNMT3B1–4 are identical. DNMT1, DNMT2, and DNMT3L show a significant difference in the sequence of the catalytic site with respect to the rest of the isoforms. Therefore, it can be anticipated that is possible to identify or design selective inhibitors for these isoforms.

SECTION 3: FUNCTION AND MECHANISM OF DNMTs

As outlined in Section 2, cytosine-5 DNMTs catalyze the addition of methylation marks to genomic DNA. All DNMTs have a related catalytic mechanism that is featured by the formation of a covalent adduct intermediate between the enzyme and the substrate base. All DNMTs use *S*-adenosyl-*L*-methionine (SAM) as the methyl group donor (Vilkaitis et al., 2001; Du et al., 2016). DNMT forms a complex with DNA and the cytosine which will be methylated flips out from the DNA (Klimasauskas et al., 1994). A conserved cysteine performs a nucleophilic attack to the six-position of the target cytosine yielding a covalent intermediate. The five-position of the cytosine is activated and conducts a nucleophilic attack on the cofactor SAM to form the 5-methyl covalent adduct and *S*-adenosyl-*L*-homocysteine (SAH). The attack on the six-position is aided by a transient protonation of the cytosine ring at the endocyclic nitrogen atom N3, which can be stabilized by a glutamate and arginine residues. The covalent complex between the methylated base and the DNA is resolved by deprotonation at the five-position to generate the methylated cytosine and the free enzyme.

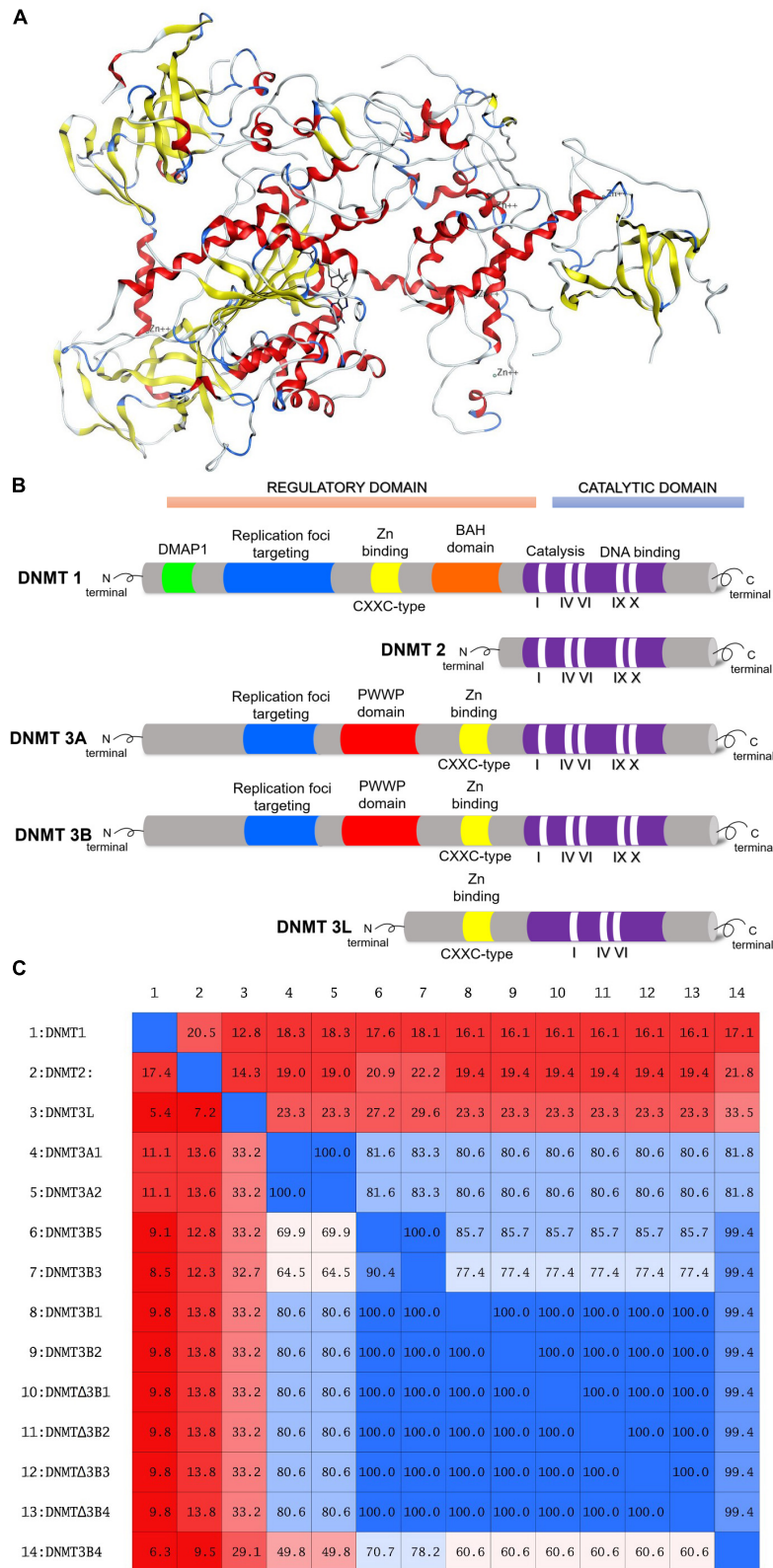


FIGURE 1 | (A) Three-dimensional model of DNMT1, amino acid residues 351–1600. Figure rendered from the Protein Data Bank PDB ID: 4WXX. **(B)** Schematic diagram of the structure of human DNMT1, DNMT2, DNMT3A, DNMT3B, and DNMT3L. **(C)** Identity matrix of the catalytic site of 14 DNMTs isoforms. Note that there is a significant difference in the sequence of DNMT1, DNMT2, and DNMT3L.

SECTION 4: KNOWN INHIBITORS OF DNMTs FROM NATURAL SOURCES

Thus far more than 500 compounds have been tested as inhibitors of DNMTs. The structural diversity and coverage in chemical space has been analyzed using chemoinformatic methods (Fernandez-de Gortari and Medina-Franco, 2015). The chemical space of DNMT inhibitors has been compared with inhibitors of other epigenetic targets (Naveja and Medina-Franco, 2018). Furthermore, the structure-activity relationships (SAR) of DNMT inhibitors using the concept of activity landscape has been documented (Naveja and Medina-Franco, 2015).

DNA methyltransferase inhibitors have been obtained from a broad number of different strategies including organic synthesis, virtual, and high-throughput screening (Medina-Franco et al., 2015). Organic synthesis has been employed in several instances for lead optimization (Castellano et al., 2008; Kabro et al., 2013; Davide et al., 2016). Natural products and food chemicals have also been a major source of active compounds. Natural products that are known to act as DNMT inhibitors or demethylating agents have been extensively reviewed by Zwergel et al. (2016). These natural products are of the type polyphenols, flavonoids, anthraquinones, and other classes. Some of the first natural products described were curcumin, (-)-epigallocatechin-3-gallate (EGCG), mahanine, genistein, and quercetin. Other natural products that have described as inhibitors of DNMT or demethylating agents are silibinin, luteolin, kazinol Q, laccic acid, hypericin, boswellic acid, and lycopene. **Figure 2** shows the chemical structure of representative DNMT inhibitors with emphasis on compounds from natural origin.

The bioactivity profile and potency in enzymatic and/or cell-based assays of these natural products have been discussed in detail by Zwergel et al. (2016). Of note, it will be valuable if all natural products could have been screened under the same conditions. For few natural products the selectivity has been characterized being nanaomycin A an exception (*vide infra*). Indeed, for about eight natural products the IC_{50} has been measured in enzymatic based assays. Despite the fact that the potency of the natural products with DNMTs is not very high in enzymatic-based assays, e.g., IC_{50} between 0.5 and 10 μ M, several natural products have shown promising activity in cell-based assays. Notably, natural products have distinct chemical scaffolds that could be used as a starting point in lead optimization efforts. Moreover, quercetin in combination with green tea extract has advanced into phase I clinical trials for the treatment of prostate cancer.

Most of the natural products with demethylating activity or ability to inhibit DNA methyltransferases in enzymatic assays have been identified fortuitously. However, as discussed in this work, there are efforts toward the identification of bioactive demethylating agents using systematic approaches such a virtual screening. Indeed, the natural product nanaomycin A (**Figure 2**) was identified from a virtual screening campaign initially focused on the identification of inhibitors of DNMT1. The quinone-based antibiotic isolated from *Streptomyces* showed antiproliferative effects in three human tumor cell lines, HCT116, A549, and HL60 after 72 h of treatment. Moreover, nanaomycin A showed

reduced global methylation levels in all three cell lines when tested at concentrations ranging from 0.5 to 5 μ M. Nanaomycin A reactivated the transcription of the RASSF1A tumor suppressor gene inducing its expression up to 18-fold at 5 μ M, higher than the reference drug 5-azacytidine (sixfold at 25 μ M). In an enzymatic inhibitory assay, nanaomycin A was selective toward DNMT3B with an $IC_{50} = 0.50 \mu$ M.

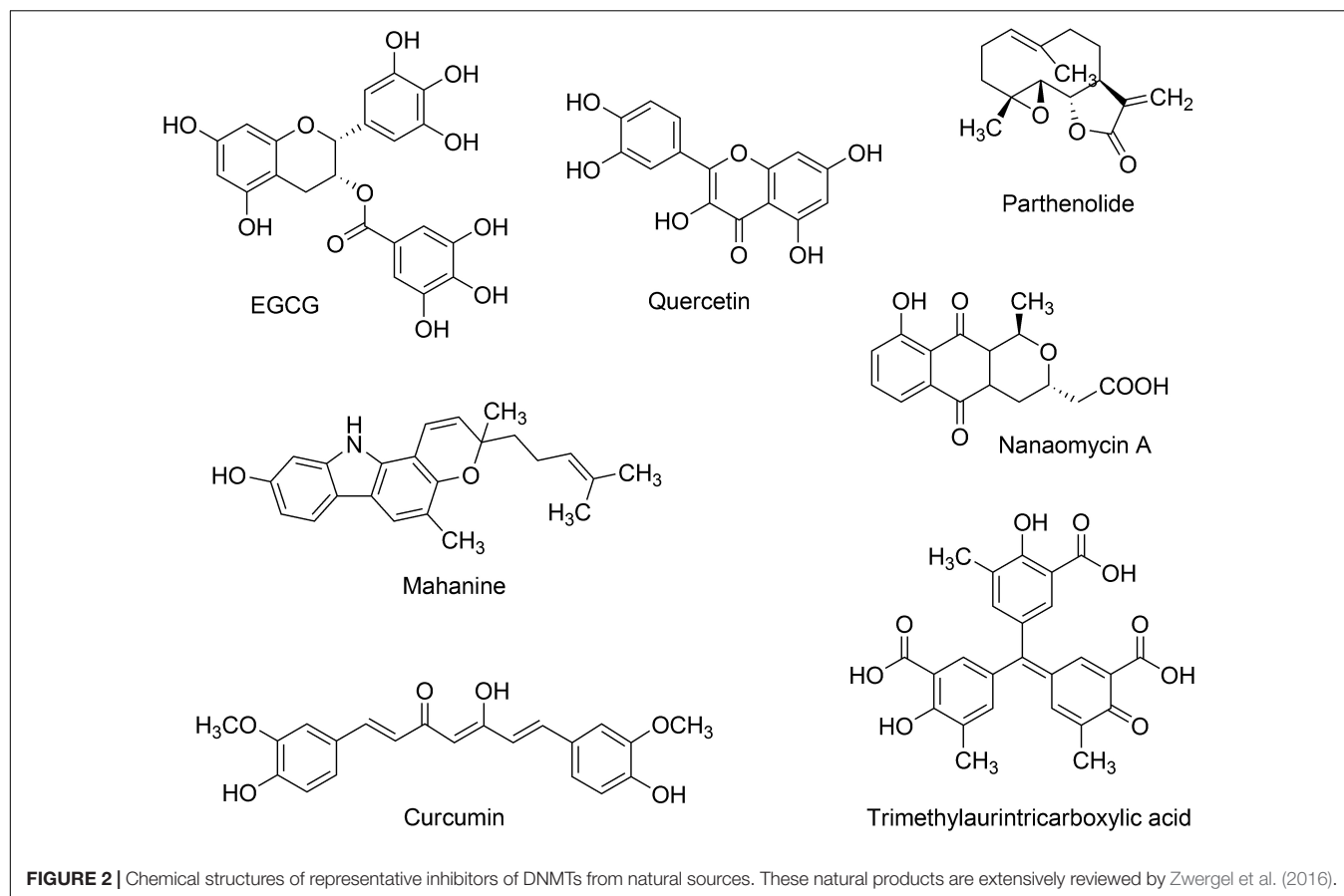
Section 4.1: Natural Products and Food Chemicals

It is remarkable that several natural products are used as dietary sources such as curcumin, caffeic acid and chlorogenic acid found in *Coffea arabica*, genistein found in soybean, quercetin found in fruits, vegetables, and beverages. Of course, there is a large overlap between the chemical space of food chemicals and natural products (Naveja et al., 2018). This has given rise to systematically screen food chemical databases for potential regulators of epigenetic targets.

SECTION 5: EPIGENETIC RELEVANT CHEMICAL SPACE OF NATURAL PRODUCTS: FOCUS ON DNMT INHIBITORS

In drug discovery it is generally accepted that a major benefit of natural products vs. purely synthetic organic molecules is, overall, the feasibility of the former to exert a biological activity and increased chemical diversity (Ho et al., 2018). The chemical space of natural products is vast and its molecular diversity has been quantified over the past few years (López-Vallejo et al., 2012; Olmedo et al., 2017; Shang et al., 2018). A major contribution to these studies has been the increasing availability of natural products collections in the public domain (Medina-Franco, 2015). Examples of major compound collections are the Traditional Chinese Medicine (Chen, 2011), natural products from Brazil – NuBBE (Pilon et al., 2017), AfroDb (Ntie-Kang et al., 2013) or collections available for screening in a medium to high-throughput screening mode. The large importance of natural products in drug discovery has boosted the development of open access applications to mine these rich repositories. Few examples are ChemGPS-NP, TCManalyzer, and other resources described elsewhere (Rosen et al., 2009; Chen et al., 2017; Gonzalez-Medina et al., 2017; Liu et al., 2018).

The chemical space of natural products from different sources has been compared to several other collections including the chemical space of drugs approved for clinical use and synthetic compounds (Olmedo et al., 2017; Shang et al., 2018). These studies demonstrate that the chemical space of natural products is vast, that there is a notable overlap with the chemical space of drugs, and that natural products cover novel regions of the chemical space. The overlap with the chemical space of approved drugs is not that surprising since there are a large percentage of drugs from natural origin. **Figure 3** shows a visual representation of the chemical space of 15 representative DNMT inhibitors from natural sources vs. 4103 compounds from a commercial



vendor library of natural products, 206 fungi metabolites, and 6253 marine natural products (Krishna et al., 2017). The visual representation was generated with principal component analysis of six physicochemical properties of pharmaceutical relevance, namely molecular weight (MW), topological surface area (TPSA), number of hydrogen bond donors and acceptors (HBD/HBA), number of rotatable bonds (RB), and octanol/water partition coefficient (logP). The first two principal components capture about 90% of the total variance. The visual representation of the chemical space in this figure indicates that marine natural products (data points in blue) cover a broader area of the chemical space followed by natural products in the vendor collection (orange) and by fungi metabolites (green). DNMT inhibitors from natural origin (purple) are, in general, inside the subspace of the DNMT1 inhibitors (red). This visualization of the chemical space indicates that there would be expected to identify more DNMT1 inhibitors in the marine and vendor collections, as well as in the data set of fungi metabolites.

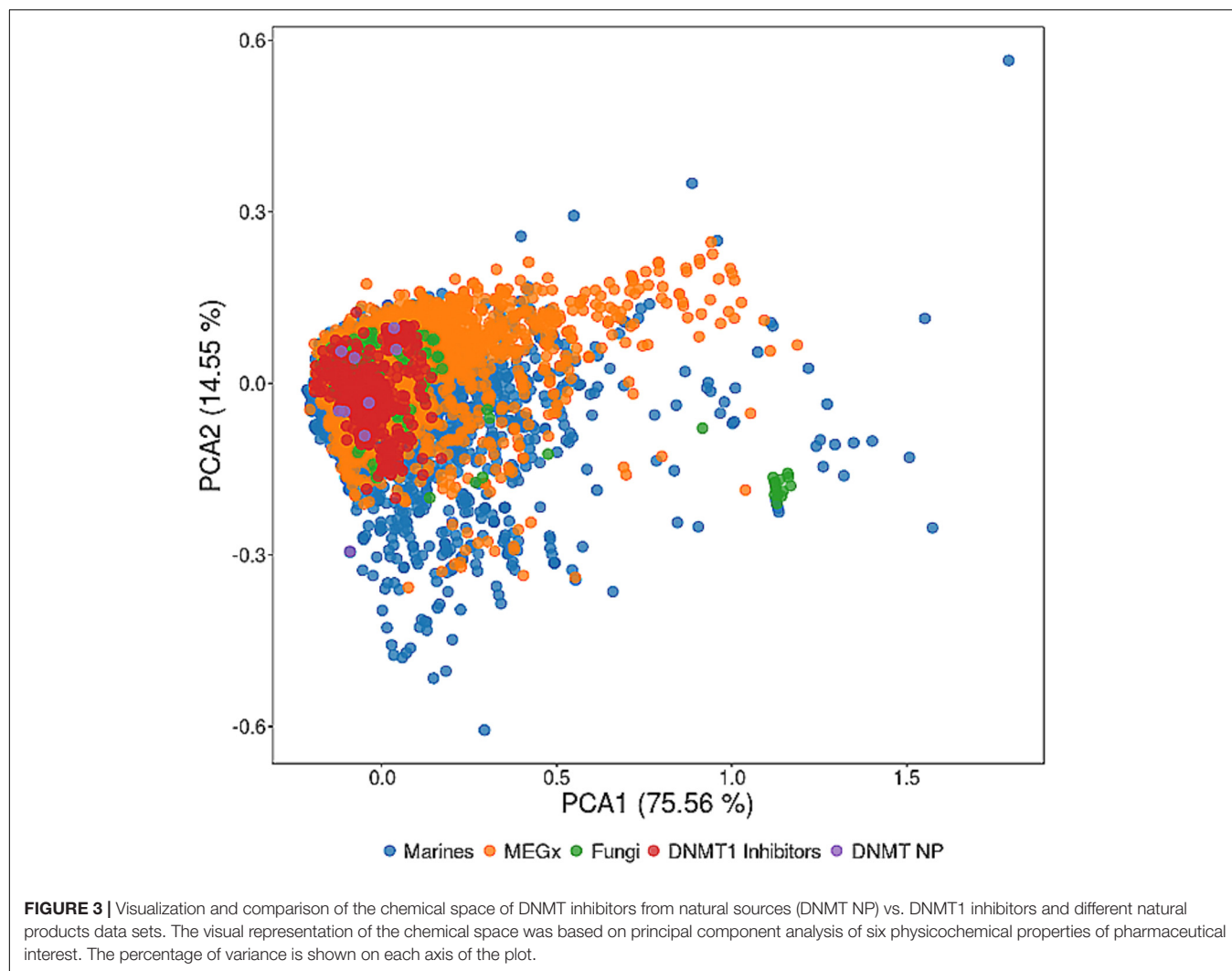
SECTION 6: OPPORTUNITIES FOR SEARCHING FOR NATURAL PRODUCTS AS DNMT INHIBITORS

Most of the DNMT inhibitors from natural sources have been identified by serendipity. As discussed in Section 5, the chemical

space of natural products and food chemicals can be explored in a systematic manner using computational approaches. A classical and general approach is using virtual screening. The main aim of virtual screening is filtering compound data sets to select a reduced number of compounds with increased probability to show biological activity. Virtual screening has proven to be useful to identify hit compounds (Clark, 2008; Lavecchia and Di Giovanni, 2013). **Table 1** summarizes representative case studies where virtual screening has led to the identification of active compounds with novel scaffolds. In other studies, virtual screening has uncovered potential active compounds but experimental validation still needs to be conducted. Examples of these studies are further discussed in the following sections.

