



UNIVERSIDAD NACIONAL AUTÓNOMA DE MEXICO

PROGRAMA DE MAESTRÍA Y DOCTORADO EN CIENCIAS QUÍMICAS

**PRODUCTOS NATURALES COMO RECURSO PARA EL CRIBADO
VIRTUAL E IDENTIFICACIÓN DE COMPUESTOS BIOACTIVOS**

**PROYECTO DE INVESTIGACIÓN
PARA OPTAR POR EL GRADO DE**

MAESTRA EN CIENCIAS

PRESENTA

Q.F.B. FERNANDA ISABEL SALDÍVAR GONZÁLEZ

**DR. JOSÉ LUIS MEDINA FRANCO
DEPARTAMENTO DE FARMACIA, FACULTAD DE QUÍMICA, UNAM**

Ciudad de México, Julio 2019



Universidad Nacional
Autónoma de México

Dirección General de Bibliotecas de la UNAM

Biblioteca Central



UNAM – Dirección General de Bibliotecas
Tesis Digitales
Restricciones de uso

DERECHOS RESERVADOS ©
PROHIBIDA SU REPRODUCCIÓN TOTAL O PARCIAL

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

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



UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

PROGRAMA DE MAESTRÍA Y DOCTORADO EN CIENCIAS QUÍMICAS

**PRODUCTOS NATURALES COMO RECURSO PARA EL CRIBADO
VIRTUAL E IDENTIFICACIÓN DE COMPUESTOS BIOACTIVOS**

PROYECTO DE INVESTIGACIÓN

PARA OPTAR POR EL GRADO DE

MAESTRA EN CIENCIAS

PRESENTA

Q.F.B. FERNANDA ISABEL SALDÍVAR GONZÁLEZ

DR. JOSÉ LUIS MEDINA FRANCO
DEPARTAMENTO DE FARMACIA, FACULTAD DE QUÍMICA, UNAM



Ciudad de México, Julio 2019

El presente trabajo se realizó principalmente en el grupo DIFACQUIM ubicado en el cubículo 117 del Edificio F de la Facultad de Química de la UNAM, durante el periodo de agosto de 2017 y junio de 2019 bajo de supervisión del Dr. José Luis Medina Franco. El diseño y enumeración de bibliotecas químicas se llevó a cabo en la Universidad de Florencia, Italia, bajo la tutoría de los doctores Andrea Trabocchi y Elena Lenci durante el periodo de diciembre de 2018 a marzo 2019.

Vo.Bo. Dr. José Luis Medina Franco

QFB. Fernanda Isabel Saldívar González

El jurado asignado a este trabajo está integrado por:

Dr. Eduardo Guillermo Delgado Lamas	Instituto de Química, UNAM
Dra. María Isabel Aguilar Laurents	Facultad de Química, UNAM
Dr. Francisco Hernández Luis	Facultad de Química, UNAM
Dr. Jaime Pérez Villanueva	UAM-Xochimilco
Dr. Mario Alberto Figueroa Saldívar	Facultad de Química, UNAM

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

Publicados:

- Saldívar-González FI, Valli M, Andricopulo AD, da Bolzani Vd and Medina-Franco JL. 2019. Chemical Space and Diversity of NuBBE Database: A Chemoinformatic Characterization. *J. Chem. Inf. Model*, 59, 74–85.
- Pílon-Jiménez BA, Saldívar-González FI, Diaz-Eufracio BI and Medina-Franco JL. 2019. BIO-FACQUIM: A Mexican compound database of natural products. *Biomolecules*, 9, 31.
- Lenci E, Menchi G, Saldívar-González FI, Medina-Franco JL and Trabocchi A. 2019. Bicyclic Acetals: Biological Relevance, Scaffold Analysis, and Applications in Diversity-Oriented Synthesis. *Org. Biomol. Chem.*, 17, 1037-1052.
- Saldívar-González FI, Gómez-García A, Chávez-Ponce de León D.E, Sánchez-Cruz N, Ruiz-Rios J, Pílon-Jiménez B.A, Medina-Franco J.L. 2018. Inhibitors of DNA methyltransferases from natural sources: A computational perspective. *Front. Pharmacol.*, 9, 1144.

En preparación:

- Saldívar-González FI, Lenci E, Trabocchi A, Medina-Franco JL. Exploring the chemical space and the bioactivity profile of lactams: a chemoinformatic study. *ACS Omega*. *Enviado*.
- Saldívar-González FI, Lenci E, Trabocchi A, Medina-Franco JL. Computational-aided design of a diversity chemical libraries

Capítulos en libros:

- Naveja JJ, Saldívar-González FI, Sánchez-Cruz N., Medina-Franco JL. 2018. Cheminformatics Approaches to Study Drug Polypharmacology. In: Roy K. (eds) *Multi-Target Drug Design Using Chem-Bioinformatic Approaches*. Methods in Pharmacology and Toxicology. Humana Press, New York, NY.
- Saldívar-González FI, Pílon-Jiménez BA and Medina-Franco JL. 2018. Chemical Space of Naturally Occurring Compounds. In: Fidele Ntie-Kang (ed.) *Cheminformatics of Natural Products*. Physical Sciences Reviews. DOI: 10.1515/psr-2018-0103.
- Saldívar-González FI, and Medina-Franco JL. 2019. Chemoinformatics to assess chemical diversity and complexity of small molecules. In: Andrea Trabocchi and Elena Lenci (eds.) *Small Molecule Drug Discovery: Methods, molecules and applications*. Elsevier. *Enviado*.

Congresos:

- "Productos Naturales como recurso para el cribado virtual e identificación de compuestos bioactivos: Cuantificación de alertas PAINS" Saldívar-González FI, Medina-Franco JL. 52 Congreso Mexicano de Química y 36 Congreso Nacional de Educación Química. 26-29 de septiembre, 2017, Puerto Vallarta, Jalisco, México (cartel).
- "Natural product databases: Chemical space, diversity and suitability of virtual screening" Saldívar-González FI, Medina-Franco JL. 256th American Chemical Society 2018 National Meeting and Exposition. CINP Division. 19- 23 de agosto, 2018, Boston, MA, EUA (oral).

Agradecimientos

A la Universidad Nacional Autónoma de México por haber recibido de ella mi formación profesional y por haberme permitido conocer a personas que quiero y admiro tanto.

Mi más sincero agradecimiento al Dr. José Luis Medina Franco por haberme inspirado a ser una mejor persona y una mejor profesional. Gracias por las enseñanzas, consejos y por la dedicación y amor que imprime en su trabajo, esto posibilitó darle un rumbo a mi investigación e intereses profesionales.

De igual forma, agradezco discusiones y comentarios de mis compañeros de DIFACQUIM. En especial agradezco a Norberto Sánchez, Oscar Palomino, Bárbara Díaz, Eduardo Cortes y Edgar López por su amistad y por todos los momentos que compartimos dentro y fuera del grupo.

Al Dr. Andrea Trabocchi y la Dra Elena Lenci por la calidez y orientación recibida durante mi estancia en la Universidad de Florencia en Italia.

A la Dra. Marilia Vali de la Universidad Estatal de São Paulo en Brasil, por la colaboración y por proporcionarnos la base de datos de NuBBE.

A los sinodales, por tomarse el tiempo de leer y realizar las sugerencias pertinentes para enriquecer y mejorar este informe.

A mis padres Daniel Saldivar y Janet González y a mi hermano Diego por el apoyo incondicional que me han brindado durante mis estudios. Gracias por todo su amor y paciencia.

Finalmente, agradezco al Consejo Nacional de Ciencia y Tecnología (CONACyT) por la beca otorgada para mis estudios de maestría (No. 629458) y por la beca mixta otorgada para realizar la estancia de investigación en la Universidad de Florencia. Se agradece también el financiamiento al proyecto de Ciencia Básica CONACyT 282785.

Resumen

Los productos naturales (PNs) continúan brindando una fuente diversa y única de compuestos bioactivos para el descubrimiento de fármacos. Sin embargo, la investigación en este campo es más compleja, costosa e ineficiente en comparación con la investigación de moléculas pequeñas obtenidas por síntesis. Por lo tanto, el uso de métodos computacionales para encontrar nuevas estructuras bioactivas a partir de PNs representa una alternativa para superar estos problemas y realizar búsquedas más dirigidas y menos costosas.

En este trabajo, mediante diversas herramientas quimioinformáticas, se analizó la diversidad estructural, la complejidad molecular y la distribución en el espacio químico de diferentes bases de datos de PNs como recursos para hacer cribado virtual. Núcleos estructurales base (*scaffolds*) de importancia biológica en PNs como los acetales bicíclicos y las lactamas fueron identificados y clasificados de manera sistemática mediante flujos de trabajo desarrollados en KNIME. La información generada fue de utilidad para la construcción de un flujo de trabajo que permite el diseño de nuevas bibliotecas químicas bajo un enfoque de síntesis orientada en diversidad (DOS, por sus siglas en inglés). Las bibliotecas químicas diseñadas e inspiradas en PNs, representan fuentes de inicio para hacer cribado virtual.