There are several published studies of virtual screening of natural products to identify DNMT inhibitors and/or demethylating agents. In an early work, Medina-Franco et al. (2011) reported the screening of a lead-like subset of natural products available in ZINC. Authors of that work implemented a multistep virtual screening approach selecting consensus hits identified from three different docking programs. One computational hit showed DNMT1 activity in a previous study. Other candidate compounds were identified for later experimental validation (Medina-Franco et al., 2011).

In a separate work, Maldonado-Rojas et al. (2015) developed a QSAR model based on linear discriminant analysis to screen 800 natural products. Hits selected were further

**TABLE 1** | Summary of virtual screening hits as inhibitors of DNMTs.

Study	<i>In silico</i> approach	Major outcome	Reference
Structure-based screening of a lead-like subset of NP from ZINC	Cascade docking followed by a consensus approach	One computational had reported activity. Additional natural products were identified for screening.	Medina-Franco et al., 2011
Ligand- and structure-based screening of 800 NP	QSAR model based on linear discriminant analysis and consensus docking.	Six consensus hits were identified as potential inhibitors.	Maldonado-Rojas et al., 2015
Structure-based screening of 111,121 molecules.	Docking-based screening of synthetic screening compounds.	Identification of a low micromolar hit with a novel scaffold. Further similarity searching led to the identification of two more potent hits.	Chen et al., 2014
Ligand-based screening of 500 compounds.	Pharmacophore-based virtual screening.	Identification of one inhibitor of DNMT1 with activity in the low micromolar range. The hit showed some selectivity vs. DNMT3B.	Hassanzadeh et al., 2017
Structure- and ligand-based screening of 53,000 synthetic compounds.	Pharmacophore model, a Naive Bayesian classification model, and ensemble docking.	Two compounds showed DNMT1 inhibitory activity at single but low concentration of 1 μ M.	Krishna et al., 2017

NP: natural products.

docked with two crystallographic structures of human DNMT employing two docking programs. Six consensus hits were identified as potential inhibitors (Maldonado-Rojas et al., 2015).

Virtual screening of synthetic libraries has also been reported to identify active compounds with novel scaffolds and suitable for lead optimization. For instance, Chen et al. (2014) reported a docking-based virtual screening of the commercial screening compound library SPECS with 111,121 molecules (after filtering compounds with undesirable physicochemical properties). Results of that work led to the identification of a compound with a novel scaffold with low micromolar IC_{50} (10.3 μ M). Starting from the computational hit, similarity searching led to the identification of two more potent compounds.

Hassanzadeh et al. (2017) recently reported a pharmacophore-based virtual screening of a compound database with 500 compounds. The pharmacophore was generated using a ligand-based approach by superimposing a group of active nucleoside analogs. Selected hits, which are structurally related to the barbituric acid, were docked into the substrate binding site of DNMT1. One compound was identified with a novel chemical scaffold that inhibits DNMT1 in the low micromolar range ($IC_{50} = 4.1 \mu$ M). The compound also showed some selectivity on DNMT1 over DNMT3 enzymes (Hassanzadeh et al., 2017).

Krishna et al. (2017) implemented a virtual screening protocol using several structure- and ligand-based approaches. Methods included a pharmacophore model, a Naïve Bayesian classification model, and ensemble docking. Three out of ten selected compounds from a commercial library of synthetic molecules (e.g., Maybridge with 53,000 small drug-like compounds), showed DNMT1 inhibitory activity at compound concentration of 20 μ M. Two of these molecules showed activity at 1 μ M (Krishna et al., 2017).

In addition to the studies discussed above and summarized in **Table 1**, the next subsections discuss other approaches that can be explored. Case studies for each strategy are outlined briefly.

Section 6.1: Similarity-Based Virtual Screening of Natural Products

Similarity searching is a commonly used approach for identifying new hit compounds. Major goals are identifying starting points for later optimization or expand the SAR of analog series. Since similarity searching is fast it can be used to filter large chemical databases and it can be used in combination with other computational approaches such as molecular docking.

Similarity searching involves two major components: a molecular representation and a similarity coefficient. In practice, one of the most common molecular representations are two-dimensional (2D) fingerprints. A fingerprint is generally a string of zeros and ones that indicate the presence or absence of molecular features, respectively. In turn, one of the most common similarity coefficients is Tanimoto's (Bajusz et al., 2015). Full discussion of molecular representations and similarity coefficients are published elsewhere (Willett et al., 1998; Maggiora et al., 2014).

A novel approach to encode the chemical structures of data sets is the database fingerprint (DBFP) (Fernández-de Gortari et al., 2017). The rationale of DBFP is account for the most structural features encoded in bit positions of an entire data set. In principle, virtually any data set can be represented. For instance, it can be a small or large chemical database of screening compounds or a group of active compounds. DBFP can be used in visual representation of the chemical space (Naveja and Medina-Franco, 2018) and similarity searching (Fernández-de Gortari et al., 2017). More recently, this approach was further refined into the so-called statistical based database fingerprint (SB-DFP). This approach has the same underlying idea and application of DBFP. A key improvement is the approach to account for the most relevant structural features that are derived from a statistical comparison between the structural features of a data set of interest vs. a database of reference.

Section 6.2: Pharmacophore-Based

Thus far, several pharmacophore modeling studies have been conducted for inhibitors of DNMT1. Different approaches and input molecules have been used to develop these models. Most of the pharmacophore models have been employed to virtually screen chemical databases and identify novel hit compounds.

Yoo and Medina-Franco (2011) reported one of the first pharmacophore models for inhibitors of DNMT1. The model was generated based on the docking poses of 14 known inhibitors available at that time. The docking was conducted with a homology model of the catalytic domain of DNMT1. Of note, at the time of that study the crystallographic structure of human DNMT1 was not available. Known DNMT inhibitors used to develop the pharmacophore model included the natural products curcumin, parthenolide, EGCG and mahanine (Yoo and Medina-Franco, 2011). A year later was reported that trimethylaurintricarboxylic acid (**Figure 2**) showed a good agreement with this structure-based pharmacophore model. This compound is structurally related to 5,5'-methylenedisalicylic acid that has an inhibition of DNMT1 in a low micromolar range ($IC_{50} = 4.79 \mu$ M) (Yoo and Medina-Franco, 2012; Yoo et al., 2012).

More recently, as described in the first part of Section 6, Hassanzadeh et al. (2017) developed a pharmacophore model based on a ligand-based approach by 3D superimposition of active nucleoside analogs. That model was used to do virtual screening (*vide supra*). In the same year, with the aid of the Hypogen module of the software DS4.1, Krishna et al. (2017) developed a ligand-based pharmacophore model using the structures of 20 compounds obtained from the literature. The model was validated with the classification of an external set with known active and inactive compounds. The validated pharmacophore models were employed as part of a combined strategy to identify novel active molecules (Krishna et al., 2017).

Section 6.3: De novo Design

De novo design is a technique currently explored for DNMT inhibitors on a limited basis. Here we briefly outline two promising perspectives related to natural product research.

The first one is a strategy that provides a structural diversity classification of natural products scaffolds through generative topographic map algorithm implementation often so-called chemographies. Chemographies allow the visualization of the landscape distribution of the chemical space of natural products and their synthetic mimetic compounds (Miyao et al., 2015). Since chemographies could be generated from pharmacophoric features and molecular descriptors, it would be feasible to do scaffold hopping based on the structures of natural products (Rodrigues et al., 2016). The second approach is based on scaffold simplification that could be adapted to generate fragment-like natural products focused on DNMT inhibitors. This strategy reduces the molecular framework of natural products through the implementation of a scaffold tree algorithm based on rule-based decomposition of ring systems (Bajorath, 2018).

SECTION 7: CONCLUSION

Epigenetic targets are attractive to develop therapeutic strategies. DNA methyltransferases are the major enzyme family being one of the first epigenetic targets studied, in particular for the treatment of cancer. However, over the past few years, more therapeutic opportunities related to the modulation of DNMTs are emerging. Therefore, there is a growing interest in the scientific community to identify and develop small molecules that can be used as epi-drugs or epi-probes targeting DNMTs. Virtual screening has become more used in recent years to uncover natural products as inhibitors of DNMTs and/or demethylating agents. To this end, well established structure- and ligand-based virtual screening approaches are being used such as automated docking, QSAR and similarity searching. Also, novel chemoinformatic approaches are being developed. Of course, the computational methods should be validated with rigorous experiments *in vitro* and *in vivo* experiments to support their application.

Natural products have a well established history as inhibitors of DNMTs and demethylating molecules. However, most of the active natural products have been identified by serendipity. The knowledge of the three-dimensional structures of DNMTs in combination with increased *in silico* approaches and better computational resources are boosting the systematic search of bioactive molecules from natural origin. In addition, the increasing availability of natural product databases facilitates the discovery of epi-drugs and epi-probes targeting DNMTs.

SECTION 8: PERSPECTIVES

Natural products inside or outside of the traditional drug-like chemical space represent a large promise to develop novel compounds with DNMT inhibitory activity or demethylating properties. This is because the traditional chemical space is highly represented by small molecules

that over the past few years have not been very successful. A notable example in this direction is the reemergence of peptide-based drug discovery. Indeed, linear, cyclic peptides and peptidomimetics are regaining interest in drug discovery (Fosgerau and Hoffmann, 2015; Henninot et al., 2018).

Other promising an emerging avenue are the modulators of protein–protein interactions (PPIs) (Díaz-Eufracio et al., 2018). DNMTs are known to be involved in several PPIs (Díaz-Eufracio et al., 2018). Modulation of such interactions can be conveniently achieved with natural products. This is because PPIs are “difficult targets” not easily addressed by small molecules from the traditional chemical space (Villoutreix et al., 2014). In other words, since PPIs have unique features these can be approached with novel chemical libraries. Natural products collections represent excellent candidates for this purpose.

We foresee an augmented hit and led identification efforts based on natural products combining approaches such as high-throughput screening, structure-, ligand-based *in silico* screening, structure-based optimization, similarity searching, and scaffold hopping (Schneider et al., 1999). As part of the search for novel and more potent compounds is crucial to consider potential toxicity since toxicity issues play a major part in the lack of success of drug discovery projects.

DISCLAIMER

A similar version of this manuscript was deposited in a Pre-Print server on July 6, 2018. The reference is: Saldívar-González, F. I.; Gómez-García, A.; Sánchez-Cruz, N.; Ruiz-Rios, J.; Pilon-Jiménez, B. A.; Medina-Franco, J. L. Computational Approaches to Identify Natural Products as Inhibitors of DNA Methyltransferases. *Preprints* 2018, 2018070116 (doi: 10.20944/preprints201807.0116.v1).

AUTHOR CONTRIBUTIONS

All authors contributed to methodology and formal analysis. FS-G, JR-R, and BP-J contributed to data curation. AG-G, FS-G, DC-PdL, and JM-F contributed to writing-original draft preparation. AG-G, FS-G, NS-G, and JM-F contributed to writing-review and editing. AG-G, FS-G, and BP-J contributed to visualization. JM-F contributed to project administration.

FUNDING

This research was funded by *Consejo Nacional de Ciencia y Tecnología* (CONACYT, Mexico) grant number 282785, the *Programa de Apoyo a Proyectos de Investigación e Innovación Tecnológica* (PAPIIT) grant IA203718, and by *Programa de Apoyo a Proyectos para*

la Innovación y Mejoramiento de la Enseñanza (PAPIME) grant PE200118, UNAM.

ACKNOWLEDGMENTS

FS-G, AG-G, and NS-C acknowledge *Consejo Nacional de Ciencia y Tecnología* (CONACyT, Mexico) for the graduate

scholarships. DC-PdL and JR-R thanks the *Programa de Apoyo a Proyectos para la Innovación y Mejoramiento de la Enseñanza* (PAPIME) for the undergraduate scholarship. The authors also thank Chanachai Sae-Lee for providing the sequences used in **Figure 1**. They also acknowledge all current and past members of the DIFACQUIM research group for their comments and discussions that enriched this manuscript.

REFERENCES

- Bajorath, J. (2018). Improving the utility of molecular scaffolds for medicinal and computational chemistry. *Future Med. Chem.* 10, 1645–1648. doi: 10.4155/fmc-2018-0106
- Bajusz, D., Rácz, A., and Héberger, K. (2015). Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *J. Cheminform.* 7:20. doi: 10.1186/s13321-015-0069-3
- Berger, S. L., Kouzarides, T., Shiekhattar, R., and Shilatifard, A. (2009). An operational definition of epigenetics. *Genes Dev.* 23, 781–783. doi: 10.1101/gad.1787609
- Castellano, S., Kuck, D., Sala, M., Novellino, E., Lyko, F., and Sbardella, G. (2008). Constrained analogues of procaine as novel small molecule inhibitors of DNA methyltransferase-1. *J. Med. Chem.* 51, 2321–2325. doi: 10.1021/jm7015705
- Castillo-Aguilera, O., Depreux, P., Halby, L., Arimondo, P., and Goossens, L. (2017). DNA methylation targeting: the DNMT/HMT crosstalk challenge. *Biomolecules* 7:3. doi: 10.3390/biom7010003
- Chen, C. Y.-C. (2011). TCM Database@Taiwan: the world's largest traditional Chinese medicine database for drug screening *in silico*. *PLoS One* 6:e15939. doi: 10.1371/journal.pone.0015939
- Chen, S. J., Wang, Y. L., Zhou, W., Li, S. S., Peng, J. L., Shi, Z., et al. (2014). Identifying novel selective non-nucleoside DNA methyltransferase 1 inhibitors through docking-based virtual screening. *J. Med. Chem.* 57, 9028–9041. doi: 10.1021/jm501134e
- Chen, Y., de Bruyn Kops, C., and Kirchmair, J. (2017). Data resources for the computer-guided discovery of bioactive natural products. *J. Chem. Inf. Model.* 57, 2099–2111. doi: 10.1021/acs.jcim.7b00341
- Choi, S. H., Heo, K., Byun, H.-M., An, W., Lu, W., and Yang, A. S. (2011). Identification of preferential target sites for human DNA methyltransferases. *Nucleic Acids Res.* 39, 104–118. doi: 10.1093/nar/gkq774
- Clark, D. E. (2008). What has virtual screening ever done for drug discovery? *Expert Opin. Drug Discov.* 3, 841–851. doi: 10.1517/17460440802281978
- Davide, G., Sandra, A., Emily, B., Mattia, C., Marta, G., Annalisa, C., et al. (2016). Design and synthesis of N-benzoyl amino acid derivatives as DNA methylation inhibitors. *Chem. Biol. Drug Des.* 88, 664–676. doi: 10.1111/cbdd.12794
- Díaz-Eufracio, B. I., Naveja, J. J., and Medina-Franco, J. L. (2018). Chapter three - protein-protein interaction modulators for epigenetic therapies. *Adv. Protein Chem. Struct. Biol.* 110, 65–84. doi: 10.1016/bs.apcsb.2017.06.002
- Du, Q., Wang, Z., and Schramm, V. L. (2016). Human DNMT1 transition state structure. *Proc. Natl. Acad. Sci. U.S.A.* 113, 2916–2921. doi: 10.1073/pnas.1522491113
- Dueñas-González, A., Jesús Naveja, J., and Medina-Franco, J. L. (2016). Chapter 1 - Introduction of Epigenetic Targets in Drug Discovery and Current Status of Epi-Drugs and Epi-Probes, in *Epi-Informatics*. Boston, MA: Academic Press, 1–20. doi: 10.1016/B978-0-12-802808-7.00001-0
- Fernández-de Gortari, E., García-Jacas, C. R., Martínez-Mayorga, K., and Medina-Franco, J. L. (2017). Database fingerprint (DFP): an approach to represent molecular databases. *J. Cheminform.* 9:9. doi: 10.1186/s13321-017-0195-1
- Fernandez-de Gortari, E., and Medina-Franco, J. L. (2015). Epigenetic relevant chemical space: a chemoinformatic characterization of inhibitors of DNA methyltransferases. *RSC Adv.* 5, 87465–87476. doi: 10.1039/C5RA19611F
- Fosgerau, K., and Hoffmann, T. (2015). Peptide therapeutics: current status and future directions. *Drug Discov. Today* 20, 122–128. doi: 10.1016/j.drudis.2014.10.003
- Ganesan, A. (2018). Epigenetic drug discovery: a success story for cofactor interference. *Philos. Trans. R. Soc. B Biol. Sci.* 373:20170069. doi: 10.1098/rstb.2017.0069
- Gonzalez-Medina, M., Naveja, J. J., Sanchez-Cruz, N., and Medina-Franco, J. L. (2017). Open chemoinformatic resources to explore the structure, properties and chemical space of molecules. *RSC Adv.* 7, 54153–54163. doi: 10.1039/C7RA11831G
- Gordon, C. A., Hartono, S. R., and Chédin, F. (2013). Inactive DNMT3B splice variants modulate de novo DNA methylation. *PLoS One* 8:e69486. doi: 10.1371/journal.pone.0069486
- Henninot, A., Collins, J. C., and Nuss, J. M. (2018). The current state of peptide drug discovery: back to the future? *J. Med. Chem.* 61, 1382–1414. doi: 10.1021/acs.jmedchem.7b00318
- Ho, T. T., Tran, Q. T. N., and Chai, C. L. L. (2018). The polypharmacology of natural products. *Future Med. Chem.* 10, 1361–1368. doi: 10.4155/fmc-2017-0294
- Hwang, J.-Y., Aromolaran, K. A., and Zukin, R. S. (2017). The emerging field of epigenetics in neurodegeneration and neuroprotection. *Nat. Rev. Neurosci.* 18, 347–361. doi: 10.1038/nrn.2017.46
- Jeltsch, A. (2002). Beyond Watson and crick: DNA methylation and molecular enzymology of DNA methyltransferases. *ChemBioChem* 3, 274–293. doi: 10.1002/1439-7633(20020402)3:4<274::AID-CBIC274>3.0.CO;2-S
- Jurkowska, R. Z., Jurkowski, T. P., and Jeltsch, A. (2011). Structure and function of mammalian DNA methyltransferases. *ChemBioChem* 12, 206–222. doi: 10.1002/cbic.201000195
- Kabro, A., Lachance, H., Marcoux-Archambault, I., Perrier, V., Dore, V., Gros, C., et al. (2013). Preparation of phenylethylbenzamide derivatives as modulators of DNMT3 activity. *MedChemComm* 4, 1562–1570. doi: 10.1039/c3md00214d
- Klimasauskas, S., Kumar, S., Roberts, R. J., and Cheng, X. D. (1994). HHAL methyltransferase flips its target base out of the DNA helix. *Cell* 76, 357–369. doi: 10.1016/0092-8674(94)90342-5
- Krishna, S., Shukla, S., Lakra, A. D., Meeran, S. M., and Siddiqi, M. I. (2017). Identification of potent inhibitors of DNA methyltransferase 1 (DNMT1) through a pharmacophore-based virtual screening approach. *J. Mol. Graph. Model.* 75, 174–188. doi: 10.1016/j.jmgm.2017.05.014
- Lan, J., Hua, S., He, X. N., and Zhang, Y. (2010). DNA methyltransferases and methyl-binding proteins of mammals. *Acta Biochim. Biophys. Sin.* 42, 243–252. doi: 10.1093/abbs/gmq015
- Lavecchia, A., and Di Giovanni, C. (2013). Virtual screening strategies in drug discovery: a critical review. *Curr. Med. Chem.* 20, 2839–2860. doi: 10.2174/09298673113209990001
- Liu, Z., Du, J., Yan, X., Zhong, J., Cui, L., Lin, J., et al. (2018). TCMAnalyzer: a chemo- and bioinformatics web service for analyzing traditional Chinese medicine. *J. Chem. Inf. Model.* 58, 550–555. doi: 10.1021/acs.jcim.7b00549
- López-Vallejo, F., Giulianotti, M. A., Houghten, R. A., and Medina-Franco, J. L. (2012). Expanding the medicinally relevant chemical space with compound libraries. *Drug Discov. Today* 17, 718–726. doi: 10.1016/j.drudis.2012.04.001
- Lu, W., Zhang, R., Jiang, H., Zhang, H., and Luo, C. (2018). Computer-aided drug design in epigenetics. *Front. Chem.* 6:57. doi: 10.3389/fchem.2018.00057
- Lyko, F. (2017). The DNA methyltransferase family: a versatile toolkit for epigenetic regulation. *Nat. Rev. Genet.* 19, 81–92. doi: 10.1038/nrg.2017.80
- Maggiore, G., Vogt, M., Stumpfe, D., and Bajorath, J. (2014). Molecular similarity in medicinal chemistry. *J. Med. Chem.* 57, 3186–3204. doi: 10.1021/jm40141z
- Maldonado-Rojas, W., Olivero-Verbel, J., and Marrero-Ponce, Y. (2015). Computational fishing of new DNA methyltransferase inhibitors from natural products. *J. Mol. Graph. Model.* 60, 43–54. doi: 10.1016/j.jmgm.2015.04.010