Índice general

1. Antecedentes	1
1.1. Productos naturales en el área de diseño de fármacos	1
1.2. Estudios quimioinformáticos de PNs	1
1.3. Estructuras privilegiadas en bases de datos de PNs: acetales bicíclicos y lactamas . .	2
1.4. Construcción de bibliotecas moleculares para cribado virtual	2
1.5. Diseño y construcción <i>in silico</i> de bibliotecas enfocadas en diversidad (DOS)	3
2. Objetivos	4
2.1. General	4
2.2. Particulares	4
3. Metodología	5
3.1. Análisis quimioinformático de bases de PNs	5
3.1.1. Propiedades fisicoquímicas	5
3.1.2. Análisis de <i>scaffolds</i>	6
3.1.3. Entropía de Shannon	6
3.1.4. Huellas digitales moleculares (<i>molecular fingerprints</i>)	6
3.1.5. <i>Consensus Diversity Plot</i> (CDPlot)	7
3.2. Identificación, clasificación y cuantificación de estructuras de relevancia terapéutica	7
3.3. Perfil biológico de estructuras de relevancia terapéutica: Lactamas	7
3.3.1. Factor de enriquecimiento (EF) de <i>scaffolds</i>	7
3.4. Diseño y enumeración de bibliotecas químicas enfocadas en diversidad	9
4. Resultados y discusión	11
4.1. Análisis quimioinformático de bases de PNs	11
4.2. Identificación, clasificación y cuantificación de estructuras de relevancia farmacéutica	12
4.3. Perfil biológico de las lactamas	14
4.4. Diseño y enumeración de bibliotecas químicas enfocadas en diversidad	16
5. Conclusiones y Perspectivas	19
5.1. Conclusiones	19
5.2. Perspectivas	19
Bibliografía	20

Capítulo 1

Antecedentes

En los últimos años se ha incrementado la cantidad reportada de metabolitos provenientes de PNs y el número de dianas moleculares relevantes en la terapia de trastornos humanos.^{1,2} En este escenario, herramientas quimioinformáticas son de gran utilidad para tratar con bibliotecas con numerosas estructuras de potenciales moléculas bioactivas.³ No obstante, son pocos los análisis que se han hecho a PNs. Dada la importancia de los compuestos de origen natural como fuente para encontrar candidatos a fármacos, la construcción de bases de datos de PNs y el subsecuente análisis de su diversidad, complejidad molecular y distribución en el espacio químico, representan el primer paso para que estas colecciones sean utilizadas en estudios de cribado virtual. Estructuras privilegiadas y sistemas de anillos innovadores que ofrecen PNs han hecho que estos compuestos sean usados ampliamente como punto de partida para inspirar y guiar el diseño de bibliotecas químicas y el diseño basado en fragmentos. En este sentido, la construcción de flujos de trabajo automatizados que integren la identificación, la clasificación y el análisis de estructuras de importancia biológica permite enfocar el diseño de nuevas estructuras químicas a espacios poco explorados y con mayor oportunidad de éxito.

1.1. Productos naturales en el área de diseño de fármacos

Si bien el proceso de descubrimiento y desarrollo de fármacos se ha revolucionado con el advenimiento de tecnologías más eficientes como la química combinatoria y el *high-throughput screening* (HTS), los PNs han atraído una vez más la atención de académicos e investigadores enfocados en la química farmacéutica. Esto es debido a que los PNs son una fuente más prometedora de fármacos que los compuestos obtenidos por química combinatoria o enfoques tradicionales de síntesis.^{4,5} Actualmente, cerca de 250 mil PNs están disponibles para estudios de cribado virtual, tanto en bibliotecas comerciales como de acceso libre.⁶ Se espera que este número aumente debido al incremento de los grupos de investigación que desarrollan bibliotecas de PNs. Un ejemplo es la biblioteca BIOFACQUIM, que contiene PNs aislados en México y que ha sido construida recientemente dentro del grupo DIFACQUIM.⁷ Una lista detallada de bases de datos de PNs puede ser consultada en la literatura científica.^{6,8}

1.2. Estudios quimioinformáticos de PNs

La quimioinformática es una disciplina que usa métodos computacionales para el manejo, la visualización y el análisis sistemático de información química.^{9,10} Esta disciplina tiene aplicaciones en diferentes áreas como la química analítica, la química orgánica y más recientemente en la

química de alimentos¹¹ y de materiales.¹² Hasta ahora, la quimioinformática ha tenido su mayor impacto en el descubrimiento y desarrollo de fármacos.¹³ En este contexto, herramientas quimioinformáticas se han utilizado para caracterizar bibliotecas de PNs en términos de descriptores moleculares, diversidad y complejidad molecular, comparando estas bibliotecas con otras de referencia como compuestos de síntesis y fármacos aprobados.¹⁴ Otras aplicaciones incluyen la visualización del espacio químico de los PNs,¹⁵ el cual permite a su vez la clasificación de los compuestos por efecto farmacológico, diana molecular, o biosíntesis, y la selección de compuestos para su cribado virtual. Como se mencionó, los PNs contienen sistemas de anillos innovadores con geometrías adecuadas para el posicionamiento espacial de las cadenas laterales. Por tanto, muchos PNs se han empleado para el diseño de bibliotecas químicas y el diseño basado en fragmentos.^{16,17} Esfuerzos recientes en esta área se están enfocando en la elucidación de estructuras,¹⁸ la predicción de actividad, en la identificación de compuestos líderes a partir de PNs y en la elucidación de la biosíntesis de metabolitos.¹⁹

1.3. Estructuras privilegiadas en bases de datos de PNs: acetales bicíclicos y lactamas

Los PNs y los fármacos derivados de éstos destacan por tener una mayor diversidad estructural que los fármacos sintéticos y presentan un área mayor del espacio químico.⁶ Se ha establecido que los PNs provienen de la adaptación de un organismo a su entorno y tienen un propósito biológico específico basado en la evolución. Por lo tanto, tiene sentido que los PNs generalmente tengan estructuras con relevancia biológica. Las subestructuras o los *scaffolds* de los PNs se consideran con frecuencia estructuras privilegiadas^{20–22} lo que significa que pueden ofrecer una actividad mayor o específica contra diversas dianas o blancos biológicos. Además, en un estudio reciente se encontró que casi 1300 *scaffolds* encontrados en PNs (83 % de todos los *scaffolds*), no se encuentran en bibliotecas de compuestos comerciales.²³ Este resultado revela el potencial de las características estructurales de PNs que no se han explotado en términos de bibliotecas combinatorias o de síntesis en general.

Entre las diferentes estructuras químicas en la naturaleza, los acetales bicíclicos y las lactamas son particularmente relevantes en la química farmacéutica, debido a su papel clave en varias interacciones biológicas y la diversidad química que proviene de las muchas combinaciones posibles de anillos. Estructuralmente, un grupo acetal bicíclico consiste en al menos dos anillos donde dos átomos de oxígeno que pertenecen a diferentes anillos están unidos a través de un átomo de carbono en común. Estos compuestos están presentes en una variedad de PNs de diferente origen, incluyendo insectos, organismos marinos, hongos y plantas.^{24,25} Por otra parte, una lactama es una amida cíclica que dependiendo el número de átomos que integren el anillo pueden clasificarse en *beta* (4 miembros), *gamma* (5 miembros), *delta* (seis miembros) y *epsilon* (siete miembros).

1.4. Construcción de bibliotecas moleculares para cribado virtual

Cada año aumenta el número de compuestos disponibles para realizar cribado virtual. Las bibliotecas de compuestos pueden ser consultadas de forma gratuita o comercial, o bien, pueden ser generadas *in silico*. Los avances en la capacidad de procesamiento computacional y de almacenamiento han permitido a investigadores generar bibliotecas químicas virtuales que contienen miles de moléculas. Algunos ejemplos de bibliotecas virtuales son *Generated Data-Base* (GDB-17) con 166 miles de millones de compuestos,²⁶ *Fragment Database* (FDB-17) con 10 millones de moléculas

con propiedades *fragment-like*,²⁷ y *Screenable Chemical Universe Based on Intuitive Data OrganizatiOn* (SCUBIDOO) con 21 millones de compuestos obtenidos de la aplicación de 58 reacciones robustas a un grupo de 18 561 *building blocks*.²⁸ Aunque estos números parecen grandes, únicamente una pequeña fracción de éstas moléculas orgánicas podrían potencialmente ser sintetizadas.²⁹ Respecto a los PNs, se ha observado que la tasa de PNs descubiertos aumenta con el tiempo, no obstante, ocurre lo contrario con la tasa de PNs con estructuras novedosas.³⁰ En este sentido hay que tener en cuenta que más allá del enfoque terapéutico obvio, la novedad es igualmente importante³¹ a la hora de realizar el diseño de bibliotecas químicas.