- Hassanzadeh, M., Kasymov, R., Mahernia, S., Adib, M., Emperle Michael Dukatz, M., Bashtrykov, P., et al. (2017). Discovery of novel and selective DNA methyltransferase 1 inhibitors by pharmacophore and docking-based virtual screening. *ChemistrySelect* 2, 8383–8392. doi: 10.1002/slct.201701734
- Medina-Franco, J. L. (2015). "Discovery and development of lead compounds from natural sources using computational approaches," in *Evidence-Based Validation of Herbal Medicine*, ed. P. Mukherjee (New York, NY: Elsevier), 455–475. doi: 10.1016/B978-0-12-800874-4.00021-0
- Medina-Franco, J. L., López-Vallejo, F., Kuck, D., and Lyko, F. (2011). Natural products as DNA methyltransferase inhibitors: a computer-aided discovery approach. *Mol. Divers.* 15, 293–304. doi: 10.1007/s11030-010-9262-5
- Medina-Franco, J. L., Méndez-Lucio, O., Yoo, J., and Dueñas, A. (2015). Discovery and development of DNA methyltransferase inhibitors using in silico approaches. *Drug Discov. Today* 20, 569–577. doi: 10.1016/j.drudis.2014.12.007
- Miyao, T., Reker, D., Schneider, P., Funatsu, K., and Schneider, G. (2015). Chemography of natural product space. *Planta Med.* 81, 429–435. doi: 10.1055/s-0034-1396322
- Naveja, J. J., and Medina-Franco, J. L. (2015). Activity landscape sweeping: insights into the mechanism of inhibition and optimization of DNMT1 inhibitors. *RSC Adv.* 5, 63882–63895. doi: 10.1039/C5RA12339A
- Naveja, J. J., and Medina-Franco, J. L. (2018). Insights from pharmacological similarity of epigenetic targets in epipolypharmacology. *Drug Discov. Today* 23, 141–150. doi: 10.1016/j.drudis.2017.10.006
- Naveja, J. J., Rico-Hidalgo, M. P., and Medina-Franco, J. L. (2018). Analysis of a large food chemical database: chemical space, diversity, and complexity. *Food Res.* 7:993. doi: 10.12688/f1000research.15440.2
- Ntie-Kang, F., Zofou, D., Babiaka, S. B., Meudom, R., Scharfe, M., Lifongo, L. L., et al. (2013). AfroDb: a select highly potent and diverse natural product library from African medicinal plants. *PLoS One* 8:e78085. doi: 10.1371/journal.pone.0078085
- Olmedo, D. A., González-Medina, M., Gupta, M. P., and Medina-Franco, J. L. (2017). Cheminformatic characterization of natural products from panama. *Mol. Divers.* 21, 779–789. doi: 10.1007/s11030-017-9781-4
- Ostler, K. R., Davis, E. M., Payne, S. L., Gosalia, B. B., Expósito-Céspedes, J., Beau, M. M. L., et al. (2007). Cancer cells express aberrant DNMT3B transcripts encoding truncated proteins. *Oncogene* 26, 5553–5563. doi: 10.1038/sj.onc.1210351
- Pilon, A. C., Valli, M., Dametto, A. C., Pinto, M. E. F., Freire, R. T., Castro-Gamboa, I., et al. (2017). NuBBEDB: an updated database to uncover chemical and biological information from Brazilian biodiversity. *Sci. Rep.* 7:7215. doi: 10.1038/s41598-017-07451-x
- Rodrigues, T., Reker, D., Schneider, P., and Schneider, G. (2016). Counting on natural products for drug design. *Nat. Chem.* 8, 531–541. doi: 10.1038/nchem.2479
- Rosen, J., Lovgren, A., Kogej, T., Muresan, S., Gottfries, J., and Backlund, A. (2009). ChemGPS-NPWeb: chemical space navigation online. *J. Comput. Aided Mol. Des.* 23, 253–259. doi: 10.1007/s10822-008-9255-y
- Sacconay, L., Angleviel, M., Randazzo, G. M., Queiroz, M. M., Queiroz, E. F., Wolfender, J. L., et al. (2014). Computational studies on sirtuins from *Trypanosoma cruzi*: structures, conformations and interactions with phytochemicals. *PLoS Negl. Trop. Dis.* 8:e2689. doi: 10.1371/journal.pntd.0002689
- Schneider, G., Neidhart, W., Giller, T., and Schmid, G. (1999). Scaffold-hopping by topological pharmacophore search: a contribution to virtual screening. *Angew. Chem. Int. Ed.* 38, 2894–2896. doi: 10.1002/(SICI)1521-3773(19991004)38:19<2894::AID-ANIE2894>3.0.CO;2-F
- Shang, J., Hu, B., Wang, J., Zhu, F., Kang, Y., Li, D., et al. (2018). Cheminformatic insight into the differences between terrestrial and marine originated natural products. *J. Chem. Inf. Model.* 58, 1182–1193. doi: 10.1021/acs.jcim.8b00125
- Tough, D. F., Tak, P. P., Tarakhovskiy, A., and Prinjha, R. K. (2016). Epigenetic drug discovery: breaking through the immune barrier. *Nat. Rev. Drug Discov.* 15, 835–853. doi: 10.1038/nrd.2016.185
- Vilkaitis, G., Merkiene, E., Serva, S., Weinhold, E., and Klimasauskas, S. (2001). The mechanism of DNA cytosine-5 methylation - kinetic and mutational dissection of Hhai methyltransferase. *J. Biol. Chem.* 276, 20924–20934. doi: 10.1074/jbc.M101429200
- Villoutreix, B. O., Kuenemann, M. A., Poyet, J. L., et al. (2014). Drug-like protein-protein interaction modulators: challenges and opportunities for drug discovery and chemical biology. *Mol. Inf.* 33, 414–437. doi: 10.1002/minf.201400400
- Waddington, C. H. (2012). The epigenotype. *Int. J. Epidemiol.* 41, 10–13. doi: 10.1093/ije/dyr184
- Wang, L., Wang, J., Sun, S., Rodriguez, M., Yue, P., Jang, S. J., et al. (2006). A novel DNMT3B subfamily, Δ DNMT3B, is the predominant form of DNMT3B in Non-small cell lung cancer. *Int. J. Oncol.* 29, 201–207. doi: 10.3892/ijo.29.1.201
- Willett, P., Barnard, J., and Downs, G. (1998). Chemical similarity searching. *J. Chem. Inf. Comput. Sci.* 38, 983–996. doi: 10.1021/ci9800211
- Yoo, J., Kim, J. H., Robertson, K. D., and Medina-Franco, J. L. (2012). Molecular modeling of inhibitors of human DNA methyltransferase with a crystal structure: discovery of a novel DNMT1 inhibitor. *Adv. Protein Chem. Struct. Biol.* 87, 219–247. doi: 10.1016/B978-0-12-398312-1.00008-1
- Yoo, J., and Medina-Franco, J. L. (2011). Homology modeling, docking, and structure-based pharmacophore of inhibitors of DNA methyltransferase. *J. Comp. Aided Mol. Des.* 25, 555–567. doi: 10.1007/s10822-011-9441-1
- Yoo, J., and Medina-Franco, J. L. (2012). Trimethylaurintricarboxylic acid inhibits human DNA methyltransferase 1: insights from enzymatic and molecular modeling studies. *J. Mol. Model.* 18, 1583–1589. doi: 10.1007/s00894-011-1191-4
- Zhang, Z.-M., Liu, S., Lin, K., Luo, Y., Perry, J. J., Wang, Y., et al. (2015). Crystal structure of human DNA methyltransferase 1. *J. Mol. Biol.* 427, 2520–2531. doi: 10.1016/j.jmb.2015.06.001
- Zwergel, C., Valente, S., and Mai, A. (2016). DNA methyltransferases inhibitors from natural sources. *Curr. Top. Med. Chem.* 16, 680–696. doi: 10.2174/1568026615666150825141505

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2018 Saldívar-González, Gómez-García, Chávez-Ponce de León, Sánchez-Cruz, Ruiz-Rios, Pilón-Jiménez and Medina-Franco. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Journal of Cheminformatics

A general approach for retrosynthetic molecular core analysis

--Manuscript Draft--

Manuscript Number:	CHIN-D-19-00021	
Full Title:	A general approach for retrosynthetic molecular core analysis	
Article Type:	Methodology	
Funding Information:	Dirección General de Asuntos del Personal Académico, Universidad Nacional Autónoma de México (MX) ((PAPIIT) IA203718)	Dr. Jose L. Medina-Franco
Abstract:	<p>Background Scaffold analysis of compound data sets has reemerged as a chemically interpretable alternative to machine learning for chemical space and structure-activity relationships analysis. In this context, analog series-based scaffolds (ASBS) are synthetically relevant core structures that represent individual series of analogs. As an extension to ASBS, we herein introduce the development of a general conceptual framework that considers all putative cores of molecules in a compound data set, thus softening the often applied "single molecule - single scaffold" correspondence.</p> <p>Methods A putative core is here defined as any substructure of a molecule complying with two basic rules: a) the size of the core is a significant proportion of the whole molecule size, and b) the substructure can be reached from the original molecule through a succession of retrosynthesis rules. Thereafter, a bipartite network consisting of molecules and cores can be constructed for a database of chemical structures. Compounds linked to the same cores are considered analogs.</p> <p>Results We present case studies illustrating the potential of the general framework. The applications range from inter- and intra- core diversity analysis of compound data sets, structure-property relationships, and identification of analog series and ASBS.</p> <p>Conclusions and perspectives The molecule-core network herein presented is a general methodology with multiple applications in scaffold analysis. New statistical methods are envisioned that will be able to draw quantitative conclusions from these data. The code to use the method presented in this work is freely available as a Supplementary File. Follow-up applications include analog searching and core structure-property relationships analyses.</p>	
Corresponding Author:	Jose L. Medina-Franco, Ph.D. Universidad Nacional Autonoma de Mexico MEXICO	
Corresponding Author Secondary Information:		
Corresponding Author's Institution:	Universidad Nacional Autonoma de Mexico	
Corresponding Author's Secondary Institution:		
First Author:	J. Jesús Naveja, MD	
First Author Secondary Information:		
Order of Authors:	J. Jesús Naveja, MD	
	B. Angélica Pilon-Jiménez, BSc	
	Jürgen Bajorath, Ph.D.	
	Jose L. Medina-Franco, Ph.D.	

Order of Authors Secondary Information:	
Suggested Reviewers:	<p data-bbox="578 155 1500 279">Mark Johnson, Ph.D. Consultant mkallyn@mindspring.com Expert in scaffold analysis from an industrial and academic perspective.</p> <p data-bbox="578 289 1500 499">Terry Stouch, Ph.D. Consultant, Science For Solutions, LLC tstouch@gmail.com "25 years experience in drug discovery research in large pharma and biotech with specialization in drug design, molecular property prediction, molecular and biomolecular structure, computational sciences, pharmaceutical data evaluation and modeling, and scientific software design"</p> <p data-bbox="578 510 1500 604">John Van Drie, Ph.D. Consultant, Van Drie Research, LLC vandrie.john@gmail.com</p> <p data-bbox="578 615 1500 709">Noel OBoyle, Ph.D. NextMove Software Ltd noel@nextmovesoftware.com</p>
Opposed Reviewers:	

A general approach for retrosynthetic molecular core analysis

J. Jesús Naveja^{1,2*}, B. Angélica Pilon-Jiménez², Jürgen Bajorath³, José L. Medina-Franco^{2,*}

¹ PECEM, Faculty of Medicine, Universidad Nacional Autónoma de México, Avenida Universidad 3000, Mexico City 04510, Mexico.

² Department of Pharmacy, School of Chemistry, Universidad Nacional Autónoma de México, Avenida Universidad 3000, Mexico City 04510, Mexico.

³ Department of Life Science Informatics, B-IT, LIMES Program Unit Chemical Biology and Medicinal Chemistry, Rheinische Friedrich-Wilhelms-Universität, Endenicher Allee 19c, D-53115 Bonn, Germany.

* Corresponding authors: naveja@comunidad.unam.mx (JJN), medinajl@unam.mx (JLMF)

Abstract

Background: scaffold analysis of compound data sets has reemerged as a chemically interpretable alternative to machine learning for chemical space and structure-activity relationships analysis. In this context, analog series-based scaffolds (ASBS) are synthetically relevant core structures that represent individual series of analogs. As an extension to ASBS, we herein introduce the development of a general conceptual framework that considers all putative cores of molecules in a compound data set, thus softening the often applied “single molecule - single scaffold” correspondence.

Methods: a putative core is here defined as any substructure of a molecule complying with two basic rules: a) the size of the core is a significant proportion of the whole molecule size, and b) the substructure can be reached from the original molecule through a succession of retrosynthesis rules. Thereafter, a bipartite network consisting of molecules and cores can be constructed for a database of chemical structures. Compounds linked to the same cores are considered analogs.

Results: We present case studies illustrating the potential of the general framework. The applications range from inter- and intra- core diversity analysis of compound data sets, structure-property relationships, and identification of analog series and ASBS.

Conclusions and perspectives: the molecule-core network herein presented is a general methodology with multiple applications in scaffold analysis. New statistical methods are envisioned that will be able to draw quantitative conclusions from these data. The code to use the method presented in this work is freely available as a Supplementary File. Follow-up applications include analog searching and core structure-property relationships analyses.

Keywords: analog series-based scaffold, analog searching, core structure-property relationships (CSPR), RECAP, scaffold, virtual screening.

List of abbreviations

ASBS: analog series-based scaffold; CSAR: core structure-activity relationship; CSPR: core structure-property relationship; RECAP: retrosynthetic combinatorial analysis procedure; SAR: structure-activity relationship; SMILES: simplified molecular-input line-entry system; SPR: structure-property relationship.

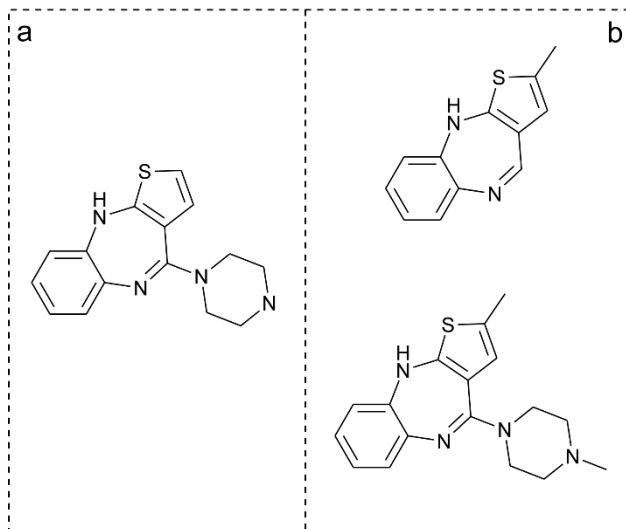
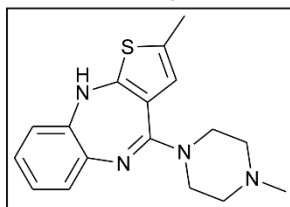
1. Background

A general trend in drug discovery through big data is emerging [1]. In this context, many exploratory analyses for finding correlations between chemical data and biological activity have been applied, often with satisfactory results [2]. Nonetheless, many of such models require simple numerical data, as opposed to the complex information enclosed in a chemical structure [3]. Chemical fingerprints, a widely applied representation for converting chemical structures into information vectors, simplify even complex structures [4]. It is common that such methods detect chemical similarity between molecules even when a synthetic chemist would struggle to find substantial structure commonalities [5].

In contrast to structural fingerprints, molecular scaffolds (and sub-structure methods in general) are alternative representations intuitively interpretable by a chemist, and scaffold analysis is a more chemically conservative approach than a computational prediction of structural resemblance [5]. Several approaches have been proposed to define and generate scaffolds in a consistent

manner [6–8]. One of the earliest and still most common scaffold concept was proposed by Bemis and Murcko [9] and is exemplified in Figure 1. Section “a” of this figure shows the Bemis and Murcko scaffolds for olanzapine and albendazole. This and other classic definitions of scaffolds consider only ring systems, a rather inconvenient feature since it is not uncommon that small rings are conceptualized as side chains or part of substituents by synthetic chemists. Considering the limitations of classical scaffolds, Bajorath et al., developed a novel scaffold concept: the analog series-based scaffold (ASBS) [10] illustrated in section “b” of Figure 1. In general, ASBS are found through a process that incorporates retrosynthetic information and restrictions in the core/molecule size ratio, thus allowing the identification of chemical analogs that can be summarized in meaningful R-group tables [11, 12]. Hence, ASBS leverage the chemical synthesis and biological relevance of scaffolds [13]. A shortcoming of the current implementation of ASBS is that it depends on the specific dataset [6]. We show below that this is a direct consequence of following the “single molecule – single scaffold” paradigm during the ASBS generation. A critical issue is that analyzing scaffold diversity or comparing scaffolds found in different datasets becomes challenging.

Olanzapine



Albendazole

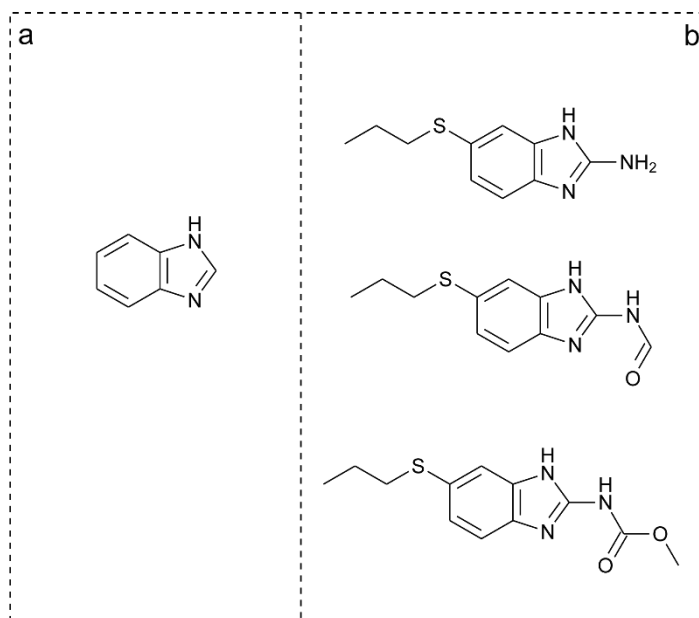
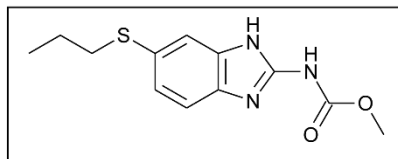


Fig. 1 Two scaffolds definitions are applied to two exemplary molecules (olanzapine and albendazole). **a)** Bemis-Murcko scaffold; **b)** Putative cores.

The goal of this work is to show how softening the “single molecule – single scaffold” paradigm can lead to consistent core results that can extend the ASBS to core diversity analysis and core-property relationships analysis. Furthermore, original ASBS can be obtained on the basis of the generalized approach. Building upon the ASBS approach, we propose a conservative yet flexible general framework able to obtain synthetically relevant cores from chemical libraries, allowing applications such as analog searching through matching of shared cores, diversity, and structure-property relationship (SPR) analyses.

This Methodology paper is organized into two major sections. First, we describe the general approach for constructing molecules-cores networks. In the second section, we illustrate the application of the method using two case studies, namely: core overlap analysis of two natural products datasets and core structure-activity relationship (CSAR) analysis of an analog series of Akt2 inhibitors. Perspectives for the methodology include, for example, chemical core diversity analysis, advanced SPR, and chemical analog searching. The approach has been used already for the identification of analog series and corresponding scaffolds [12].

2. Methods

2.1 Core definition

For any given molecule, a putative core is defined by two criteria [10], herein termed relevance and synthetic feasibility, further clarified as follows:

1. The relative size of the core as compared to the whole molecule is significant (relevance criterion),
and
2. The core is either the whole molecule or a substructure obtained from the original molecule through a series of predefined retrosynthetic steps (synthetic feasibility criterion).

These two criteria ultimately require the user’s input to be further specified. Regarding the first criterion, previous determinations of ASBS have considered a 2:1 ratio of the scaffold vs. all substituents’ atoms [10]. The second criterion requires predefining sets of retrosynthesis rules,

such as the widely used RECAP rules [14]. A user may implement other sets of available rules [15] or proprietary retrosynthetic schemes.

Importantly, given the newly proposed framework, the “single molecule – single core” paradigm underlying various scaffold definitions is no longer compulsory. On the contrary, all substructures of a molecule complying with the two criteria above are considered as putative cores, illustrated in Figure 1b for an exemplary molecule.

A direct consequence of computing putative cores for one or more datasets of molecules is analyzing the core structures in light of scaffold criteria. Major differences compared to the scaffold concept by Bemis and Murcko (Figure 1), are presented in Table 1.

Table 1. Comparison of the Bemis-Murcko scaffold and the core framework proposed in this work.

Feature	Bemis - Murcko scaffold	Core framework
Number of cores per molecule	0 or 1	1 or more
Rings can be substituents	No	Yes
Considers retrosynthesis rules	No	Yes
The core is a major component of the molecule	Yes/No	Yes

2.2 Molecules - cores network

If the core definition described above is applied to a set of compounds, a bipartite network $G = (U, V, E)$ can be drawn, where U is the set of molecules, V the set of putative cores, and E the set of edges linking molecules to their putative cores. By definition, if two molecules $u_1, u_2 \in U$ can be mapped to the same $v_1 \in V$, they are considered analogs. An example of a core network is illustrated in Figure 2, where a set of six exemplary molecules is mapped to all possible cores. Separate clusters represent series. If all compounds in a series can be mapped to a single core, then the series is an analog series, and the comprehensive core is its ASBS. It has been shown

that not all sets of related compounds form analog series applying this formalism since in some cases, no single core represents all compounds [12]. Moreover, to a pre-defined analog series represented by a single core, new molecules might be difficult to add. On the contrary, the use of expandable series with multiple cores makes it easy to include new compounds, which need only to be decomposed according to the same criteria and incorporated into the network. This is a consequence of accounting for all possible molecule-core relationships.

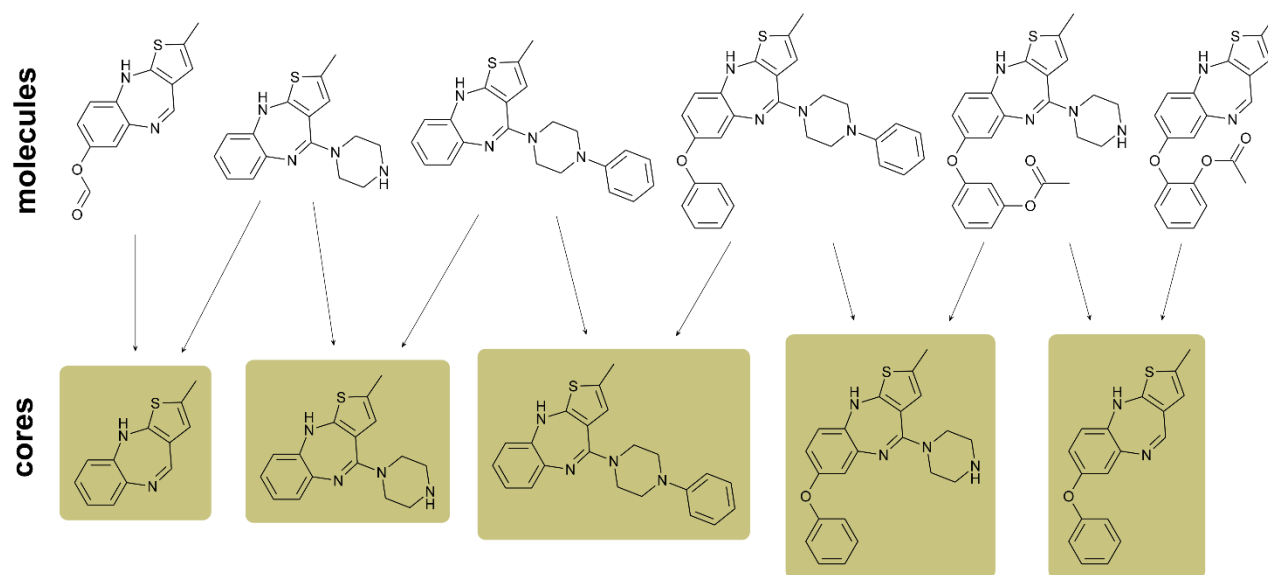


Fig. 2 Construction of a core - molecule network for an exemplary dataset. Each molecule is connected to all of its putative cores. Thus, series can be formed if at least two molecules share a core. Note that not all molecules in a series need be pairwise analogs of each other, but a sequence of analogs must exist. For this example, only putative cores mapping to more than a single molecule are included.

2.3 Computational implementation

An RDKit - Python implementation of the algorithm is made available in the Supplementary Material (see also section Availability of data and materials). The algorithm flow is depicted in Figure 3. The code is fully parallelized and runs mostly off-memory, which means it can be used to process large chemical libraries. The input is a file with molecular structures represented as SMILES strings as well as an identifier. A “washing” script was added to remove salts, retain the

largest molecular component, generate canonical SMILES, and omit stereochemistry information by default. However, stereochemistry can be retained by modifying the data preparation script. Canonical SMILES are annotated with an identifier (WID). Then, each molecule is fragmented independently, and only fragments complying with the core definition (see Methods) are saved. Unique cores are annotated with another identifier (MID). Finally, through network analysis, analog series are identified as disjoint subgraphs (clusters). The output is: 1) a file containing molecule - core associations (suffix: "cores.tsv"); 2) a file containing analog series - molecule associations (suffix: "ASW.tsv"); 3) a file containing analog series - cores associations (suffix: "ASM.tsv").

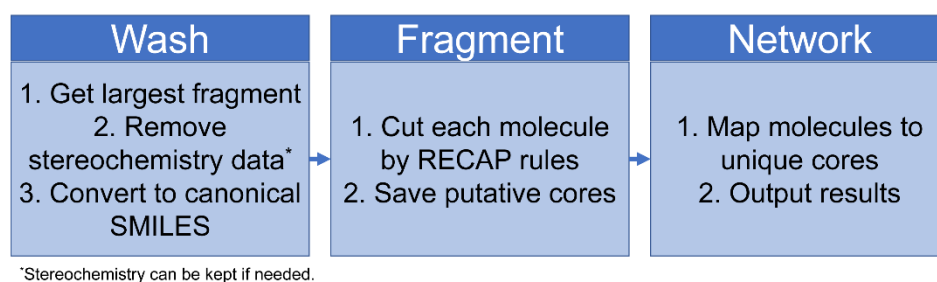


Fig. 3 Algorithm steps for the generation of core - molecule associations.

3. Results

The newly introduced framework has a number of potential applications such as structural analysis of compound databases including structural diversity analysis (based on the new cores), structure-property(-activity) relationships (SP(A)R), virtual screening focusing on "core hopping" and "core hunting" potential (in analogy to "scaffold hopping" and the "scaffold hunting" [16]). In this section, we discuss selected applications of the core framework.

3.1 Core content analysis

3.1.1 Exemplary core overlap analysis in natural product data sets

To illustrate a core overlap analysis we present an example using two publicly available natural product datasets including NuBBE_{DB} [17] and BIOFACQUIM [18], which contain information about Brazilian and Mexican natural products, respectively.

The motivation of pursuing a scaffold overlap analysis would be to identify common and unique chemotypes in these databases. As shown in Table 2, NuBBE_{DB} and BIOFACQUIM share 49 (~5%) Bemis-Murcko scaffolds and around 106 (~1%) cores. By design, the number of unique Bemis-Murcko scaffolds can only be as high as the total number of unique molecules, while this is the minimum number of cores that can be found. This explains why more cores than Bemis-Murcko scaffolds are found. Remarkably, if a core is shared between two databases, an analog series might be constructed for that core (Figure 4a). On the other hand, a shared Bemis-Murcko scaffold might not represent a meaningful analog series (Figure 4b).

Table 2. Core and Bemis-Murcko scaffold overlap of NuBBE_{DB} vs BIOFACQUIM databases.

	Measurement	BIOFACQUIM	NuBBE _{DB}	Both
	Unique molecules intraDB	399	2018	2417
	Unique molecules interDB	344	1963	2362 (55 shared)
Cores	Cores intraDB	1356	15,758	17,114
	Unique cores intraDB	1153	11,738	12,289
	Unique cores interDB	1047	11,632	12,785 (106 shared)
Bemis-Murcko scaffolds	Scaffolds intraDB	396	1921	2317
	Unique scaffolds intraDB	176	754	930
	Unique scaffolds interDB	127	705	881 (49 shared)

Similar overlap analysis can be performed with other larger natural product databases such as the Dictionary of Natural Products [19], the Universal Natural Product Data Set [20] or basically any other compound collection. Here, we illustrate the method with two natural product datasets as examples. Of note, quantitative diversity metrics remain to be developed, similar to those available to quantify scaffold diversity based on Bemis-Murcko scaffolds [21].

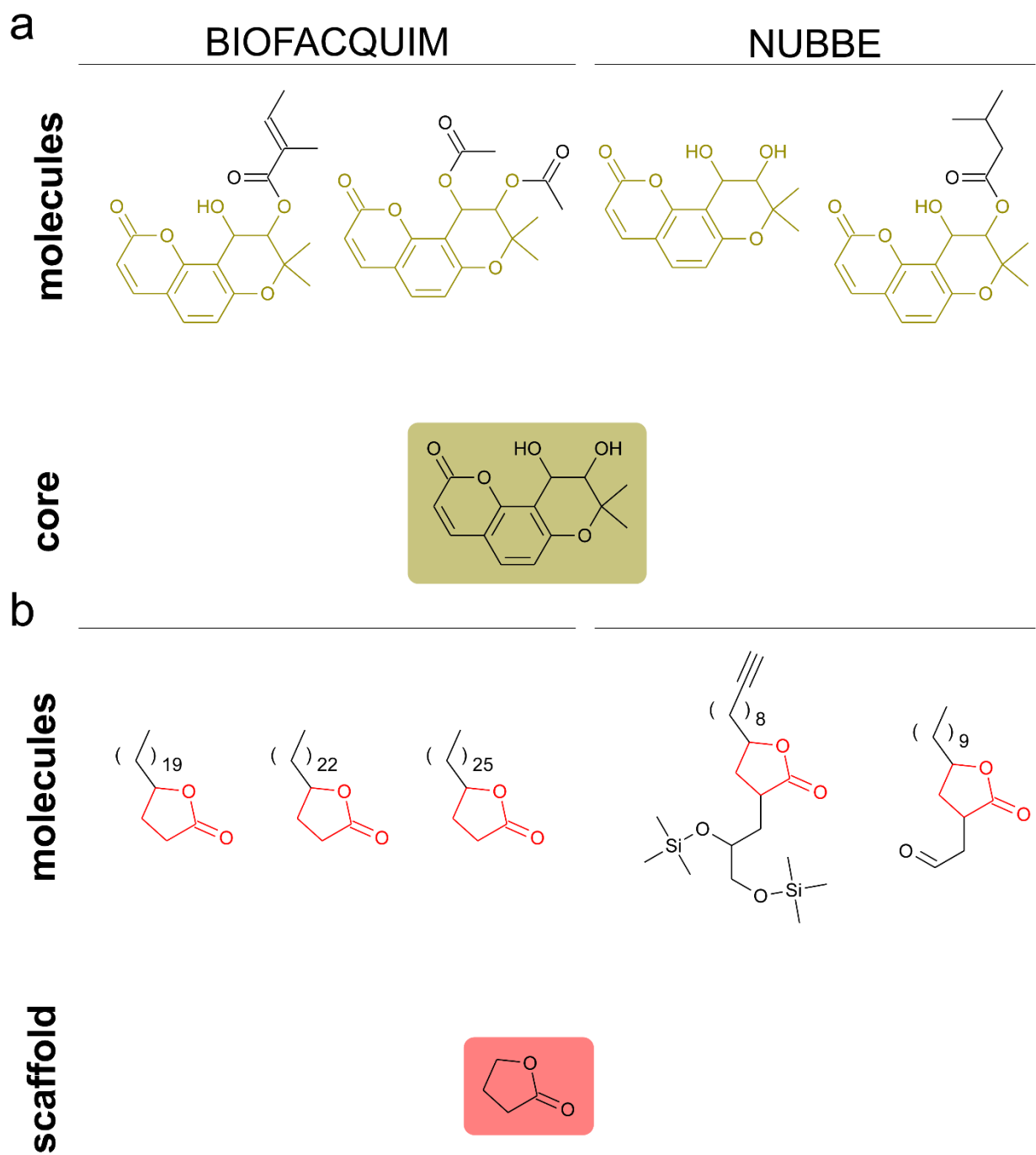


Fig. 4 Exemplary overlapping cores and scaffolds from two datasets. **a)** For any overlapping core, an analog series can be found with the core itself as its ASBS; **b)** This is not necessarily the case for overlapping Bemis-Murcko scaffolds.

3.2 Core structure-property (activity) relationship analysis: “hit-to-lead cores”

Substructure and scaffold-based representations are commonly used in many areas of chemistry. An example is R-group tables to assist in the analysis of SPRs [22, 23]. Considering cores changes the view of SPR analysis. For instance, every collection of molecules linked to a single core can be considered an analog series, for which SPR can be conducted using an R-group table. Moreover, molecules can be assigned to more than a single core. Therefore, the progression of an analog series can be readily visualized from the cores perspective (Figure 5). Analyzing a database and identifying the most relevant analog series with a given activity, can be considered “*de novo* lead discovery”. Such an approach prioritizes activity of the analog series over its size measured in the number of analogs it contains. This can be accomplished best by considering the properties in the whole molecules-cores network and then selecting enriched cores. Such cores will represent an analog series where the desired property tends to appear, plus different decorations on the scaffold retain the property. Therefore, these cores could be considered leads for drug discovery programs. We call these cores “hit-to-lead cores”, as they can also remind us of a hit in the sense that it can be found from exploratory and high-throughput drug discovery campaigns.