1.5. Diseño y construcción *in silico* de bibliotecas enfocadas en diversidad (DOS)

La síntesis orientada a la diversidad (DOS, por sus siglas en inglés) es un área que proporciona acceso a moléculas pequeñas, complejas y diversas, que prometen modular la actividad de muchas dianas o blancos biológicos que han estado fuera del alcance de las colecciones de compuestos tradicionales.³² Desde los inicios de DOS, se han reconocido dos estrategias para generar colecciones diversas:³² 1) enfoque basado en los reactivos, donde una molécula dada se somete a un rango de condiciones de reacción que permiten la síntesis de un número de compuestos distintos y 2) enfoque basado en los sustratos, donde un número materiales de partida que contienen información esquelética precodificada se transforman bajo las mismas condiciones en una gama de estructuras moleculares. Algunas de estas ideas fueron refinadas por Schreiber et al. cuando describieron la estrategia de *Build/Couple/Pair* para generar diversidad estructural.³³ La Figura 1.1 describe las principales estrategias aplicadas en DOS. Dentro de DOS también se han hecho esfuerzos para el desarrollo de bibliotecas basadas en andamios privilegiados.³⁴

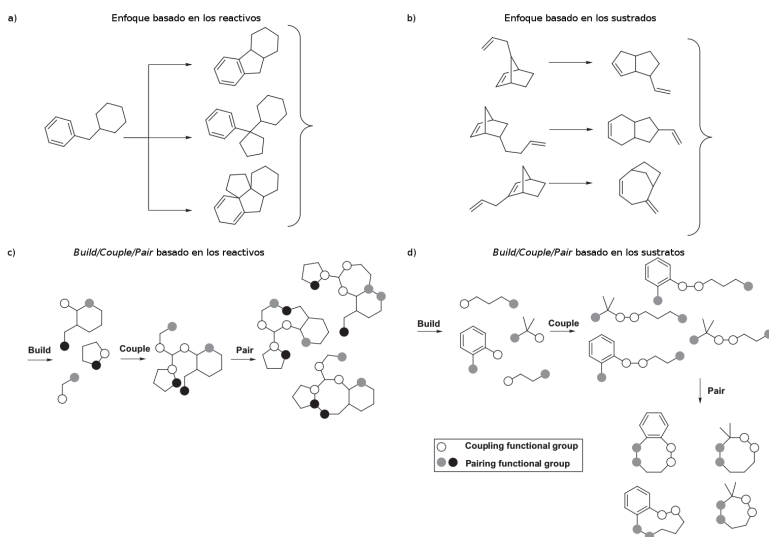


Figura 1.1: Estrategias aplicadas en el diseño orientado en diversidad: enfoque basado en los reactivos, b) enfoque basado en los sustratos, c) *Build/Couple/Pair* basado en los reactivos y d) *Build/Couple/Pair* basado en los sustratos. Imagen tomada de O'Connell y Warren.³²

Capítulo 2

Objetivos

2.1. General

Contribuir al estudio de los PNs mediante el análisis quimioinformático de diferentes bases de datos de compuestos de origen natural disponibles para estudios de cribado virtual, y generar flujos de trabajo para identificar, clasificar y cuantificar la presencia de estructuras de relevancia biológica en PNs y en otras bases de datos de interés farmacéutico.

2.2. Particulares

- Analizar la diversidad molecular, la complejidad estructural y la distribución en el espacio químico de bases de datos de PNs.
- Identificar, clasificar y cuantificar la presencia de acetales bicíclicos y lactamas en PNs y otras bases de datos de relevancia farmacéutica.
- Describir el perfil biológico de estos compuestos de acuerdo a su clasificación química.
- Identificar núcleos estructurales enriquecidos y con pocos análogos, para generar una biblioteca de nuevos compuestos químicos.
- Automatizar el diseño de bibliotecas químicas enfocado en diversidad, mediante la integración de métodos quimioinformáticos de acceso libre.

Capítulo 3

Metodología

3.1. Análisis quimioinformático de bases de PNs

Para analizar la diversidad de los PNs, se examinaron varias bases de datos moleculares bajo métricas reportadas en la literatura, entre ellas: propiedades fisicoquímicas, huellas digitales moleculares (*molecular fingerprints*) y análisis de núcleos base (*scaffolds*). Las bibliotecas de compuestos utilizadas en este trabajo se describen en la Tabla 3.1. Para un análisis simultáneo integrando las diferentes métricas, se hizo uso de los gráficos CDP (*Consensus Diversity Plots*).³⁵ Por su parte, la complejidad molecular fue analizada mediante el cálculo de los descriptores: fracción de carbonos sp³ (Fsp³), peso molecular, y fracción de carbonos quirales (FCC).

Tabla 3.1. Bases utilizadas para el análisis de la diversidad y la complejidad de los PNs

Base de datos	No. de compuestos únicos	Referencia
Fármacos aprobados	1806	www.drugbank.ca
Metabolitos de cianobacterias	473	<i>In-house</i>
Metabolitos de hongos	206	³⁶
PNs marinos	6253	³⁰
PNs disponibles comercialmente (MEGx)	4103	ac-discovery.com
Compuestos semi-sintéticos (NATx)	26318	ac-discovery.com
PNs de Brasil	2214	³⁷
PNs de medicina tradicional China (TCM)	17986	³⁸
<i>Universal Natural Products Database</i> (UNPD)	209574	³⁹

3.1.1. Propiedades fisicoquímicas

Se calcularon siete descriptores moleculares relevantes desde el punto de vista farmacéutico y que están asociados con frecuencia a factores que contribuyen a una buena disponibilidad oral de fármacos.^{40,41} Estos descriptores son: número de átomos aceptores y donadores de puentes de hidrógeno (HBA y HBD, respectivamente), coeficiente de partición octanol/agua (log P), peso molecular, número de enlaces rotantes, área de superficie polar topológica (TPSA) y fracción de carbonos con hibridación sp³ (FCsp³). Se obtuvo la distribución de cada una de las propiedades calculadas, así como el valor de su media, mediana, rango intercuartílico y desviación estándar mediante el programa R.⁴² Para facilitar la representación visual de los descriptores moleculares, se realizó un análisis de componentes principales (PCA, por sus siglas en inglés)⁴³ en el programa KNIME.⁴⁴

3.1.2. Análisis de *scaffolds*

Otra estrategia para cuantificar la diversidad de las bases de datos fue contar los *scaffolds* obtenidos con la definición de Murcko.⁴⁵ De acuerdo con esta definición, a cada compuesto se le remueve de forma sistemática todos aquellos vértices con grado uno, dando como resultado gráficos cíclicos. En caso de existir más de un ciclo, se incluyen los átomos que conectan a dichos sistemas. Con los resultados obtenidos del conteo de *scaffolds*, se graficó la fracción de sistemas cíclicos contra la fracción acumulativa en la base de datos (gráficos *Cyclic System Retrieval* (CSR)) y de esta manera se realizó la comparación directa con el contenido y diversidad en otras bases de datos. Los casos extremos que ayudan a interpretar los gráficos CSR son los siguientes: a) en el caso de contener un sistema cíclico diferente para cada uno de los compuestos en la biblioteca, dará como resultado una diagonal, lo que corresponderá a la máxima diversidad de *scaffolds* y por ende presentará una área bajo la curva de 0.5; y b) en el caso opuesto, de encontrar un solo sistema cíclico que englobe a todos los compuestos se obtendrá un escalón con su máximo en uno, el mínimo de diversidad posible y con un área bajo la curva de 1.

3.1.3. Entropía de Shannon

Para obtener mayor información y realizar un análisis más completo de la diversidad de *scaffolds*, una métrica reportada y muy utilizada es la entropía de Shannon (SE).⁴⁶ Esta métrica, a diferencia de las curvas CSR, considera la distribución específica de las moléculas en un dado n número de *scaffolds* más poblados. La SE de una población de compuestos P distribuidos en n sistemas se define como:

$$SE = - \sum_{i=1}^n p_i \log_2 p_i \quad p_i = \frac{c_i}{P} \quad , \quad (3.1)$$

donde p_i es la probabilidad estimada de la ocurrencia de un *scaffold* específico i en una población de P compuestos que contienen un total de n sistemas acíclicos y cíclicos, y c_i es el número de moléculas que contienen un quimiotipo particular. Para normalizar SE a los diferentes n , la entropía Shannon escalada (SSE) se define como:

$$SEE = \frac{SE}{\log_2 n} \quad (3.2)$$

El valor de SSE oscila entre cero cuando todos los compuestos tienen el mismo *scaffold* (diversidad mínima) y 1.0 cuando todos los compuestos están distribuidos uniformemente entre los n sistemas acíclicos y/o cíclicos (diversidad máxima). Para probar la dependencia de SSE con varios números máximos de quimiotipos, se consideraron diferentes números de n (5-70).