3.2.1 Exemplary CSAR analysis

Herein, we illustrate the application of CSAR analysis with a dataset of Akt2 inhibitors extracted from ChEMBL 24 [24, 25]. For preprocessing of the data, only compounds with reported IC₅₀ values and standard type “=” were considered. Furthermore, duplicates were removed and the maximum ChEMBL activity values were kept. The dataset was first run through the *cores.py* script (see Supplementary Materials) and the output was used for CSAR analysis. A Jupyter Notebook with the CSAR analysis is provided as a Supplementary File as well.

79 series had at least two compounds, and 24 series had at least five. The largest series contained 42 compounds. We analyzed the SAR of this series and found that only six cores were connected to more than a single compound. As shown in Figure 5a, a bipartite network is constructed, where one part of the network is the molecules and the other their putative cores.

Edges map molecules to their putative cores. In this way, for any given property, a statistical distribution can be obtained for each core through analogs mapping to the core. Also, the bipartite network allows examining the relevance of the cores. In the example shown in Figure 5a, the core labeled **M406** represents a larger subset of molecules (represented with red dots at the top of the figure). Note that the cores labeled **M807**, **M808**, **M160**, and **M161** are mapped to the same subset of molecules (Figure 5a).

The molecule-core bipartite network can be condensed to a core network representation. Figure 5b illustrates a molecule-core network taken the information from Figure 5a. The network shows the relationship of the core labeled **M406** with other five cores. An edge between two cores means that they share at least one molecule. As in Figure 5a, the dots in Figure 5b are colored by the median of the pIC_{50} of the associated molecules using a continuous color scale. The cores network shows that three subregions in the CSAR can be found. Furthermore, in this case, there is a gradient, where the most active cores (**M807** and **M808**) are connected to cores with medium activity (**M406**) but not to those with low activity (**M160** and **M161**).

Figure 5c shows a more detailed CSAR visualization for this series in Figure 5a, adding the chemical structures to the core's network and removing redundant cores by keeping only the largest. In this example, Figure 5c indicates that the four Akt2 inhibitors sharing the core **M161** with an amine substitution in the imidazopyridine ring (average $pIC_{50} = 6.51$) are less active than the two molecules having the related core **M808** but with a substituent with negative partial charges (average $pIC_{50} = 7.05$).

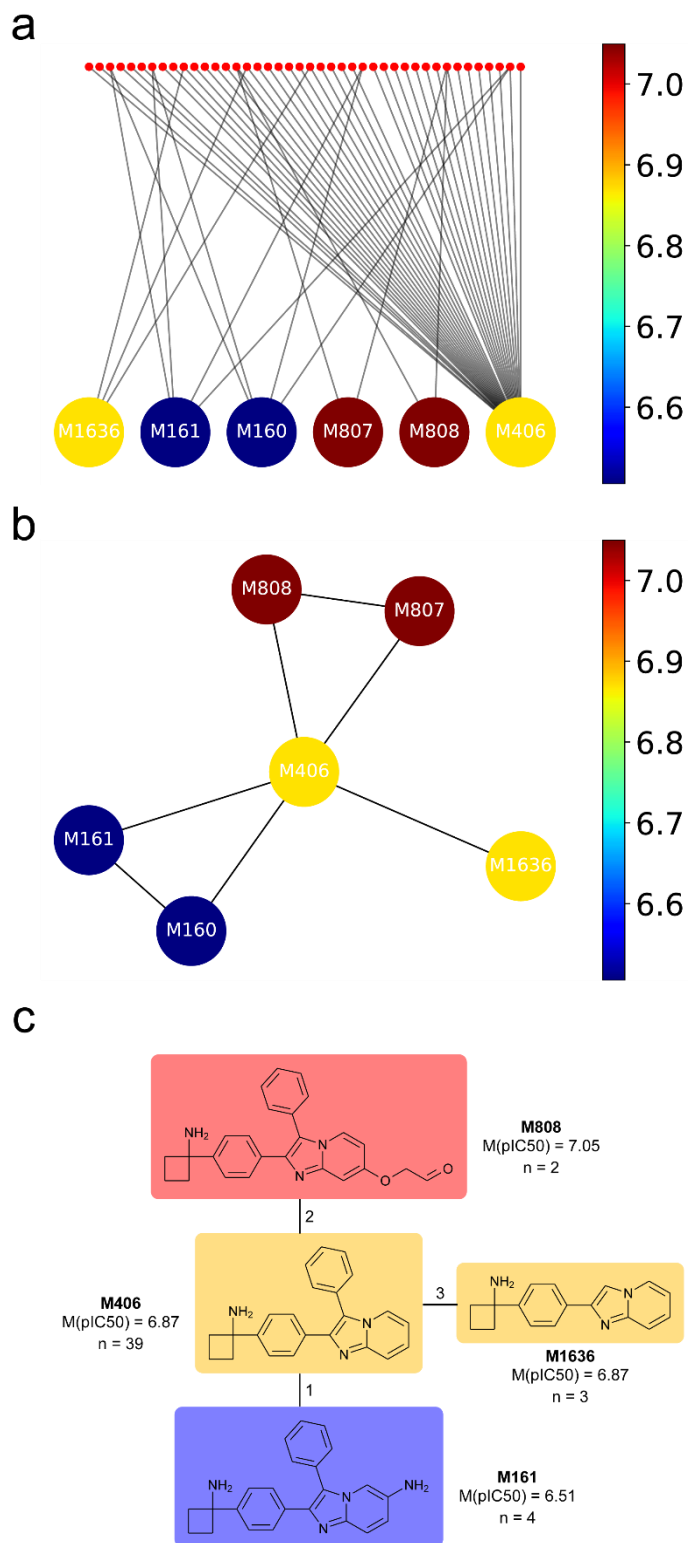


Fig. 5 Core structure-activity relationship visualization for the largest series in a dataset of Akt2 inhibitors. **a)** Molecule-core bipartite network. Molecules are shown as small red dots, while cores are represented as larger dots and colored by the median of the pIC₅₀ of the molecules represented by them. **b)** Core network obtained from the molecules-cores bipartite network. Nodes

are putative cores and edges are drawn between nodes that share at least one compound in the dataset; c) Final CSAR visualization. Redundant cores were omitted and chemical structures were added to the core's network.

3.3 Identification of analog series and corresponding scaffolds

In a recent publication, a direct application of the core framework for finding ASBS was introduced [12]. By definition, analog series must have a common scaffold and be disjoint from each other according to the paradigm of “single molecule - single scaffold” paradigm. To this end, the initial bipartite network of molecules and their putative cores can be used as a starting point. Then, the number of putative cores has to be reduced to the minimum, and subnetworks are not allowed to overlap. This can be achieved by an iterative greedy selection of cores according to which cores that are more represented in the dataset persist and disqualify secondary cores.

4. Discussion

Scaffold content and diversity analysis are common practice to explore the chemical space of compound data sets and perform classifications based on a structure representation that is highly intuitive [26–28]. There are multiple ways of defining chemical scaffolds or cores (see [29] for a comprehensive review). Of note, hierarchical scaffolds might allow each molecule to have more than a single scaffold. Nevertheless, the level a scaffold occupies in the hierarchy is arbitrary and depends on the dataset. In our general core approach, core structures are followed horizontally as they progress.

Herein, we have introduced a novel framework for performing scaffold analysis, which is an extension and generalization of the ASBS approach. Several exemplary applications of the were presented. Our approach avoids a possible information loss as a consequence of not considering all possible molecule-core relationships. Only in the context of a chemical dataset, cores can be chosen that represent as many molecules as possible. Reducing the number of cores might be feasible for SPR analysis, but not for comprehensively comparing core overlap between databases.

Among the limitations of the newly presented core framework is the often increased computational cost compared to chemical fingerprint methods or conventional scaffold analysis following Bemis and Murcko. Nonetheless, the off-memory and parallel nature of the scripts make it feasible to process a database as large as ChEMBL_24 on a desktop computer in less than 24 hours. For many applications, it is also possible to generate a library of molecules and their cores, such that only as new molecules need to be fragmented.

5. Conclusions and perspectives

In this study, a new and general method inspired by the ASBS concept is introduced. Exemplary applications are shown to establish proof-of-concept using data from medicinal and natural product chemistry. Scaffold content and diversity analysis are fundamental to characterize compound databases. The results of the recently developed definition of ASBS have proven the chemical and biological usefulness of identifying core scaffolds through retrosynthetic rules and size restrictions. Other applications include the identification of ASBS for hit identification and structure-property analysis.

Going forward, the new core framework might be systematical to analog searching and core hopping. A meaningful methodological extension would be exploring instances in which individual molecules could be reduced to more than one core.

DECLARATIONS

Availability of data and material: Source code for getting core data is provided using the free RDKit Python package as a Supplementary File. Requirements: Linux OS, an RDKit environment, packages: pandas, NetworkX, Dask. A .zip file containing a Jupyter Notebook with the exemplary CSAR analysis for the Akt2 dataset is provided as well, including the output data from the script.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was funded by the *Programa de Apoyo a Proyectos de Investigación e Innovación Tecnológica* (PAPIIT) IA203718.

Authors' contributions

All authors participated in the conception and conceptualization of the study. JJN carried out the analysis and wrote the first draft; AP participated in the scaffold overlap analyses; JLMF and JB revised the manuscript.

Acknowledgments

Helpful discussions with Martin Vogt, Dagmar Stumpfe, Filip Miljković and Swarit Jasial are much appreciated. JJN is thankful to CONACyT for the granted scholarship number 622969 and to DAAD (program 53378443).

References

1. Lusher SJ, McGuire R, van Schaik RC, Nicholson CD, de Vlieg J. Data-driven medicinal chemistry in the era of big data. *Drug Discov Today*. 2014;19:859–68. doi:10.1016/j.drudis.2013.12.004.
2. Lavecchia A. Machine-learning approaches in drug discovery: methods and applications. *Drug Discov Today*. 2015;20:318–31. doi:10.1016/j.drudis.2014.10.012.
3. Vogt M, Bajorath J. Chemoinformatics: a view of the field and current trends in method development. *Bioorg Med Chem*. 2012;20:5317–23. doi:10.1016/j.bmc.2012.03.030.
4. Lo Y-C, Rensi SE, Torng W, Altman RB. Machine learning in chemoinformatics and drug discovery. *Drug Discov Today*. 2018;23:1538–46. doi:10.1016/j.drudis.2018.05.010.
5. Bajorath J. Exploring activity cliffs from a chemoinformatics perspective. *Mol Inform*. 2014;33:438–42. doi:10.1002/minf.201400026.
6. Bajorath J. Improving the utility of molecular scaffolds for medicinal and computational

- chemistry. *Future Med Chem.* 2018;10:1645–8. doi:10.4155/fmc-2018-0106.
7. Schneider P, Schneider G. Privileged structures revisited. *Angew Chem Int Ed Engl.* 2017;56:7971–4. doi:10.1002/anie.201702816.
 8. Hu Y, Stumpfe D, Bajorath J. Lessons learned from molecular scaffold analysis. *J Chem Inf Model.* 2011;51:1742–53. doi:10.1021/ci200179y.
 9. Bemis GW, Murcko MA. The properties of known drugs. 1. Molecular frameworks. *J Med Chem.* 1996;39:2887–93. doi:10.1021/jm9602928.
 10. Stumpfe D, Dimova D, Bajorath J. Computational method for the systematic identification of analog series and key compounds representing series and their biological activity profiles. *J Med Chem.* 2016;59:7667–76. doi:10.1021/acs.jmedchem.6b00906.
 11. Dimova D, Bajorath J. Collection of analog series-based scaffolds from public compound sources. *Future Science OA.* 2018;4:FSO287. doi:10.4155/fsoa-2017-0135.
 12. Naveja JJ, Vogt M, Stumpfe D, Medina-Franco JL, Bajorath J. Systematic extraction of analogue series from large compound collections using a new computational compound–core relationship method. *ACS Omega.* 2019;4:1027–32. doi:10.1021/acsomega.8b03390.
 13. Dimova D, Stumpfe D, Hu Y, Bajorath J. Analog series-based scaffolds: computational design and exploration of a new type of molecular scaffolds for medicinal chemistry. *Future Science OA.* 2016;2:FSO149. doi:10.4155/fsoa-2016-0058.
 14. Lewell XQ, Judd DB, Watson SP, Hann MM. RECAP - Retrosynthetic combinatorial analysis procedure: a powerful new technique for identifying privileged molecular fragments with useful applications in combinatorial chemistry. *J Chem Inf Comput Sci.* 1998;38:511–22. doi:10.1021/ci970429i.
 15. Watson IA, Wang J, Nicolaou CA. A retrosynthetic analysis algorithm implementation. *J Cheminform.* 2019;11:1. doi:10.1186/s13321-018-0323-6.
 16. Schäfer T, Kriege N, Humbeck L, Klein K, Koch O, Mutzel P. Scaffold Hunter: a comprehensive visual analytics framework for drug discovery. *J Cheminform.* 2017;9:28. doi:10.1186/s13321-

017-0213-3.

17. Pilon AC, Valli M, Dametto AC, Pinto MEF, Freire RT, Castro-Gamboa I, et al. NuBBEDB: an updated database to uncover chemical and biological information from Brazilian biodiversity. *Sci Rep.* 2017;7:7215. doi:10.1038/s41598-017-07451-x.
18. Pilón-Jiménez BA, Saldívar-González FI, Díaz-Eufracio BI, Medina-Franco JL. BIOFACQUIM: A Mexican compound database of natural products. *Biomolecules.* 2019;9:31. doi:10.3390/biom9010031.
19. Taylor and Francis CP. Dictionary of Natural Products. Dictionary of Natural Products. <http://dnp.chemnetbase.com/faces/chemical/ChemicalSearch.xhtml>. Accessed 12 Feb 2019.
20. Gu J, Gui Y, Chen L, Yuan G, Lu H-Z, Xu X. Use of natural products as chemical library for drug discovery and network pharmacology. *PLoS ONE.* 2013;8:e62839. doi:10.1371/journal.pone.0062839.
21. González-Medina M, Prieto-Martínez FD, Owen JR, Medina-Franco JL. Consensus Diversity Plots: a global diversity analysis of chemical libraries. *J Cheminform.* 2016;8:63. doi:10.1186/s13321-016-0176-9.
22. Khire UR, Bankston D, Barbosa J, Brittelli DR, Caringal Y, Carlson R, et al. Omega-carboxypyridyl substituted ureas as Raf kinase inhibitors. *Bioorg Med Chem Lett.* 2004;14:783–6. doi:10.1016/j.bmcl.2003.11.041.
23. Wang M, Xu S, Wu C, Liu X, Tao H, Huang Y, et al. Design, synthesis and activity of novel sorafenib analogues bearing chalcone unit. *Bioorg Med Chem Lett.* 2016;26:5450–4. doi:10.1016/j.bmcl.2016.10.029.
24. Naveja JJ, Oviedo-Osornio CI, Trujillo-Minero NN, Medina-Franco JL. Chemoinformatics: a perspective from an academic setting in Latin America. *Mol Divers.* 2018;22:247–58. doi:10.1007/s11030-017-9802-3.
25. Gaulton A, Hersey A, Nowotka M, Bento AP, Chambers J, Mendez D, et al. The ChEMBL database in 2017. *Nucleic Acids Res.* 2017;45:D945–54. doi:10.1093/nar/gkw1074.

26. Shang J, Sun H, Liu H, Chen F, Tian S, Pan P, et al. Comparative analyses of structural features and scaffold diversity for purchasable compound libraries. *J Cheminform.* 2017;9:25. doi:10.1186/s13321-017-0212-4.
27. Koch MA, Schuffenhauer A, Scheck M, Wetzel S, Casaulta M, Odermatt A, et al. Charting biologically relevant chemical space: a structural classification of natural products (SCONP). *Proc Natl Acad Sci USA.* 2005;102:17272–7. doi:10.1073/pnas.0503647102.
28. Medina-Franco JL, Petit J, Maggiora GM. Hierarchical strategy for identifying active chemotype classes in compound databases. *Chem Biol Drug Des.* 2006;67:395–408. doi:10.1111/j.1747-0285.2006.00397.x.
29. Langdon SR, Brown N, Blagg J. Scaffold diversity of exemplified medicinal chemistry space. *J Chem Inf Model.* 2011;51:2174–85. doi:10.1021/ci2001428.

Fernanda I. Saldívar-González¹ / B. Angélica Pilon-Jiménez¹ / José L. Medina-Franco¹

Chemical space of naturally occurring compounds

¹ Department of Pharmacy, School of Chemistry, Universidad Nacional Autónoma de Mexico, Av. Universidad 3000, Mexico City 04510, Mexico, E-mail: fer.saldivarg@gmail.com, jose.medina.franco@gmail.com

Abstract:

The chemical space of naturally occurring compounds is vast and diverse. Other than biologics, naturally occurring small molecules include a large variety of compounds covering natural products from different sources such as plant, marine, and fungi, to name a few, and several food chemicals. The systematic exploration of the chemical space of naturally occurring compounds have significant implications in many areas of research including but not limited to drug discovery, nutrition, bio- and chemical diversity analysis. The exploration of the coverage and diversity of the chemical space of compound databases can be carried out in different ways. The approach will largely depend on the criteria to define the chemical space that is commonly selected based on the goals of the study. This chapter discusses major compound databases of natural products and cheminformatics strategies that have been used to characterize the chemical space of natural products. Recent exemplary studies of the chemical space of natural products from different sources and their relationships with other compounds are also discussed. We also present novel chemical descriptors and data mining approaches that are emerging to characterize the chemical space of naturally occurring compounds.

Keywords: biodiversity, BioFacQuim, cheminformatics, consensus diversity plots, drug discovery, foodinformatics, molecular diversity, natural products

DOI: 10.1515/psr-2018-0103

1 Introduction

Chemical space is a concept that helps to address questions such as: How many compounds exist? How many components can be synthesized? Similarly, the concept of chemical space is highly attached to the relationship among collections of chemical compounds. Currently, there is no single or unique criterion to define chemical space. Dobson stated that the chemical space “encompasses all possible small organic molecules, including those present in biological systems” [1]. Lipinski and Hopkins described it “as being analogous to the cosmological universe in its vastness, with chemical compounds populating space instead of stars” [2]. The concept of chemical space has become more relevant as compounds and their information increase over time [3].

For practical applications, chemical space can be used as a “tool” that helps to find associations in complex data and rapidly exploit the increasing information available for the discovery of drugs and other research areas such as food science [4]. More specific applications of the chemical space include evaluating the diversity of different data sets, exploring the relationships between compound collections and assessing the potential to cover other regions in the chemical space yet to be explored. Likewise, this tool is useful to design novel compound libraries and in the selection of compounds from existing libraries for computational and/or experimental screening [5].