3.1.4. Huellas digitales moleculares (*molecular fingerprints*)

Las bases de datos fueron analizadas con huellas digitales moleculares para evaluar su similitud intra-colección. Para ello, se calcularon en KNIME las huellas digitales ECFP4, un *fingerprint* topológico circular y MACCS keys, un *fingerprint* basado en un diccionario de 166 -bits. Posteriormente, se calculó la matriz de similitud utilizando el índice de Tanimoto.⁴⁷ Los valores fuera de la diagonal de la matriz de similitud se utilizaron para graficar la función de distribución acumulativa para cada una de las bases de datos.

3.1.5. Consensus Diversity Plot (CDPlot)

Los gráficos CDPlots (*Consensus Diversity Plots*)³⁵ son una nueva propuesta de metodología que permite representar en dos dimensiones la diversidad global de las bases de datos. Estos gráficos pueden ser construidos usando diversas métricas individuales o combinadas de diversidad, por ejemplo métricas del análisis de *scaffolds*, huellas moleculares (*molecular fingerprints*) y propiedades fisicoquímicas, permitiendo así un análisis más completo.

3.2. Identificación, clasificación y cuantificación de estructuras de relevancia terapéutica

La identificación y clasificación de PNs que contienen *scaffolds* con acetales bicíclicos y lactamas se llevó a cabo mediante SMARTS, un lenguaje útil para describir patrones moleculares y para realizar búsqueda de subestructuras.⁴⁸ La Figura 3.1 muestra las subestructuras definidas para la identificación y clasificación de los acetales bicíclicos en las bases de PNs. Para este estudio se analizaron bases de datos de PNs de acceso libre: MEGx, UNPD, PNs marinos, metabolitos de hongos, PNs presentes en la base de datos ZINC⁴⁹ y en PNs provenientes de medicina tradicional China (TCM) (Tabla 3.1).

La Figura 3.2 muestra las subestructuras definidas para la identificación y la clasificación de lactamas en las bases de datos UNPD,³⁹ fármacos aprobados⁵⁰ y ChEMBL.⁵¹

3.3. Perfil biológico de estructuras de relevancia terapéutica: Lactamas

Las lactamas provenientes de la base de datos ChEMBL se asociaron con sus datos de actividad biológica. Para este trabajo, sólo se incluyeron los compuestos con actividad reportada como IC₅₀ y porcentaje de inhibición. Aquellos compuestos con un pIC₅₀ ($-\log(IC_{50})$) mayor que 6 (<1000 nM) y con un porcentaje de inhibición mayor que 60 % se consideraron “activos”.

3.3.1. Factor de enriquecimiento (EF) de *scaffolds*

Dado que los valores de actividad para las lactamas fueron disponibles, también fue posible construir los gráficos del factor de enriquecimiento (EF).⁵³ Estos gráficos son útiles porque proporcionan información general y cuantitativa sobre el número de análogos reportados para cada *scaffold*, así como su actividad promedio con respecto a los otros *scaffolds* presentes. El valor de EF da la proporción de compuestos activos para un sistema cíclico o *scaffold* en un determinado blanco biológico, y la frecuencia nos da una idea de la cantidad de análogos que hay para un *scaffold* en particular. Para llevar a cabo la gráfica de EF, a cada compuesto se le calculó su *scaffold* según la definición de Bemis y Murcko⁴⁵ y los compuestos se agruparon por las dianas biológicas contra las que eran activos. Los valores de EF se calcularon usando la expresión:

$$EF(C_l) = \frac{Act(C_l)}{Act(C)} \quad , \quad (3.3)$$

$$\text{donde, } Act(C_l) = \frac{(C_l^+)}{(C_l)} \quad Act(C) = \frac{(C^+)}{(C)}$$

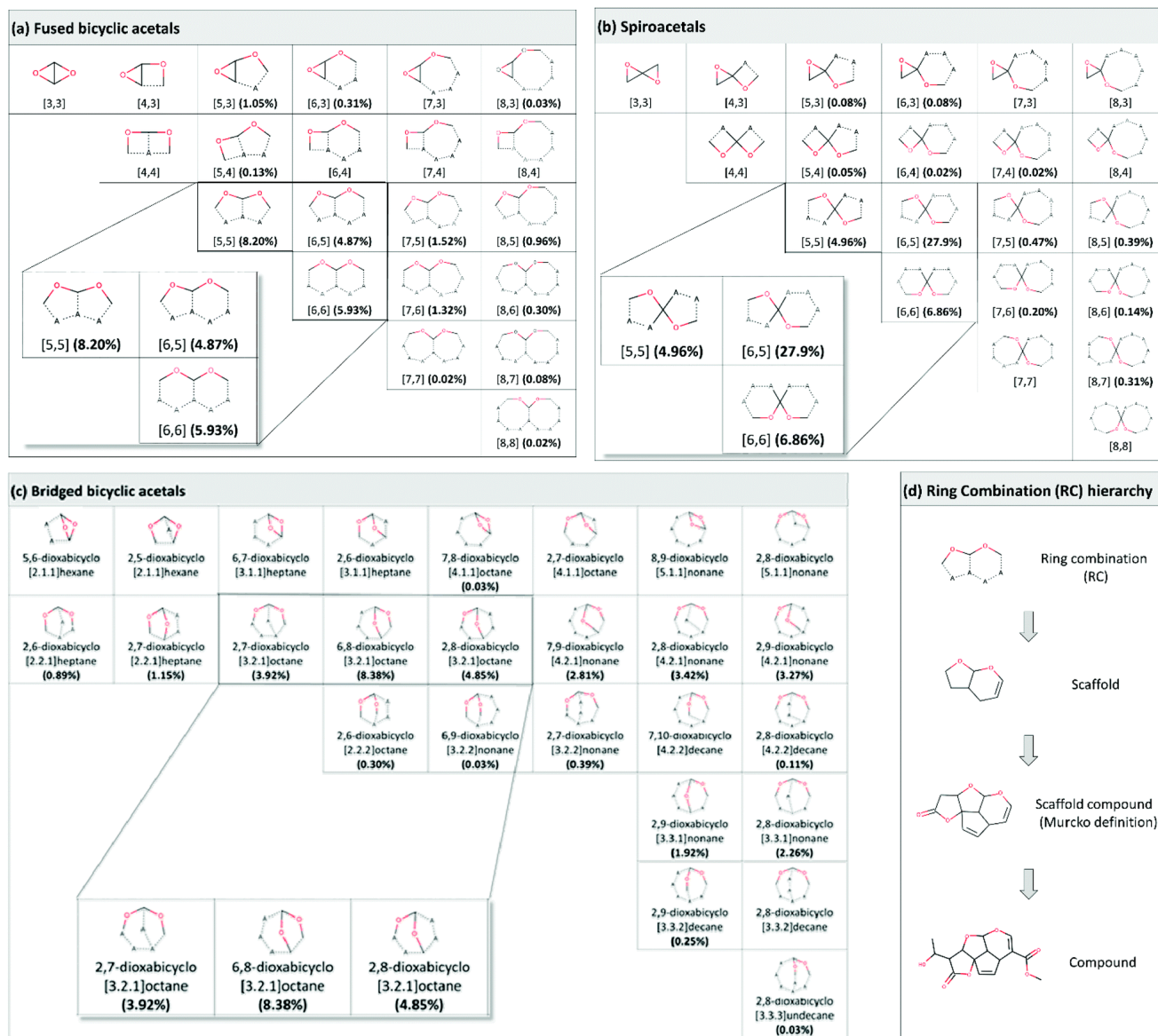


Figura 3.1: Clasificación y porcentaje de acetales bicíclicos a) *fused*, b) *spiro* y c) *bridged* en bases de PNs. Las tres combinaciones de anillos más frecuentes para cada clase de acetales bicíclicos se muestran en recuadros ampliados. El porcentaje mostrado es relativo al número total de marcos de acetal bicíclicos identificados en la base de datos (6369) y no al número total de productos naturales que contienen acetal (4699), ya que algunos PNs contienen más de un tipo de acetal bicíclico en su estructura. (d) Ejemplo de la combinación de anillo [5,6]-fused acetal. Imágenes tomadas de Lenci et al.⁵².

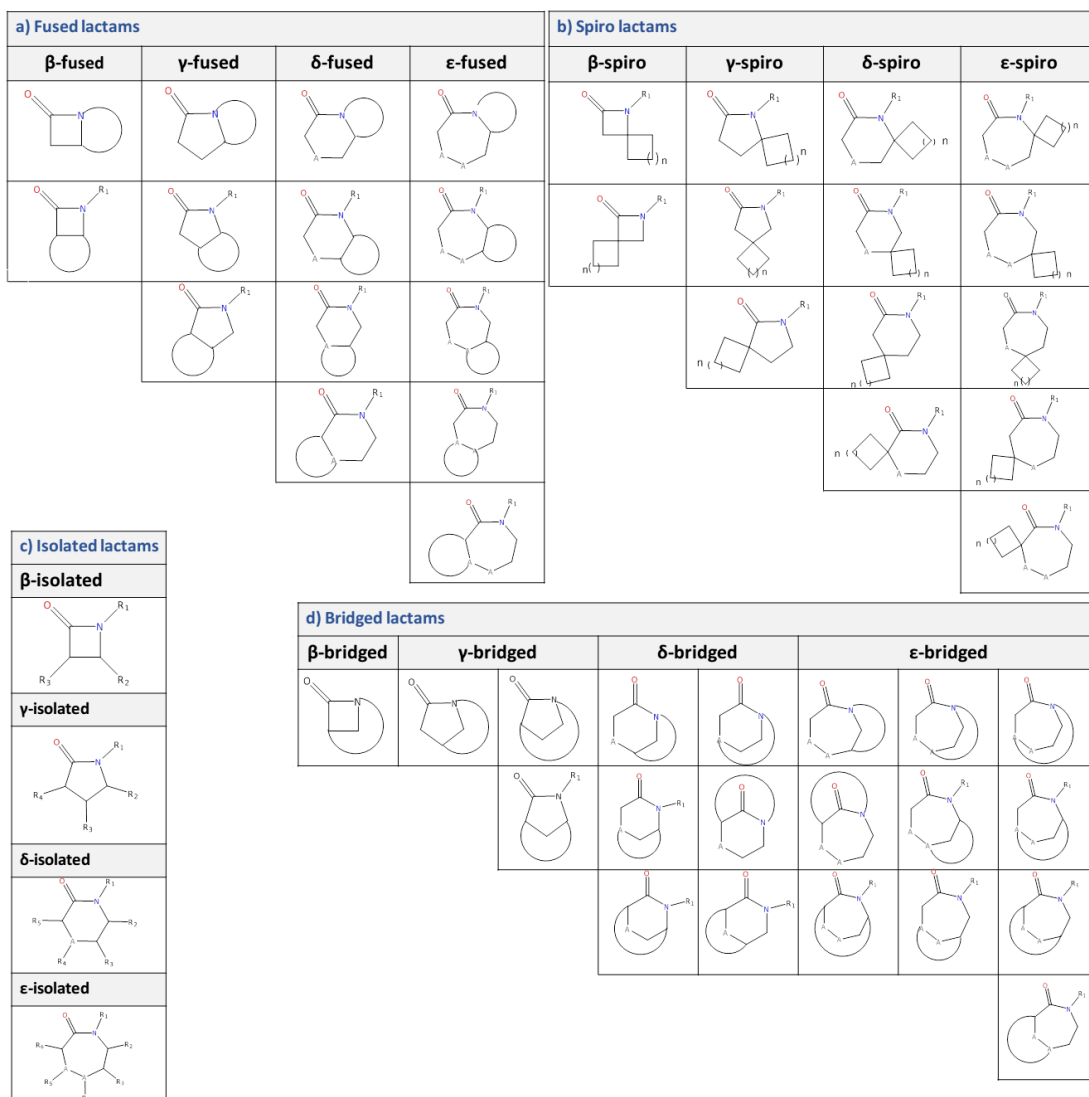
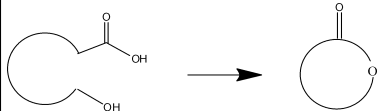
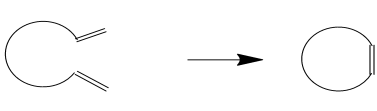
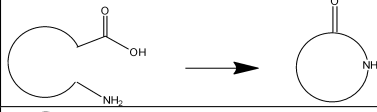
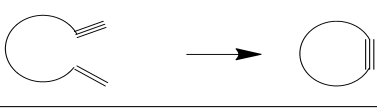
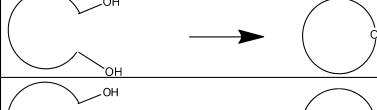
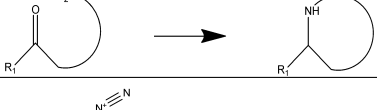
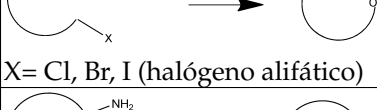
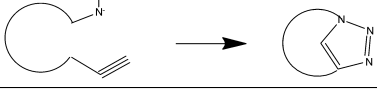
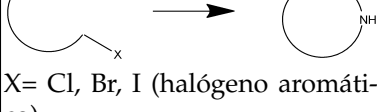


Figura 3.2: Subestructuras definidas para la identificación y la clasificación de las *beta*-, *gamma*-, *delta*- y *epsilon*- lactamas, que a su vez están clasificadas como *isolated*, *bridged*, *fused* y *spiro* lactamas.

3.4. Diseño y enumeración de bibliotecas químicas enfocadas en diversidad

Para el diseño de la biblioteca de lactamas se siguió una estrategia aplicada en la síntesis orientada en diversidad llamada *Build/Couple/Pair* (B/C/P).⁵⁴ Como punto de partida se seleccionó la biblioteca de *building blocks* de Enamine.⁵⁵ Esta biblioteca contiene 437,625 compuestos. En la fase de acoplamiento o *coupling* únicamente se seleccionaron *building blocks* con dos o más grupos funcionales. Para enfocar la biblioteca en lactamas, solo se consideró en la fase de acoplamiento la reacción de formación de carboxamidas entre ácidos carboxílicos y aminas primarias y secundarias. Finalmente, para la fase de emparejamiento o *pairing* se consideraron las reacciones descritas en la Tabla 3.2.

Tabla 3.2. Reacciones intramoleculares usadas en la reacción de emparejamiento

Reacción	Nombre de la reacción	Reacción	Nombre de la reacción
	Lactonización		Metátesis
	Lactamización		Metátesis de eninos
	Condensación de alcoholes		Aminación reductiva
 X= Cl, Br, I (halógeno alifático)	Síntesis de Williamson		Reacción de química click
 X= Cl, Br, I (halógeno aromático)	Acoplamiento cruzado Buchwald-Hartwig		

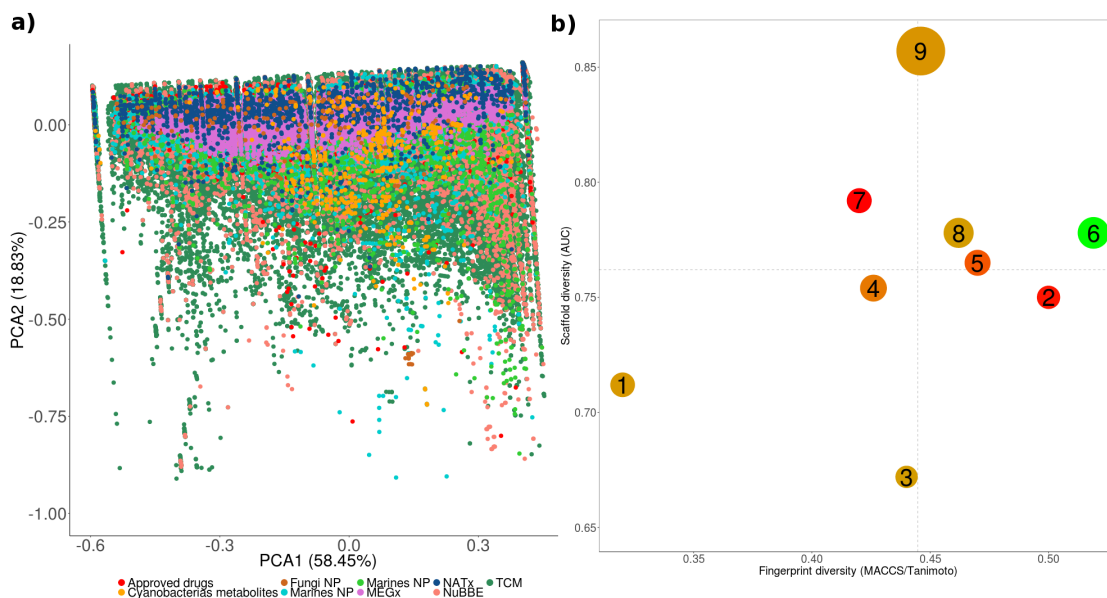
Capítulo 4

Resultados y discusión

4.1. Análisis quimioinformático de bases de PNs

En la Figura 4.1a se muestran los resultados del análisis de propiedades fisicoquímicas y la visualización del espacio químico de PNs. Como se observa en ésta figura, la base de datos de UNPD cubre la mayor parte del espacio químico y es también la base de datos con la mayor diversidad de propiedades fisicoquímicas. En contraste, NuBBE_{DB} y los productos semisintéticos (NATx) ocupan una región más enfocada en el espacio químico, que a su vez se incluye dentro del espacio de fármacos aprobados y UNPD.