To generate a visual representation of the chemical space of compound collection two main components are required: the molecular representation of the molecules to define the multidimensional descriptor space, and a visualization technique used to reduce the multidimensional space into two or three dimensions.

Descriptors based on the structure (constitution, configuration, and conformation of the molecule) or descriptors based on properties (physical, chemical and biological) can be used to represent molecules. The interpretations and predictions that can be made will depend on the type of descriptor used [6].

Regarding data visualization, there are several established methods to generate approximate representations of the chemical space [7–9]. All these methods are applicable to virtually any molecular library. However, the choice of method depends on the expected qualities of the visualization, or on the ability of the method to provide useful graphics.

Fernanda I. Saldívar-González, José L. Medina-Franco are the corresponding authors.

© 2018 Walter de Gruyter GmbH, Berlin/Boston.

In this chapter different applications of the visualization of the chemical space in the study of natural products (NPs) are discussed, as well as the cheminformatic approaches used and those that are emerging to characterize the chemical space of natural compounds. Recent exemplary studies of the chemical space of NPs products from different sources and their relationships with other compounds are also commented.

1.1 Importance of chemical space of natural products

NPs continue to be an important resource of drug discovery [10]. Despite the advent of efficient technologies such as combinatorial chemistry and high-throughput screening (HTS), over the past few years, NPs have attracted again the attention of academics and researchers focused on pharmaceutical chemistry. This is because NPs have proven to be a more promising source of drugs and novel structures than the compounds obtained by combinatorial chemistry [11]. Indeed, progress on technical advances and genomic and metabolomic approaches are largely contributing to further enhance the interest in natural product-based, drug discovery [12]. Figure 1 shows the chemical structures of approved drugs from natural origin approved in the last four years. At the time of writing, (September 2018) Migalastat (Galafold®) is the most recent small molecule drug approved by the Food and Drug Administration (FDA) of the United States which was isolated from the fungus *Streptomyces lydicus* PA-5726 [13]. It was found by Amicus Therapeutics and it is used for the treatment of Fabry disease, restoring the activity of specific mutant forms of α -galactosidase [14]. The high-cost and multi-step chemical processes to synthesize migalastat led to the development of a low-cost and sustainable process of fermentation with *Streptomyces lydicus* PA-5726.

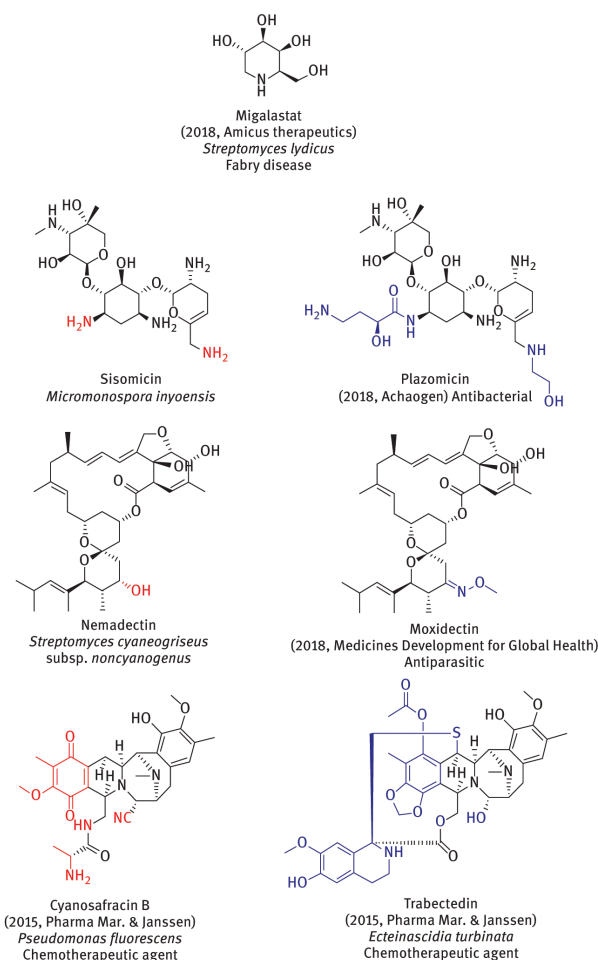


Figure 1: Chemical structures of a natural product (top) and derivatives (right) recently approved for clinical use. The source, therapeutic indication, year of approval and pharmaceutical company are indicated. Migalastat is one of the most recent isolated natural products approved by the FDA in the past four years. In blue are the modifications that were made to compounds isolated from natural products (left) and were approved by the FDA.

Plazomicin (Zemdri®) is a drug derived from a NP (Figure 1). It was developed by Achaogen as a next-generation aminoglycoside that inhibits bacterial protein synthesis. It was synthetically derived from si-

somicin, which is isolated from the aerobic fermentation of *Micromonospora inyoensis*, by appending a hydroxyaminobutyric acid substituent at position 1 and a hydroxyethyl substituent at position 6' [15].

Moxidectin (Moxidectin®) is another case of an approved drug derived from a compound previously isolated from a NP. It is a macrocyclic lactone derived from nemadectin that was isolated from *Streptomyces cyanogriseus* subsp. *noncyanogenus* but Moxidectin has the addition of a methoxime moiety at C-23 [16]. It is used in the treatment of against River blindness [17] or onchocerciasis due to *Onchocerca volvulus*. Interestingly, of the first cheminformatic analysis of natural products were focused on macrocycles and macrolides [18, 19]. In those publications, authors highlighted the importance of biologically active macrocycles in drug discovery.

Trabectedin (ET-743, Yondelis®) was developed by PharmaMar [20]. It was discovered and isolated from the Caribbean Sea squirt *Ecteinascidia turbinata*. This drug is an alkylating agent that has shown significant broad-spectrum potential as a single agent second-line drug alone or in combination particularly in the treatment of liposarcomas and leiomyosarcomas [21]. PharmaMar searched for a suitable source because of the poor yields of Trabectedin [20] obtained from Mediterranean aqua farms, plus the economic impact of the extraction and purification processes. The company developed some synthesis process but, even with these improvements, it was not suitable for manufacturing ET-743 at an industrial scale. Nevertheless, PharmaMar found a semi-synthetic process starting from cyanosafraicin B, an antibiotic obtained by fermentation from the bacteria *Pseudomonas fluorescens* obtaining good yields and it was economically profitable.

In this sense, the traditional approach to search for active compounds from NPs is being modified to take advantage of technological advances and explore the biologically relevant chemical space through chemometric approaches [22–24]. Also, studies of chemical space visualization have been shown to be useful when dealing with large libraries of potential bioactive molecules. For instance, it is estimated that about 250k NPs can be found in virtual libraries [25]. It is expected that this number will increase in the coming years and enrich the databases that can be currently searched.

So far, the characterization of the chemical space of NPs has been conducted with different approaches from which we can recognize its importance (cf. Table 1). As elaborated in the next sections of this manuscript, one of the most widespread uses has been to make comparisons between these compounds with other reference libraries such as synthetic compounds or drugs approved for clinical use. This type of comparisons had led to the conclusion that NPs have chemical structures with increased diversity and complexity as compared to other compound collections. In addition, the characterization of the chemical space of NPs can be used to classify bioactive compounds according to their biological properties. The rationale is that similar molecules have similar bioactivity. Based on this hypothesis, an important application in the area of NPs has been the selection of compounds for virtual screening and the emergence of biology-oriented synthesis (BIOS): BIOS is focused on the “islands of bioactivity” that have composite data sets containing central structures of compound classes that are biologically relevant [26, 27].

Table 1: Representative studies of the chemical space of natural products.

Library	Properties calculated	Visualization method	Analysis	Ref.
NPs, drugs, and compounds from combinatorial libraries.	Topological.	PCA	Diversity.	[41]
NPs, drugs, bioactive molecules, Lipinski's rule of five compliant, compounds from DOS and molecular fragments.	Topological and physicochemical.	Radar plots and PCA	Diversity.	[42]
Merck's NP collection, the company's sample collection, and 200 top-selling drugs of 2006.	Topological and physicochemical.	PCA	Diversity.	[43]
NPs, human metabolites, drugs, clinical candidates, and known bioactive compounds.	Topological and physicochemical.	Pie charts, scatterplots	Comparative.	[44]
NPs, bioactive, and organic drug-like molecules.	Physicochemical and structural.	PCA	Diversity.	[45]
Fragment-sized and No fragment-sized NPs.	Physicochemical and structural.	PCA, SOMs	Diversity.	[46]

NPs in UNPD and, approved drugs.	Physicochemical.	PCA, Network-based	Diversity and biological activity.	[30]
Fungi metabolites, approved anticancer drugs, approved non-anticancer drugs, compounds in clinical trials, GRAS.	Physicochemical and diversity.	PCA	Diversity and complexity.	[47]
18 virtual and 9 physical NP libraries using the DNP as an encyclopedic reference.	Physicochemical.	PCA	Diversity.	[48]
Terrestrial and marine NPs.	Scaffolds.	Tree maps method	Comparative (difference of scaffolds).	[49]

1.2 Overview of a representative cheminformatic analysis of natural products databases

1.2.1 Databases of natural products

Several public natural product databases are assembled, curated and maintained by academic and non-profit groups. Examples are the Traditional Chinese Medicine TCM database@Taiwan [28, 29] and the Universal Natural Product Database [30]. Other compound databases focused on different geographical regions of the globe have been developed. Hereunder are described as representative examples.

AfroDb [31], developed by Nitte-Kang et al., is a major initiative that put together a subset of compounds representative of the African medicinal plants containing around 1,000 three-dimensional structures. The same group developed the ConMedNP collection [32], an extension of the previously published database CamMedNP. The augmented library ConMedNP is a compilation of 3,177 compounds from the Central African flora. NuBBE_{DB} is a database of compounds from Brazilian biodiversity [33]. NuBBE_{DB} presently contains data of 2218 compounds, mainly from plants, marine organisms, and fungi [34] comprising compounds from species from all six Brazilian biomes [35].

In Mexico, an emerging in-house compound database of natural products is being assembled by an academic group putting together natural products reported over the past 10 years by the School of Chemistry of the National Autonomous University of Mexico (UNAM). At the time of writing (September 2018) the in-house collection herein referred as BioFacQuim, has 423 compounds mostly isolated from plants and fungi. In this set 316 compounds were isolated from 49 different genus of plants, 98 compounds were isolated from 19 genus of fungi and 9 compounds were isolated from Mexican propolis (sticky dark-colored hive product collected by bees from living plant sources). BioFacQuim contains the compound name, SMILES, reference, kingdom (Plantae or Fungi), genus, and species of the natural product. This collection would be part of D-TOOLS [36]. Other compound datasets of natural products have been made accessible as supporting information of peer-reviewed publications or can be requested from the authors (cf. Table 2).

Table 2: Data sets included in the visual representation of the chemical space.

Database	Size	Reference
Cyanobacteria metabolites	473	In-house
Fungi metabolites	206	[47]
Marines	6253	[64]
Semi-synthetics (NATx)	26,318	ac-discovery.com
Drug approved	1806	www.drugbank.ca

1.2.2 Cheminformatic analysis of databases of natural products

As discussed in Section 1.1., the chemical space is a multidimensional space of descriptors, which can be measured experimentally or calculated *in silico*. Multidimensional data mining tools that are available to handle

this information are hierarchical clustering, decision trees, multidimensional scaling, genetic algorithms, neural networks, and support vector machines [37]. However, to navigate through the chemical space, it is required to use methods that allow projecting this multidimensional into a lower dimensional space to create graphics susceptible to visual inspection and analysis. So far, common visualization methods to represent chemical space are principal component analysis (PCA) and self-organizing maps (SOMs), also known as Kohonen networks. Other visualization approaches are multi-fusion similarity (MFS) maps, radar plots, Sammon mapping, activity-seeded structure-based clustering, singular value decomposition, minimal spanning tree, k-means clustering, generative topographic mapping (GTM) [38], hierarchical GTM [39] and the recently developed ChemMaps [40].

Table 1 summarizes representative examples of studies of the chemical space covered by compound databases from NPs. Some of these examples are discussed in the next sections.

1.2.3 Comparison with other compounds collections

The technique most used to study the chemical space of NPs has been PCA. Several studies reported thus far are focused on diversity analysis of which very valuable conclusions or interpretations have been obtained. For example, it has been observed that NPs have a large chemical diversity and their chemical space is clearly distinguishable from the space of synthetic compounds [41, 42, 50]. Similarly, several collections of NPs such as the Dictionary of Natural Products (DNP) and the Universal Natural Product Database (UNPD) cover a much broader region of chemical space than synthesized compounds or than approved drugs [43, 48, 50]. However, when NPs are compared to collections of bioactive compounds or approved drugs, many NPs have approximately the same coverage of chemical space [30, 41, 42, 50]. These results encourage the continued use of libraries of NPs to identify bioactive compounds for later development or optimization.

1.2.4 Analysis of different sources of natural products

The differences in the coverage of the chemical space of NPs according to the origin of the compounds have also been evaluated. For example, Muigg et al. [51] compared the chemical space regions of NPs collected from marine and terrestrial organisms with that of synthetic compounds. They found clear differences in the regions of the chemical space covered by the compounds of these three origins. For example, NPs extracted from marine organisms tend to be large and very flexible compared to synthetic compounds. In contrast, NPs that originate from terrestrial organisms are often large and rigid. Recently, Shang et al. [49] analyzed the chemical space covered by natural marine products. They discovered that long chains and macrocyclic structures are more prominent among marine natural products than on terrestrial ones. Similarly, Saldívar-González et al. [52] reported a comparison of NPs from marine sources, cyanobacteria and fungi metabolites. It was concluded that metabolites from cyanobacteria are remarkable for their high structural complexity and distinct profile based on molecular properties and sub-structural alerts that are different from other NPs.

1.3 Novel cheminformatic approaches to navigate the chemical space of natural products

In order to navigate efficiently the chemical space and the biological space associated with NPs, different research groups have developed methods toward the rapid identification of structure-activity relationships, easier navigation through the space, and facilitate the identification of new classes of compounds with a desired biological activity. To this end, ChemGPS-NP [53] was one of the first chemographic models used to describe in a global manner the physicochemical properties of NPs and has been shown to be useful for a variety of applications [51, 54, 55]. A web-based service of ChemGPS-NP model was developed to facilitate its use [56]. ChemGPS-NP is a PCA-based model of physicochemical property space, defined by training-set of carefully selected compounds acting as 'satellites'. In this model, the compound can be predicted to position and evaluated on a very large scale using PCA score prediction. A recent application of ChemGPS-NP was the assessment of datasets, characterizing their neighborhoods and then interpreting the map using the prediction of a control set of compounds with activities determined experimentally [57].

Other approaches for the visual representation of the chemical space are focused on representing the molecules beyond data points in a map to enhance the interpretation of the data. Examples of this approach are *Molecule Cloud* or *Scaffold tree*. *Molecule Cloud* allows the visual representation of the most common structural characteristics of chemical databases in the form of a cloud diagram [58]. *Scaffold tree* is a method for classifying molecules based on their scaffolds. In this approach, the molecules are converted to their frameworks,

then the rings are removed one by one according to a set of predefined rules, creating a hierarchy of scaffolds [59]. This approach is reminiscent of an earlier scaffold-based classification of compound data sets based on the scaffolds generated by Xu and Johnson [60].

Researchers in Dortmund, in cooperation with Novartis, investigated the scaffold content of NPs and then classified them hierarchically, by size and complexity, in a tree-like diagram. This structural classification of NPs (SCONP) [61] allows an easy and intuitive navigation in the universe of scaffolds for the identification of new regions of interest for the development of libraries inspired by NPs.

Another useful program for the analysis and visualization of scaffolds is Scaffold Hunter [62]. Unlike Scaffold tree, Scaffold Hunter allows including bioactivity data to identify new classes of scaffolds and compounds endowed with the desired activity. Using this tool and a fingerprint analysis Tao et al. [63] analyzed the distribution profiles of the natural product lead of drugs (NPLD) in chemical space, this in order to obtain useful clues to prioritize the efforts in the study of NPs. Useful information regarding the mechanisms that partially contribute to the formation of these profiles was identified. In particular, the trend of NPLD to join preferentially to privileged target sites is influenced collectively by potent binding to the target-sites and such additional factors as the optimization potential to reach the drug sweet spots in the chemical space with more adequate metabolic stability, metabolite safety, absorption, and physical forms.

2 Coverage of natural product chemical space from diverse sources

2.1 Plants, fungi, cyanobacteria, and marine

It is generally accepted that NPs are compounds with large diversity and structural complexity, however, these properties can vary according to the source of the compounds. In a previous study, Ertl et al. [50] classified the NPs of the DNP according to their origin, with the purpose of analyzing systems of rings that are typical according to the source of the compounds. However, the chemical space according to this classification was not discussed in detail.

Figure 2 shows an example of the visualization of the chemical space of NPs according to the origin of the compounds, which includes a database of semi-synthesis compounds and a database of approved drugs by the FDA as a reference (Table 2).

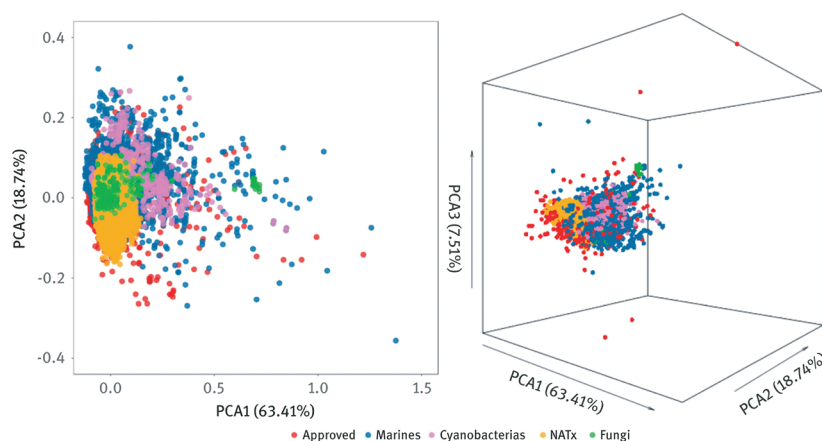


Figure 2: 2D and 3D visual representation of the chemical space of natural products databases of different origin (Table 2). The visual representation was generated with a principal component analysis of six physicochemical properties: molecular weight, hydrogen bond donors, hydrogen bond acceptors, the octanol and/or water partition coefficient, topological polar surface area and number of rotatable bonds.

As shown in Figure 2 the database of approved drugs covers mainly the chemical space and is also the database with the largest diversity in physicochemical properties. In contrast, the database of semi-synthetic products occupies a more restricted space, which in turn is included within the space of approved drugs. In this visualization can be observed that, in general, the NPs considered in this study occupy a space of traditional physicochemical properties, so that their study can lead to identify new compounds with possible therapeutic activity. Similarly, it is also observed that some of the compounds present in NPs collections occupy regions of the chemical space not yet explored and may be useful in virtual screening studies for therapeutic targets in which molecules with activity have not yet been found. Indeed, the approved drug that appears as outlier in the visual representation of the chemical space is trabectedin, the natural product that was recently approved as drug (Figure 1).

2.2 Comparison of natural products with themselves and other reference libraries

As in the case of maritime navigation where different tools or points of references are used such as the position of the stars to orient themselves; in the study of chemical space, it is sometimes useful to map the chemical space not in absolute coordinates (properties) but in relation to a set of reference points (molecules with well-defined properties).

The selection of reference libraries will always depend on the case study. However, there are collections that have been used frequently as a point of reference. Such is the case of the set of approved drugs, useful for determining the space of “safe” compounds with good physicochemical properties. Collections from synthesis have also been very useful to map the space feasible synthetically. Recently, libraries of NPs have been used as a reference in the design and development of libraries of NP-likeness compounds.

2.2.1 Commercial screening libraries of natural products

Currently, there is a large number of NP libraries, most of them can be accessed for free, for example, for virtual screening [25]. In a recent work, Chen et al. [48] analyzed the information available on the content, coverage, and relevance of individual libraries of natural products. In total 18 virtual databases (including the DNP), 9 physical libraries and the Protein Data Bank (PDB) were analyzed. Among the most important results of this study is the fact that the chemical space covered by the known NPs is substantially larger than that of readily obtainable NPs and drugs. DNP and UNPD are the known NP sets that clearly cover the larger regions of the chemical space. However, readily obtainable NPs are highly diverse (representing more than 5700 different Murcko scaffolds) and are accumulated in densely populated areas with approved drugs. With respect to individual physicochemical properties, readily obtainable NPs are substantially smaller than known NPs and drugs. In properties such as logP, the number of chiral centers and number of Csp³ readily obtainable NPs and drugs are comparable. Distinctive features of the individual databases were also observed. That analysis provided a complete and detailed overview of the known and easily obtainable NPs, which will undoubtedly be useful in the selection of data sources for computer-guided drug discovery.

2.2.2 Natural products libraries versus synthetic compounds

One of the main interests in NP-based drug discovery is differentiating compounds coming from nature with compounds obtained by synthesis. In this sense, several studies confirm that chemical space of NPs is clearly distinguishable from the space of synthesis compounds [41, 42, 50]. In comparison with synthetic molecules similar to drugs, NPs stand out for their enormous structural and physicochemical diversity [61, 65, 66]. In addition, NPs have been shown a larger structural complexity, in particular with respect to the stereochemical aspects [67].

Regions that share NPs with synthetic compounds are of particular interest for drug design. The regions in chemical space where both collections overlap may be of interest for the design of NP-inspired compounds [45].

2.2.3 Natural product libraries versus approved drugs

Collections of drugs and other databases containing information of the bioactivity profile against one or multiple biological endpoints are useful for multiple applications, including the systematic description of structure-activity relationships (also called “activity landscape modeling”) [68] and the further understanding of polypharmacology [69]. These collections are also very useful to try to delineate the boundaries of the currently explored medicinally relevant chemical space [70]. Navigation guided by the bioactivity of the chemical space allows focusing the design of libraries, which it is also known as BIOS [71].

Some prominent examples of public databases annotated with biological activity are PubChem [72], ChEMBL [73, 74], and DrugBank [75] that contain approved drugs for clinical use.

Since NPs exhibit a broad range of biological activities in different organisms, this based on their specific biological purposes in evolution, it is not surprising that drugs and NPs share parts of the chemical space. However, there is also a large part of NPs that occupy parts of the chemical space not explored yet by the available detection collections and that at the same time, are adhering to a great extent to the rule of the five [41, 76, 77].

2.2.4 Natural product libraries versus food chemicals

Food chemicals are a cornerstone in the food industry. Their study represents a step beyond the emerging field of “Food Informatics” [78]. In studies that directly compare the chemical structures of food chemicals with collections of natural products reveal that food chemicals have a high structural diversity, comparable to that of NPs and other reference libraries [79]. Likewise, there is a large overlap between the chemical space of food chemical products and NPs [4], this is somehow expected because several NPs are currently used as dietary sources.

3 Diversity in chemical space: quantification and implications

As discussed so far, the visualization of the chemical space has many applications. One of them is to determine the diversity of databases, however, it can be complemented with other cheminformatic methods for a more quantitative analysis, which provides information to prioritize the selection of libraries or sub-libraries for experimental selection. The diversity analysis helps to evaluate the structural novelty of a compound collection [80].

If the purpose of a selection project is to identify new lead compounds, then it is desirable to select collections with chemically diverse structures to increase the probability of identifying new compounds that can become leads. However, if the purpose of the campaign is to optimize one or more specific chemical scaffolds, then it is desirable to explore dense regions of the chemical space [81].

Approaches to assess the diversity can be divided into two parts. One is the analysis of structural diversity that encodes information of the structure based on fingerprints, pharmacophoric characteristics or definitions of sub-structures [82]. The second part is the analysis of chemical diversity that encodes information of macroscopic chemical properties (e.g., solubility, logP) or calculated energies, among others.

The structural diversity of NP databases using structural fingerprints and molecular scaffolds has been reported and reviewed in several papers [50, 61, 83, 84].

3.1 Molecular properties

Molecular descriptors capture information of the whole molecule and are usually straightforward to interpret. Physicochemical properties frequently used to describe chemical libraries include molecular weight (MW), number of rotatable bonds (RBs), hydrogen-bond acceptors (HBAs), hydrogen-bond donors (HBDs), topological polar surface area (TPSA), and the octanol/water partition coefficient (SlogP), which are properties commonly used as descriptors to represent lead-like, drug-like, or medically relevant chemical spaces [85, 86].

To illustrate a visual representation of the property-based chemical space Figure 3 depicts the comparison of 423 compounds from the current version of BioFacQuim (*vide supra*) with 2,214 compounds in NuBBE_{DB} [34], 885 AfroDb [31] and a subset of 3000 compounds of TCM [28]. The NP databases are compared to a collection of 1806 drugs approved for clinical use obtained from DrugBank [75]. The first two principal components capture 85.48% of the variance. For the first PC, the larger loadings correspond to SlogP followed by RB, whereas for the second PC the largest loading corresponds to HBD followed by TPSA. The visual representation of the chemical space indicates that many the NPs occupy the same space as the already approved drugs while other compounds, such as some TCM's and BioFacQuim's compounds cover neglected regions of the drug-ADME space. As commented above, interestingly, one of the main outliers in the visual representation of the chemical space in Figure 3 is the natural product recently approved as drug trabectedin (Figure 1).

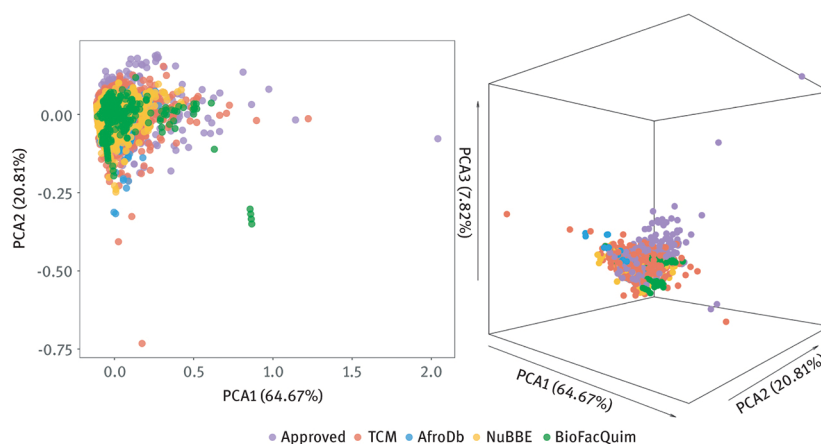


Figure 3: 2D and 3D visual representation of the chemical space of NP databases by geographical context. The visual representation was generated with a principal component analysis of six physicochemical properties: molecular weight, hydrogen bond donors, hydrogen bond acceptors, the octanol and/or water partition coefficient, topological polar surface area and number of rotatable bonds.

3.2 Structural fingerprints

Many structural features escape the very general information obtained with physicochemical and complexity descriptors. Molecular fingerprints are widely used and have been successfully applied to a number of cheminformatics and computer-aided drug design applications [87]. Fingerprints are especially useful for similarity calculations, such as database searching or clustering, generally measuring similarity with the Tanimoto coefficient. A disadvantage of some fingerprints is that they are difficult to interpret. Also, it is well-known that chemical space will depend on the types of fingerprints used. Using multiple fingerprints and representations to derive consensus conclusions have been proposed as a solution. NP databases have been largely analyzed using structural fingerprints. Representative examples are in Table 1.

3.3 Scaffolds

A complementary approach to characterize compound databases is through molecular scaffolds or ‘chemotypes’ i.e., the core structure of a molecule [88]. Same as physicochemical properties, molecular scaffolds are easy to interpret and facilitate the communication with a scientist working in different disciplines. For instance, this representation is associated with the concepts of “scaffold hopping” [89] and “privileged structures” [90]. Scaffold content analysis is broadly used to compare compound databases, to identify novel scaffolds in a compound library, to evaluate the performance of virtual screening approaches, and to analyze the structure-activity relationships of sets of molecules with measured activity.

Measuring and comparing the scaffold diversity of compound collections depends on several factors including the specific approach to describe the scaffolds, the size of the database, and the distribution of the molecules in those scaffold classes. Often, scaffold diversity is measured based on frequency counts. While these measures are correct in the way they are defined they do not provide sufficient information concerning the specific distribution of the molecules across the different scaffolds, particularly the most populated ones. Medina-Franco et al. [91] proposed the use of an entropy-based metric to measure the distribution of the molecules across different scaffolds, particularly the most populated ones, as a complementary metric for the comprehensive scaffold diversity analysis of compound data sets.

To illustrate the scaffold content and diversity analysis, Figure 4 shows the ten most frequent scaffolds found in NuBBE_{DB} (2,214 compounds in total, *vide supra*) and in BioFacQuim (423 compounds, *vide supra*). The recovery percentage of the cyclic systems are 31.5% for BioFacQuim and 19.70% for NuBBE_{DB}. Comparing the 10 most frequent scaffolds and the recovery percentage, the current version of BioFacQuim is less diverse than the NuBBE_{DB}, even though that 4.9% of NuBBE_{DB} compounds are acyclic. This result is not that surprising due to the larger number of NPs and sources gathered thus far to build the current version of NuBBE_{DB} [34].

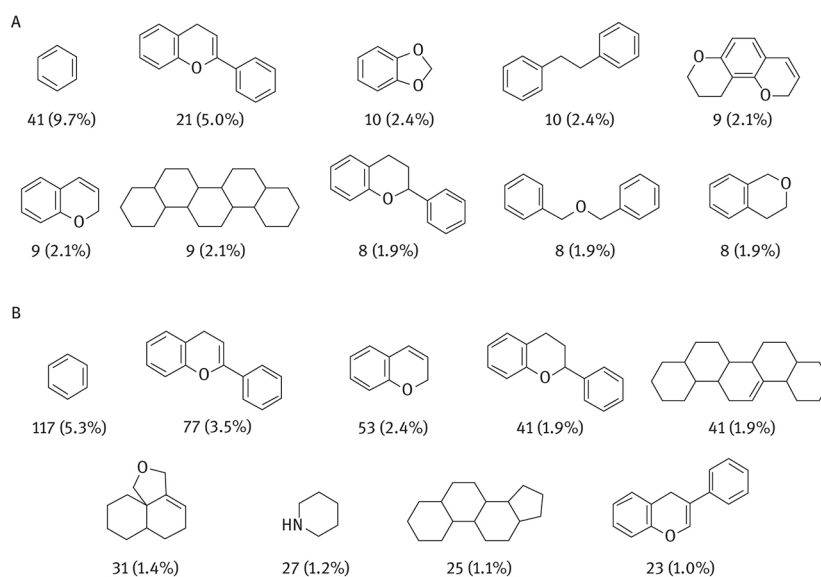


Figure 4: Most frequent cyclic systems scaffolds (Murcko scaffolds) found in (A) BioFacQuim (B) NuBBE_{DB}. The frequency and percentage are shown. Cyclic systems shown recover 31.5% and 19.70%, respectively. The second most frequent system for NuBBE_{DB} was acyclic compounds with 4.5% recovered.

Another criterion to compare the scaffold diversity of the databases is the Scaled Shannon Entropy (SSE) [91]. It is used as a measure of the specific distribution of molecules in the most populated scaffolds in a compound database. SSE values closer to 1.0 indicate that the molecules are more equally distributed in the scaffolds (high diversity) and smaller SSE values indicate that most of the molecules are distributed in fewer scaffolds (low diversity). The NuBBE_{DB} SSE is 0.931 being a little larger than the SSE of BioFacQuim which is 0.911. Based on these two metrics, it can be concluded that the current version of NuBBE_{DB} is slightly more diverse than BioFacQuim.

3.4 Consensus diversity analysis: consensus diversity plots

Since, as commented above, chemical diversity and our perception of chemical space depend on the molecular representation, an intuitive two-dimensional graph so-called Consensus Diversity (CD) Plot, was developed to represent in low dimensions the diversity of chemical libraries considering simultaneously multiple molecular representations [92]. CD Plots have already been used to characterize the global diversity of fungi metabolites [93] and NPs from Panama [94]. More recently, CD Plots were employed to compare the diversity of 23,883 food chemicals with drugs approved for clinical use, Generally Regarded as Safe molecules and screening compounds from ZINC [4]. Figure 5 shows a CD Plot comparing the global diversity of NuBBE_{DB}, BioFacQuim, TCM, AfroDb and approved drugs. BioFacQuim and AfroDb are found in the bottom right quadrant (yellow area), which indicates that the scaffolds of the molecules are the main factor that contributes to the diversity, having a relatively low diversity by fingerprints. The CD Plots also indicate that AfroDb is more diverse than BioFacQuim according to the Euclidean distance between its properties. In contrast, the approved drugs have an average diversity by fingerprints and relatively low diversity by scaffolds which indicates that the chemical characteristics of the whole molecule and/or the side chains contribute significantly to the diversity, although it is not very diverse in terms of its physicochemical properties. The area where TCM is located (bottom left quadrant) indicates that the library has the relative largest fingerprint and scaffold diversity of the data sets that are being compared. In contrast, NuBBE_{DB} (at the top-right quadrant) is the data set with the relative lowest fingerprint and scaffold diversity despite the fact it has a high intra-dataset diversity of its physicochemical properties.

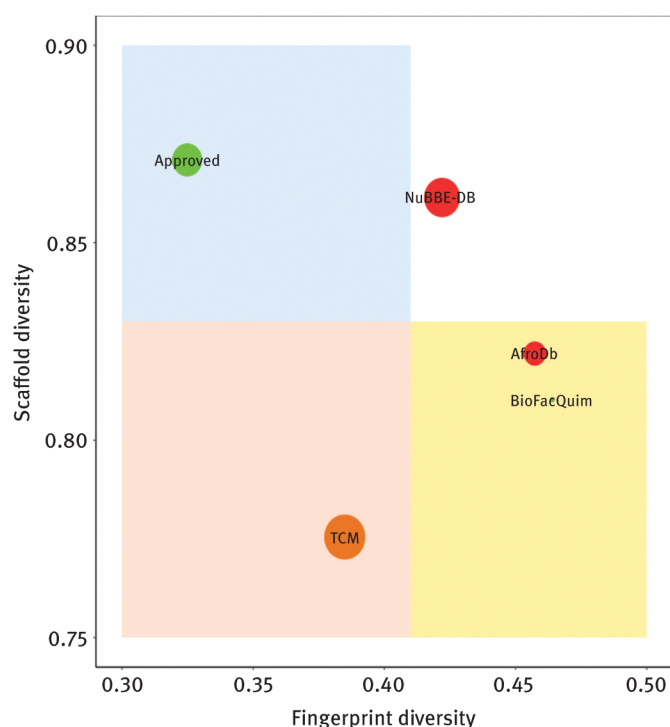


Figure 5: Consensus Diversity Plot comparing the diversity of BioFaeQuim, AfroDb, TCM, NuBBE_{DB} and approved drugs. The median Tanimoto coefficient of MACCS keys fingerprints represent the structural diversity (fingerprint diversity) of each data set and is plotted on the X-axis. The scaffold diversity of each database was defined as the area under the curve (AUC) of the respective scaffold recovery curves, and it is represented on the Y-axis. The quadrants are colored as follows: red, indicates the library is diverse considering its scaffolds and/or side chains; white, the library is not diverse; blue, the library is diverse if the chemical features of the entire molecule are considered and/or side chains contribute significantly to the diversity; yellow, the scaffolds of the molecules are the main factor contributing to the diversity and/or this set contains mostly rings with few side chains. Data points are colored by the diversity of the physicochemical properties of the data set as measured by the Euclidean distance of six properties of pharmaceutical relevance. The distance is represented with a continuous color scale from red (more diverse) to orange/brown (intermediate diversity) to green (less diverse). The relative size of the data set is represented with the size of the data point: smaller data points indicate compound data sets with fewer molecules.

4 Conclusions and future directions

NPs provide an evolutionary validated useful starting point for the design of new bioactive molecules. Several academic groups, not-for-profit, and commercial initiatives are integrating resources of NPs including compound databases. It is anticipated that more academic groups keep integrating their resources. The wealth of NPs in several countries has also motivated the interest in systematically explore the coverage of the chemical space of these collections and analyze systematically the chemical and structural diversity.

Thus far, it has been quantified distinct features of NPs such as their molecular scaffolds, property diversity, and large structural complexity. The collections of NPs would keep promoting local and global initiatives to identify bioactive compounds and potentially enrich the medicinally relevant chemical space. Furthermore, if the screening data of NPs is made accessible, in preference in the public domain, the large structure-activity information for natural compounds would facilitate the development of algorithms and eventually construct predictive models using machine learning.

Acknowledgements

This work was supported by the National Council of Science and Technology (CONACyT, Mexico) grant number 282785. FIS-G is thankful to CONACyT for the granted scholarship number 629458. BAP-J is grateful for the support given by the subprogram 127 “Basic Training in Research” of the School of Chemistry, UNAM.