Los resultados del análisis de diversidad y de complejidad de PNs indican que, en general, estas bases de datos son más diversas en cuanto a propiedades moleculares. Sin embargo, la diversidad en cuanto a *scaffolds* y *fingreprints* y la complejidad estructural varían según el origen de los compuestos. Como se observa en el CDPlot (Figura 4.1b), los metabolitos de hongos y cianobacterias tienen la mayor diversidad de *scaffolds*, mientras que NuBBE_{DB} y los PNs marinos son los más diversos considerando *fingerprint*s. También se observó que los metabolitos de cianobacterias resaltan por su alta complejidad estructural y su perfil distintivo basado en propiedades moleculares y alertas subestructurales, y son también diferentes de otros PNs (Figura 4.1c y 4.1d).



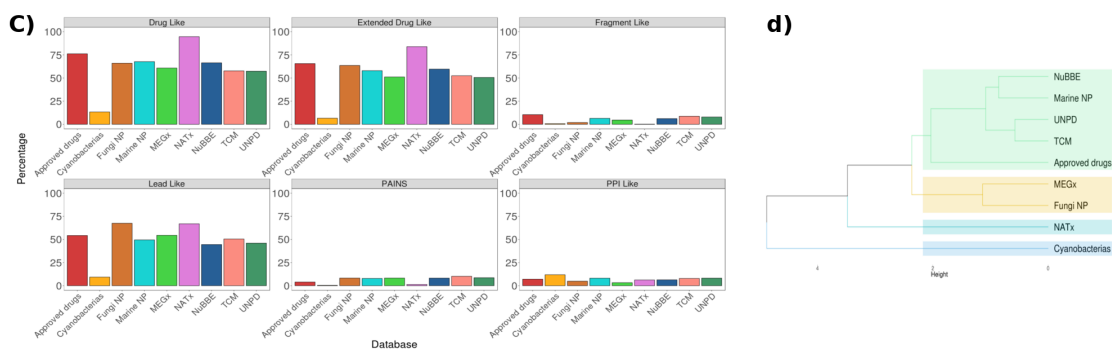


Figura 4.1: Resultados representativos del análisis quimioinformático de PNs. **a)** Representación visual del espacio químico. La representación visual se generó con un análisis de componentes principales de siete propiedades fisicoquímicas tal como se describe en métodos; **b)** CDPlot que compara la diversidad de diferentes bibliotecas de NPs. Cada punto representa una base de datos: fármacos aprobados (1), metabolitos de cianobacterias (2), metabolitos de hongos (3), PNs marinos (4), MEGx (5), NATx (6), NuBBE_{DB} (7), TCM (8) y UNPD (9). La mediana de MACCS / Tanimoto del conjunto de datos se representa en el eje X y el AUC de las curvas de recuperación del *scaffolds* en el eje Y. Los puntos de datos están coloreados por la diversidad de las propiedades fisicoquímicas del conjunto de datos medido por la distancia euclidiana de seis propiedades de relevancia farmacéutica. La distancia se representa con una escala de color continua de rojo (más diverso), a naranja/marrón (diversidad intermedia), a verde (menos diverso). El tamaño relativo del conjunto de datos se representa con el tamaño del punto de datos: los puntos de datos más pequeños indican conjuntos de datos compuestos con menos moléculas; **c)** Porcentaje de compuestos en cada uno de los 6 subconjuntos (*drug-like*, *extended drug-like*, *fragment-like*, *lead-like*, *PPI-like*, and *Pan Assay Interference Compounds - PAINS*); y **d)** Agrupamiento jerárquico (promedio/distancia euclidiana) de acuerdo con el porcentaje de compuestos de variables no correlacionadas (*drug-like*, *fragment-like*, *PPI-like*, and *PAINS*).

4.2. Identificación, clasificación y cuantificación de estructuras de relevancia farmacéutica

En la Figura 4.2 se resumen los resultados obtenidos del análisis de la diversidad topológica de los *scaffolds* de PNs que contienen acetales bicíclicos. Se observa la frecuencia de acetales bicíclicos *bridged*, *fused* y *spiro* en siete bases de datos diferentes de PNs. En total, de 466328 PNs analizados, se identificaron 4699 compuestos con acetales bicíclicos en su estructura y 45 combinaciones de anillos diferentes. Los PNs de TCM, PNs marinos y la base de UNPD son particularmente interesantes, ya que presentan un alto porcentaje de compuestos con acetales bicíclicos, con una prevalencia de [6,5]-espiroacetales en las bases de datos de UNPD y TCM y [6,6]-espiroacetales en la base de datos de PNs marinos. Este análisis también reveló que de las 45 combinaciones de anillos, los sistemas [5,5], [6,6] y [5,6] fueron los más abundantes dentro de la categoría de acetales bicíclicos fusionados, mientras que en la categoría de acetales bicíclicos *bridged*, el 6,8-dioxabicyclo[3.2.1]octano demostró ser el sistema cíclico más abundante. En cuanto a los espiroacetales, los sistemas [6,5] y [6,6] fueron las combinaciones de anillos con mayor número de compuestos. Por otro lado, los resultados de diversidad y complejidad de acetales bicíclicos sugieren que el grupo de acetales bicíclicos fusionados presenta una mayor diversidad cuando se mide bajo métricas de diversidad de *scaffolds* (Figura 4.3a), mientras que los espiroacetales destacan por su

De la misma forma, en la Figura 4.4 se observa la frecuencia de lactamas *isolated*, *bridged*, *fused* y *spiro* en las bases de datos UNPD, fármacos aprobados y ChEMBL. Para las tres bases de datos analizadas hay una fracción más grande de lactamas *fused*, en la base de fármacos aprobados, hay una mayor cantidad de *beta-fused*, mientras que en ChEMBL y UNPD hay una mayor cantidad de lactamas *gamma-fused* y *delta-fused*, respectivamente. No se encontraron lactamas *bridged* y *spiro* en las bases de PNs y fármacos aprobados, y en ChEMBL se han explorado muy poco. En cuanto a los compuestos *isolated*, los más explorados y los que se encuentran más frecuentemente en la naturaleza son los clasificados como *gamma-isolated*.

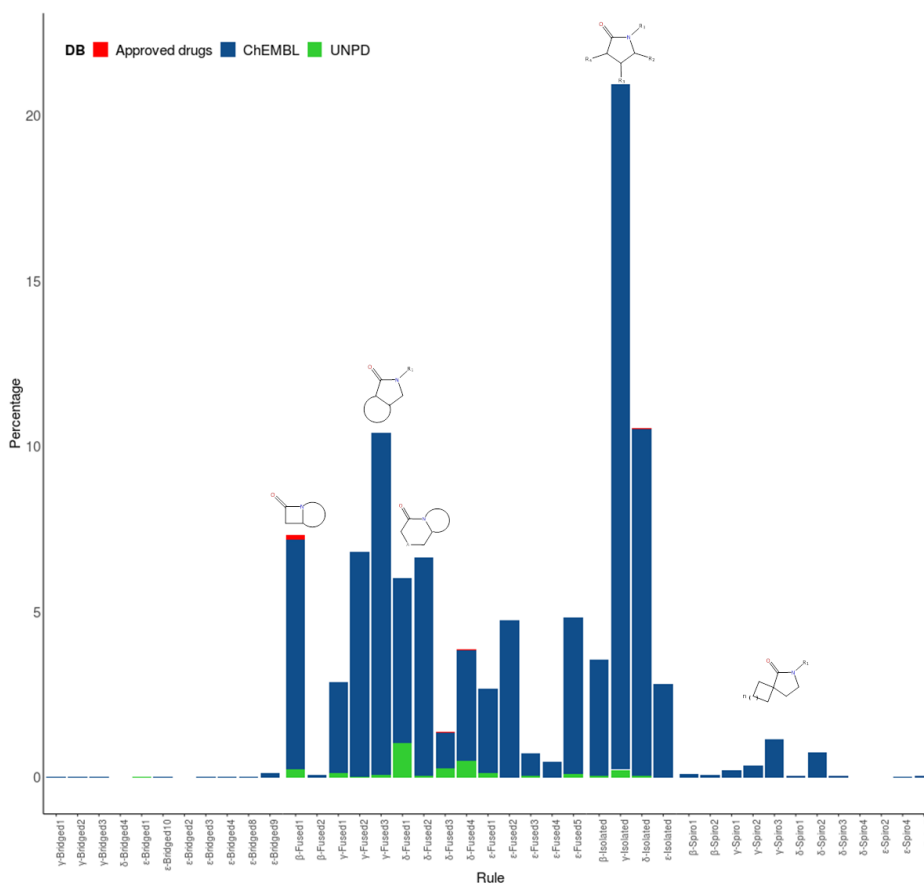


Figura 4.4: Presencia de lactamas *isolated*, *bridged*, *fused* y *spiro* en las bases de datos UNPD, fármacos aprobados y ChEMBL.

4.3. Perfil biológico de las lactamas

El perfil biológico de las lactamas provenientes de ChEMBL se resume en la Tabla 4.1. Se puede observar el número de dianas biológicas en los que se han analizado las lactamas según su clasificación química, el número de compuestos y *scaffolds* únicos, y el porcentaje de compuestos activos. La Tabla 4.1 indica que las lactamas *gamma-isolated* y *delta-fused* tienen un perfil de actividad más extenso en comparación con las lactamas *spiro* y *bridged* que se han probado en un número menor de dianas biológicas. Las dianas biológicas humanas con el mayor número de compuestos activos son los purinoreceptores P2X 2 y 7, la proteína Mdm-2, TNF-alpha, el factor de coagulación

X, el receptor metabotrópico de glutamato 5, y los bromodominios 4, 3 y 2. De la misma forma, las dianas biológicas no humanas con un mayor número de compuestos activos son *Plasmodium falciparum*, el virus de la hepatitis C, la integrasa (HIV-1), la beta lactamasa (*Enterobacter cloacae* y *E. coli*), y el virus del herpes humano.

Tabla 4.1 Resumen de la información biológica y química de las lactamas clasificadas por clase química

	<i>beta-fused</i>	<i>beta-isolated</i>	<i>beta-spiro</i>	<i>gamma-fused</i>	<i>gamma-isolated</i>	<i>gamma-spiro</i>	<i>delta-fused</i>	<i>delta-isolated</i>	<i>delta-spiro</i>	<i>epsilon-fused</i>	<i>delta-spiro</i>	<i>delta-bridged</i>
No. de dianas biológicas humanas	52	88	4	501	352	15	351	179	9	190	5	5
No. de dianas biológicas no humanas	94	52	2	172	184	11	153	71	8	76	2	6
% compuestos activos	37.18	43.58	62.74	46.76	43.39	7.87	39.93	48.76	6.32	50.69	62.16	17.20
No. compuestos únicos	627	892	41	4062	3297	198	2406	2042	255	1861	25	78
No. de scaffolds únicos	280	376	19	1638	1658	138	1259	1031	206	926	12	58

Por otra parte, el análisis de EF indica que existen 131 *scaffolds* con una frecuencia mayor a 20 y un EF mayor a 1. De estos, 45 son *gamma-fused*, 21 *delta-fused*, 18 *epsilon-fused*, 19 *gamma-isolated*, 16 *delta-isolated*, 10 *beta-isolated*, y 5 *beta-fused*. Por otro lado, se encontraron 2934 *scaffolds* con un EF mayor que 1 pero con una cantidad de análogos menor a 20. La mayoría de estos *scaffolds* son *gamma-fused* y *gamma-isolated* con 652 y 492 compuestos, respectivamente. También se encontró una gran cantidad de compuestos *epsilon-fused* (494), *delta-isolated* (473) y *delta-fused* (471). Los *scaffolds* de compuestos *spiro* (*beta*, *gamma*, *delta* y *epsilon*) y los compuestos *epsilon-bridged* destacan ya que como se mencionó, han sido poco explorados en síntesis orgánica. No obstante, algunos de estos *scaffolds* muestran actividad contra ciertos objetivos biológicos. Por ejemplo, los compuestos con *scaffolds beta-spiro* han mostrado actividad para el receptor Vanilloide y solo se han informado 11 *scaffolds* con menos de 5 compuestos cada uno. Dentro del grupo de *spiro* lactamas, *delta-spiro* y *gamma-spiro* han sido los más explorados, sin embargo, solo se identificaron 17 *scaffolds delta-spiro* y 10 *gamma-spiro* con EF mayor que 1. Los objetivos potenciales para los *scaffolds delta-spiro* son el receptor de quimiocina CC-5, el virus de inmunodeficiencia humana tipo 1, el receptor de angiotensina II tipo 1, el virus de la hepatitis C y la 11- β -hidroxiesteroide deshidrogenasa tipo 1 (11- β HD1). De igual forma, *scaffolds gamma-spiro* muestran actividad contra el receptor de proteína G acoplado 44, angiotensina II tipo 1a, la integrina II tipo 1a, la integrina alfa-11b/beta-3, la cinasa 4 dependiente de ciclina, el canal de calcio tipo T regulado por voltaje y la aldosa reductasa. Para la clase de *epsilon-spiro*, solo se identificó un *scaffold* activo contra el transportador de norepinefrina. Los compuestos con *scaffolds epsilon-bridged* han mostrado actividad en objetivos como la cinasa 4 dependiente de ciclina, el virus de inmunodeficiencia humana tipo 1 y la proteína integrina alfa-V/beta-3; para esta clase de compuestos solo se han identificado 11 *scaffolds*, los cuales también tienen menos de 5 análogos cada uno. En la Figura 4.5 se describen ejemplos de los *scaffolds* más frecuentes y con EF mayor a 1 (Figura 4.5a), así como ejemplos de *scaffolds* con menos de 20 análogos informados pero que muestran actividad contra blancos biológicos específicos (Figura 4.5b).

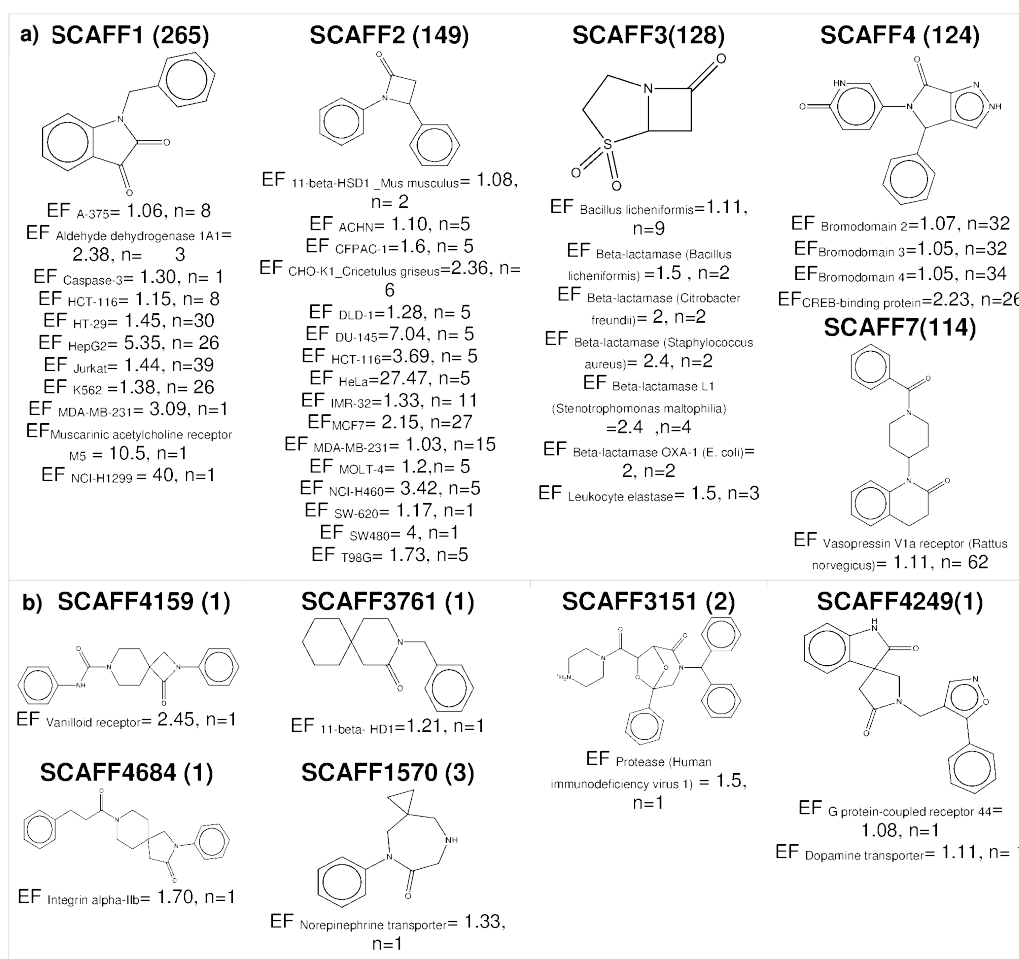


Figura 4.5: **a)** *Scaffolds* más frecuentes con EF mayor que 1 y **b)** *Scaffolds* menos frecuentes con EF mayor que 1. Se indica el ID de cada *scaffold* y entre paréntesis la frecuencia de ese *scaffold* en la base de datos. Se muestran los factores de enriquecimiento (EF) y el número (n) de compuestos totales con el *scaffold* que se han probado contra el objetivo particular.

4.4. Diseño y enumeración de bibliotecas químicas enfocadas en diversidad

En la Tabla 4.2 se resume el número de compuestos obtenidos en cada fase de la estrategia (B/C/P), para la biblioteca de *building blocks* de Enamine. Los compuestos obtenidos a partir de reacciones intramoleculares se separaron en macrociclos, compuestos con anillos de más de 7 miembros y no macrociclos, compuestos con anillos de entre 3 y 7 miembros. La Tabla 4.2 indica que se obtuvieron 9252 lactamas no macrocíclicas usando aminas primarias y ácidos carboxílicos en la reacción de acoplamiento, y 11345 lactamas cuando se usaron aminas secundarias y ácidos carboxílicos. En general, las reacciones de esterificación, lactamización y acoplamiento cruzado generaron la mayor cantidad de compuestos. Las reacciones como la metátesis de eninos y la reacción de química click generaron compuestos macrocíclicos principalmente. Cabe señalar que esta estrategia, al menos con la biblioteca de *building blocks* de Enamine, puede producir una gran di-

versidad de *scaffolds*, lo que se puede observar al comparar el número de compuestos únicos con el número de *scaffolds* únicos.