References

- [1] Dobson CM. [Chemical space and biology](#). *Nature*. 2004;432:824–8.
- [2] Lipinski C, Hopkins A. [Navigating chemical space for biology and medicine](#). *Nature*. 2004;432:855–61.
- [3] Awale M, Visini R, Probst D, Arús-Pous J, Reymond J-L. [Chemical space: big data challenge for molecular diversity](#). *Chimia*. 2017;71:661–6.
- [4] Naveja J, Rico-Hidalgo MP, Medina-Franco JL. Analysis of a large food chemical database: chemical space, diversity, and complexity. *F1000Res*. 2018;7.
- [5] López-Vallejo F, Giulianotti MA, Houghten RA, Medina-Franco JL. [Expanding the medicinally relevant chemical space with compound libraries](#). *Drug Discov Today*. 2012;17:718–26.
- [6] López-Vallejo F, Waddell J, Yongye AB, Houghten RA, Medina-Franco JL. A large scale classification of molecular fingerprints for the chemical space representation and SAR analysis. *J Cheminform*. 2012;4:P26.
- [7] Medina-Franco JL, Martínez-Mayorga K, Giulianotti MA, Houghten RA, Pinilla C. Visualization of the chemical space in drug discovery. *Current Comput - Aided Drug Des*. 2008;4:322–33.
- [8] Osolodkin DI, Radchenko EV, Orlov AA, Voronkov AE, Palyulin VA, Zefirov NS. Progress in visual representations of chemical space. *Expert Opin Drug Discov*. 2015;10:959–73.
- [9] Opassi G, Gesù A, Massarotti A. The hitchhiker's guide to the chemical-biological galaxy. *Drug Discov Today*. 2018;23:565–74.
- [10] Newman DJ, Cragg GM. Natural products as sources of new drugs from 1981 to 2014. *J Nat Prod*. 2016;79:629–61.
- [11] Bauer A, Brönstrup M. [Industrial natural product chemistry for drug discovery and development](#). *Nat Prod Rep*. 2014;31:35–60.
- [12] Harvey AL, Edrada-Ebel R, Quinn RJ. [The re-emergence of natural products for drug discovery in the genomics era](#). *Nat Rev Drug Discov*. 2015;14:111–29.
- [13] Alvarez-Ruiz E, Collis AJ, Dann AS, Forsbury AP, Reddy SJ, Vázquez Muniz MJ. Microbiological process. Patent. 2017. <https://patentimages.storage.googleapis.com/96/8b/de/87242640defaa1/CN106687596A.pdf>. Accessed: 30 Sep 2018.
- [14] Pereira DM, Valente P, Andrade PB. Tuning protein folding in lysosomal storage diseases: the chemistry behind pharmacological chaperones. *Chem Sci*. 2018;9:1740–52.
- [15] Zhanel GG, Lawson CD, Zelenitsky S, Findlay B, Schweizer F, Adam H, et al. Comparison of the next-generation aminoglycoside plazomicin to gentamicin, tobramycin and amikacin. *Expert Rev Anti Infect Ther*. 2012;10:459–73.
- [16] Cobb R, Boeckh A. Moxidectin: a review of chemistry, pharmacokinetics and use in horses. *Parasit Vectors*. 2009;2:S5.
- [17] Ca G, Ci F, Ag P, Chen C, Tipping R, Cm C, et al. Safety, tolerability, and pharmacokinetics of escalating high doses of ivermectin in healthy adult subjects. *J Clin Pharmacol*. 2002;42:1122–33.
- [18] Brandt W, Haupt VJ, Wessjohann LA. [Chemoinformatic analysis of biologically active macrocycles](#). *Curr Top Med Chem*. 2010;10:1361–79.
- [19] Wessjohann LA, Ruijter E, Garcia-Rivera D, Brandt W. What can a chemist learn from nature's macrocycles? – A brief, conceptual view. *Mol Divers*. 2005;9:171–86.
- [20] Cuevas C, Francesch A. Development of Yondelis (trabectedin, ET-743). A semisynthetic process solves the supply problem. *Nat Prod Rep*. 2009;26:322–37.
- [21] Gajdos C, Elias A. Trabectedin: safety and efficacy in the treatment of advanced sarcoma. *Clin Med Insights Oncol*. 2011;5:35–43.
- [22] Scotti L, Ferreira EI, Ms S, Mt S. [Chemometric studies on natural products as potential inhibitors of the NADH oxidase from Trypanosoma cruzi using the VolSurf approach](#). *Molecules*. 2010;15:7363–77.
- [23] Scotti MT, Scotti L. Editorial: chemometrics in drug discovery. *Comb Chem High Throughput Screen* 2015;18:702–03.
- [24] Rodrigues T, Reker D, Schneider P, Schneider G. [Counting on natural products for drug design](#). *Nat Chem*. 2016;8:531–41.
- [25] Chen Y, de Bruyn Kops C, Kirchmair J. Data resources for the computer-guided discovery of bioactive natural products. *J Chem Inf Model*. 2017;57:2099–111.
- [26] Maier ME. [Design and synthesis of analogues of natural products](#). *Org Biomol Chem*. 2015;13:5302–43.
- [27] Wilk W, Zimmermann TJ, Kaiser M, Waldmann H. Principles, implementation, and application of biology-oriented synthesis (BIOS). *Biol Chem*. 2010;391:491–97.
- [28] Cy-C. TCM Database@Taiwan: the world's largest traditional Chinese medicine database for drug screening in silico. *PLoS One*. 2011;6:e15939.
- [29] Tsai T-Y, Chang K-W, Chen CY. iScreen: world's first cloud-computing web server for virtual screening and de novo drug design based on TCM database@Taiwan. *J Comput Aided Mol Des*. 2011;25:525–31.
- [30] Gu J, Gui Y, Chen L, Yuan G, Lu H-Z XX. Use of natural products as chemical library for drug discovery and network pharmacology. *PLoS One*. 2013;8:e62839.
- [31] Ntie-Kang F, Zofou D, Babiaka SB, Meudom R, Scharfe M, Lifongo LL, et al. AfroDb: a select highly potent and diverse natural product library from African medicinal plants. *PLoS One*. 2013;8:e78085.
- [32] Ntie-Kang F, Onguéné PA, Scharfe M, Owono Owono LC, Megnassan E, Mbaze LM, et al. [ConMedNP: a natural product library from Central African medicinal plants for drug discovery](#). *RSC Adv*. 2014;4:409–19.
- [33] Valli M, Dos Santos RN, Ld F, Ch N, Castro-Gamboa I, Ad A, et al. Development of a natural products database from the biodiversity of Brazil. *J Nat Prod*. 2013;76:439–44.
- [34] Pilon AC, Valli M, Dametto AC, Pinto MEF, Freire RT, Castro-Gamboa I, et al. [NuBBE DB: an updated database to uncover chemical and biological information from Brazilian biodiversity](#). *Sci Rep*. 2017;7:7215.
- [35] NuBBE - Núcleo de Bioensaios, Biossíntese e Ecofisiologia de Produtos Naturais (Nuclei of Bioassays, Ecophysiology and Biosynthesis of Natural Products Database). <http://nubbe.iq.unesp.br/portal/nubbedb.html>. Accessed 30 Sep 2018.
- [36] Naveja J, Oviedo-Osornio CI, Trujillo-Minero NN, Medina-Franco JL. [Chemoinformatics: a perspective from an academic setting in Latin America](#). *Mol Divers*. 2018;22:247–58.
- [37] Medina-Franco JL. Chemoinformatic Characterization of the Chemical Space and Molecular Diversity of Compound Libraries. In: Trabocchi A, editor. *Diversity-Oriented Synthesis*. Hoboken, NJ, USA: John Wiley & Sons, Inc., 2013:325–52.

- [38] Gaspar HA, Sidorov P, Horvath D, Marcou G. Generative topographic mapping approach to chemical space analysis. ACS Symp Ser. 2016. <https://elibrary.ru/item.asp?id=27576908>.
- [39] Tino P, Nabney I. Hierarchical GTM: constructing localized nonlinear projection manifolds in a principled way. IEEE Trans Pattern Anal Mach Intell. 2002;24:639–56.
- [40] Naveja JJ, Medina-Franco JL. ChemMaps: towards an approach for visualizing the chemical space based on adaptive satellite compounds. *Fluorid Res*. 2017;6:1134.
- [41] Feher M, Schmidt JM. [Property distributions: differences between drugs, natural products, and molecules from combinatorial chemistry](#). J Chem Inf Comput Sci. 2003;43:218–27.
- [42] Shelat AA, Guy RK. [The interdependence between screening methods and screening libraries](#). Curr Opin Chem Biol. 2007;11:244–51.
- [43] Singh SB, Chris Culberson J. Chapter 2: chemical space and the difference between natural products and synthetics. In: Antony D Buss, Mark S Butler (editors). Natural product chemistry for drug discovery, Cambridge, UK: Royal Society of Chemistry. 2009:28–43.
- [44] Chen H, Engkvist O, Blomberg N, Li J. A comparative analysis of the molecular topologies for drugs, clinical candidates, natural products, human metabolites and general bioactive compounds. Med Chem Commun. 2012;3:312–21.
- [45] Ertl P, Schuffenhauer A. Cheminformatics analysis of natural products: lessons from nature inspiring the design of new drugs. Prog Drug Res. 2008;66:217, 219–35.
- [46] Pascolutti M, Campitelli M, Nguyen B, Pham N, Gorse A-D, Quinn RJ. Capturing nature's diversity. PLoS One. 2015;10:e0120942.
- [47] González-Medina M, Prieto-Martínez FD, Naveja JJ, Méndez-Lucio O, El-Elimat T, Pearce CJ, et al. [Chemoinformatic expedition of the chemical space of fungal products](#). Future Med Chem. 2016;8:1399–412.
- [48] Chen Y, García de Lomana M, N-O F, Kirchmair J. Characterization of the chemical space of known and readily obtainable natural products. J Chem Inf Model. 2018;58:1518–32.
- [49] Shang J, Hu B, Wang J, Zhu F, Kang Y, Li D, et al. [Cheminformatic Insight into the differences between terrestrial and marine originated natural products](#). J Chem Inf Model. 2018;58:1182–93.
- [50] Ertl P, Schuffenhauer A. Cheminformatics analysis of natural products: lessons from nature inspiring the design of new drugs. Prog Drug Res. 2008;66:217, 219–35.
- [51] Muigg P, Rosén J, Bohlin L, Backlund A. In silico comparison of marine, terrestrial and synthetic compounds using ChemGPS-NP for navigating chemical space. Phytochem Rev. 2013;12:449–57.
- [52] Saldívar-González FI, Valli M, Da Silva Bolzani V, Medina-Franco JL. Chemical diversity of NuBBE database: A chemoinformatic characterization 2018.
- [53] Larsson J, Gottfries J, Muresan S, Backlund A. [ChemGPS-NP: tuned for navigation in biologically relevant chemical space](#). J Nat Prod. 2007;70:789–94.
- [54] Rosén J, Rickardson L, Backlund A, Gullbo J, Bohlin L, Larsson R, et al. [ChemGPS-NP mapping of chemical compounds for prediction of anticancer mode of action](#). QSAR Comb Sci. 2009;28:436–46.
- [55] Korinek M, Tsai Y-H, El-Shazly M, Lai K-H, Backlund A, Wu S-F, et al. [Anti-allergic Hydroxy Fatty Acids from Typhonium blumei Explored through ChemGPS-NP](#). Front Pharmacol. 2017;8:356.
- [56] Rosén J, Lövgren A, Kogej T, Muresan S, Gottfries J, Backlund A. ChemGPS-NP(Web): chemical space navigation online. J Comput Aided Mol Des. 2009;23:253–9.
- [57] Frédéric R, Bruyère C, Vancaeynest C, Reniers J, Meinguet C, Pochet L, et al. [Novel trisubstituted harmine derivatives with original in vitro anticancer activity](#). J Med Chem. 2012;55:6489–501.
- [58] Ertl P, Rohde B. The molecule cloud - compact visualization of large collections of molecules. J Cheminform. 2012;4:12.
- [59] Schuffenhauer A, Ertl P, Roggo S, Wetzel S, Koch MA, Waldmann H. The scaffold tree—visualization of the scaffold universe by hierarchical scaffold classification. J Chem Inf Model. 2007;47:47–58.
- [60] Medina-Franco JL, Petit J, Maggiora GM. [Hierarchical strategy for identifying active chemotype classes in compound databases](#). Chem Biol Drug Des. 2006;67:395–408.
- [61] Koch MA, Schuffenhauer A, Scheck M, Wetzel S, Casaulta M, Odermatt A, et al. Charting biologically relevant chemical space: A structural classification of natural products (SCONP). Proc Natl Acad Sci USA. 2005;102:17272–77.
- [62] Schäfer T, Kriege N, Humbeck L, Klein K, Koch O, Mutzel P. Scaffold Hunter: a comprehensive visual analytics framework for drug discovery. J Cheminform. 2017;9:28.
- [63] Tao L, Zhu F, Qin C, Zhang C, Chen S, Zhang P, et al. [Clustered distribution of natural product leads of drugs in the chemical space as influenced by the privileged target-sites](#). Sci Rep. 2015;5:9325.
- [64] Pye CR, Bertin MJ, Lokey RS, Gerwick WH, Lington RG. [Retrospective analysis of natural products provides insights for future discovery trends](#). Proc Natl Acad Sci USA. 2017;114:5601–6.
- [65] Camp D, Garavelas A, Campitelli M. [Analysis of physicochemical properties for drugs of natural origin](#). J Nat Prod. 2015;78:1370–82.
- [66] Stratton CF, Newman DJ, Tan DS. [Cheminformatic comparison of approved drugs from natural product versus synthetic origins](#). Bioorg Med Chem Lett. 2015;25:4802–7.
- [67] Clemons PA, Bodycombe NE, Carrinski HA, Wilson JA, Shamji AF, Wagner BK, et al. Small molecules of different origins have distinct distributions of structural complexity that correlate with protein-binding profiles. Proc Natl Acad Sci USA. 2010;107:18787–92.
- [68] Medina-Franco JL, Navarrete-Vázquez G, Méndez-Lucio O. [Activity and property landscape modeling is at the interface of chemoinformatics and medicinal chemistry](#). Future Med Chem. 2015;7:1197–211.
- [69] Reddy AS, Zhang S. [Polypharmacology: drug discovery for the future](#). Expert Rev Clin Pharmacol. 2013;6:41–7.
- [70] Medina-Franco JL, Martínez-Mayorga K, Meurice N. Balancing novelty with confined chemical space in modern drug discovery. Expert Opin Drug Discov. 2014;9:151–65.
- [71] van Hattum H, Waldmann H. [Biology-oriented synthesis: harnessing the power of evolution](#). J Am Chem Soc. 2014;136:11853–9.
- [72] Kim S, Thiessen PA, Bolton EE, Chen J, Fu G, Gindulyte A, et al. PubChem substance and compound databases. Nucleic Acids Res. 2016;44:D1202–13.

- [73] Gaulton A, Bellis LJ, Bento AP, Chambers J, Davies M, Hersey A, et al. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* 2012;40:D1100–7.
- [74] Gaulton A, Hersey A, Nowotka M, Bento AP, Chambers J, Mendez D, et al. [The ChEMBL database in 2017](#). *Nucleic Acids Res.* 2017;45:D945–54.
- [75] Wishart DS, Feunang YD, Guo AC, Lo E, Marcu A, Grant JR, et al. DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res.* 2018;46:D1074–82.
- [76] Boufridi A, Quinn RJ. [Harnessing the properties of natural products](#). *Annu Rev Pharmacol Toxicol.* 2018;58:451–70.
- [77] Rosén J, Gottfries J, Muresan S, Backlund A, Oprea TI. [Novel chemical space exploration via natural products](#). *J Med Chem.* 2009;52:1953–62.
- [78] Martínez-Mayorga K, Medina-Franco JL, editors. *Foodinformatics: applications of chemical information to food chemistry*, Switzerland: Springer. 2014. <https://www.springer.com/gp/book/9783319102252>.
- [79] Medina-Franco JL, Martínez-Mayorga K, Peppard TL, Del Rio A. Chemoinformatic analysis of GRAS (Generally recognized as safe) flavor chemicals and natural products. *PLoS One.* 2012;7:e50798.
- [80] Medina-Franco JL. Advances in computational approaches for drug discovery based on natural products. *Revista Latinoamericana de Química.* 2013;41:95–110.
- [81] Houghten RA, Pinilla C, Giulianotti MA, Appel JR, Dooley CT, Nefzi A, et al. [Strategies for the use of mixture-based synthetic combinatorial libraries: scaffold ranking, direct testing in vivo, and enhanced deconvolution by computational methods](#). *J Comb Chem.* 2008;10:3–19.
- [82] Brown N, Jacoby E. [On scaffolds and hopping in medicinal chemistry](#). *Mini Rev Med Chem.* 2006;6:1217–29.
- [83] Singh N, Guha R, Giulianotti MA, Pinilla C, Houghten RA, Medina-Franco JL. Chemoinformatic analysis of combinatorial libraries, drugs, natural products, and molecular libraries small molecule repository. *J Chem Inf Model.* 2009;49:1010–24.
- [84] Yongye AB, Waddell J, Medina-Franco JL. Molecular scaffold analysis of natural products databases in the public domain. *Chem Biol Drug Des.* 2012;80:717–24.
- [85] Lipinski CA. Lead- and drug-like compounds: the rule-of-five revolution. *Drug Discov Today Technol.* 2004;1:337–41.
- [86] Veber DF, Johnson SR, Cheng H-Y, Smith BR, Ward KW, Kopple KD. [Molecular properties that influence the oral bioavailability of drug candidates](#). *J Med Chem.* 2002;45:2615–23.
- [87] Maldonado AG, Doucet JP, Petitjean M, Fan B-T. [Molecular similarity and diversity in chemoinformatics: from theory to applications](#). *Mol Divers.* 2006;10:39–79.
- [88] Schuffenhauer A, Varin T. Rule-based classification of chemical structures by scaffold. *Mol Inform.* 2011;30:646–64.
- [89] Schneider G, Neidhart W, Giller T, Schmid G. “Scaffold-hopping” by topological pharmacophore search: a contribution to virtual screening. *Angew Chem Int Ed.* 1999;38:2894–96.
- [90] Evans BE, Rittle KE, Bock MC, DiPardo RM, Freidinger RM, Whitter WL, et al. [Methods for drug discovery: development of potent, selective, orally effective cholecystokinin antagonists](#). *J Med Chem.* 1988;31:2235–46.
- [91] Medina-Franco JL, Martínez-Mayorga K, Bender A, Scior T. [Scaffold diversity analysis of compound data sets using an entropy-based measure](#). *QSAR Comb Sci.* 2009;28:1551–60.
- [92] González-Medina M, Prieto-Martínez FD, Owen JR, Medina-Franco JL. [Consensus diversity plots: a global diversity analysis of chemical libraries](#). *J Cheminform.* 2016;8:63.
- [93] González-Medina M, Owen JR, El-Elimat T, Pearce CJ, Oberlies NH, Figueroa M, et al. Scaffold diversity of fungal metabolites. *Front Pharmacol.* 2017;8:180.
- [94] Olmedo DA, González-Medina M, Gupta MP, Medina-Franco JL. Cheminformatic characterization of natural products from Panama. *Mol Divers.* 2017;21:779–89.