Los compuestos diseñados bajo este enfoque se clasificaron en la misma forma en la que fueron agrupadas las lactamas de fármacos aprobados, UNPD y ChEMBL. Los resultados se describen en la Tabla 4.3. Como se observa, con esta estrategia solo se obtienen lactamas de seis y siete miembros (*delta* y *epsilon* lactamas), siendo esta última clase la que contiene mayor número de compuestos. En total, utilizando la estrategia (B/C/P) se obtuvieron 4682 *epsilon*-bridged, 3310 *epsilon*-spiro, 361 *delta*-bridged y 1733 *delta*-spiro, que son las clases de lactamas que hasta ahora se han explorado menos en la síntesis orgánica. Ejemplos de compuestos diseñados bajo este enfoque se muestran en la Figura 4.6.

Tabla 4.2 Número de compuestos obtenidos en cada fase de la estrategia (B/C/P), para los *building blocks* de Enamine.

<i>Building blocks</i>	437,625 compuestos únicos			
Building blocks con más de un grupo funcional	117,700 compuestos			
<i>Coupling</i>	R-COOH + R-NH ₂ 1,000,000		R-COOH + R-NH-R 1,000,000	
<i>Pairing</i>				
Reacción	Macrociclos	No macrociclos	Macrociclos	No macrociclos
Lactonización	16525	4224	25099	6533
Lactamización	7444	1962	8633	2351
Síntesis de Williamson	4181	512	4827	1212
Acoplamiento cruzado Buchwald-Hartwig	12308	2353	9849	1044
Aminación reductiva	481	190	500	145
Condensación de alcoholes	0	0	0	0
Metátesis de alquenos	1	1	1	0
Metátesis de eninos	452	12	847	63
Reacción de química click	81	0	210	1
Compuestos únicos	41472	9252	49965	11345
Scaffolds únicos	37966	7504	45081	9382

Tabla 4.3. Número de lactamas por clase.

Clase	Amina primaria + ácido carboxílico	Amina secundaria + ácido carboxílico
<i>delta-isolated</i>	1490	1446
<i>delta-bridged</i>	5	356
<i>delta-fused</i>	742	1353
<i>delta-spiro</i>	738	995
<i>epsilon-isolated</i>	1942	1832
<i>epsilon-bridged</i>	1631	1679
<i>epsilon-fused</i>	3037	5045
<i>epsilon-spiro</i>	1631	1679

Building block 1	Building block 2	Producto de la reacción de coupling	Reacción de pairing	Producto de la reacción de pairing
			Lactamización	
			Acoplamiento cruzado Buchwald-Hartwig	
			Lactonización	
			Síntesis de Williamson	
			Reacción química click	
			Metátesis de alquenos	
			Metátesis de eninos	

Figura 4.6: Ejemplos de lactamas diseñadas bajo un enfoque DOS.

Capítulo 5

Conclusiones y Perspectivas

5.1. Conclusiones

Los PNs proporcionan un punto de partida útil y validado evolutivamente para la identificación y diseño de nuevas moléculas bioactivas. En este proyecto, los análisis quimioinformáticos realizados permitieron evaluar el espacio químico, la diversidad y la complejidad de diferentes bases de PNs. Los resultados mostraron que varias colecciones de PNs como Universal Natural Product Database (UNPD), PNs de medicina tradicional China (TCM) y los PNs marinos, cubren una región del espacio químico más amplia que los fármacos aprobados para uso clínico. No obstante, en general, la mayoría de los PNs tienen aproximadamente la misma cobertura de espacio químico que los fármacos aprobados. También se concluyó que la diversidad y la complejidad estructural varía de acuerdo al origen de los PNs. En general, las bases de PNs marinos, NuBBE_{DB} y metabolitos de hongos son las bases de datos más diversas y que representan una fuente prometedora para estudios de cribado virtual y el posterior descubrimiento de fármacos.

Estructuras de relevancia médica como acetales bicíclicos y lactamas se encuentran ampliamente distribuidos en las bases de PNs. Su identificación y clasificación permitió identificar a los *scaffolds* más frecuentes y aquellos que representan una gran oportunidad en el diseño de fármacos, ya que son estructuras con potencial biológico pero que han sido poco explorados en el área de síntesis. Los resultados del análisis biológico y quimioinformático de las lactamas indicaron que las *spiro* y *bridged* lactamas son las clases con menor número de compuestos y *scaffolds* únicos.

En cuanto al diseño de la biblioteca de lactamas, la estrategia de B/C/P fue muy útil para generar lactamas con una gran diversidad de *scaffolds*. Con la aplicación de esta estrategia se obtuvo un gran número de *spiro* y *bridged* lactamas, principalmente de seis y siete miembros.

5.2. Perspectivas

Se pretende evaluar la novedad y diversidad de las lactamas obtenidas bajo el enfoque DOS. Asimismo, en el grupo del Dr. Andrea Trabocchi (Universidad de Florencia, Italia), se realizará la validación experimental de los resultados obtenidos mediante la síntesis de algunos compuestos diseñados en este trabajo. Otra perspectiva es que las bases de PNs analizadas en este trabajo, así como así como de la biblioteca de lactamas que fue generada bajo un enfoque DOS, serán utilizadas en estudios de cribado virtual. En cuanto a la metodología, se plantean utilizar los flujos de trabajo generados en este proyecto para el análisis de otros andamios de relevancia farmacéutica y el posterior diseño y síntesis de nuevas bibliotecas químicas.

Bibliografía

1. Thomford NE, Senthebane DA, 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.
2. Newman, D. J.; Cragg, G. M. Natural Products as Sources of New Drugs from 1981 to 2014. *J. Nat. Prod.* **2016**, 79, 629–661.
3. Tetko IV, Engkvist O, Koch U, Reymond JL, Chen H. BIGCHEM: Challenges and Opportunities for Big Data Analysis in Chemistry. *Mol. Inform.* **2016**, 35, 615–621.
4. van Hattum H, Waldmann H. Biology-Oriented Synthesis: Harnessing the Power of Evolution. *J. Am. Chem. Soc.* **2014**, 136, 11853–11859.
5. Sukuru SCK, Jenkins JL, Beckwith REJ, Scheiber J, Bender A, Mikhailov D, Davies JW, Glick M. Plate-Based Diversity Selection Based on Empirical HTS Data to Enhance the Number of Hits and Their Chemical Diversity. *J. Biomol. Screen.* **2009**, 14, 690–699.
6. 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, 2099–2111.
7. Pilon-Jiménez BA, Saldívar-González FI, Díaz-Eufracio BI, Medina-Franco JL. BIOFACQUIM: A Mexican Compound Database of Natural Products. *Biomolecules.* **2019**, 9, 31.
8. Xie T, Song S, Li S, Ouyang L, Xia L, Huang J. Review of Natural Product Databases. *Cell Prolif.* **2015**, 48, 398–404.
9. Gasteiger, J. Chemoinformatics: Achievements and Challenges, a Personal View. *Molecules.* **2016**, 21, 151.
10. Willett, P. Chemoinformatics: A History. *WIREs Comput Mol Sci.* **2011**, 1, 46–56.
11. Foodinformatics - Applications of Chemical Information to Food Chemistry | Karina Martinez-Mayorga | Springer <https://www.springer.com/us/book/9783319102252> (accessed Mar 6, 2019).
12. Senderowitz H, Tropsha A. Materials Informatics. *J. Chem. Inf. Model.* **2018**, 58, 2377–2379.
13. Gillet, V. J. Applications of Chemoinformatics in Drug Discovery. In *Biomolecular and Bioanalytical Techniques*; Ramesh, V., Ed.; John Wiley Sons, Ltd: Chichester, UK, **2019**; Vol. 33, pp 17–36.

14. Stahura FL, Godden JW, Xue L, Bajorath J. Distinguishing between Natural Products and Synthetic Molecules by Descriptor Shannon Entropy Analysis and Binary QSAR Calculations. *J. Chem. Inf. Comput. Sci.* **2000**, 40, 1245–1252.
15. Saldívar-González FI, Pilon-Jiménez BA, Medina-Franco JL. Chemical Space of Naturally Occurring Compounds. *Physical Sciences Reviews.* **2018**, 0, 525.
16. Pascolutti M, Quinn RJ. Natural Products as Lead Structures: Chemical Transformations to Create Lead-like Libraries. *Drug Discov. Today.* **2014**, 19, 215–221.
17. Grabowski K, Baringhaus KH, Schneider G. Scaffold Diversity of Natural Products: Inspiration for Combinatorial Library Design. *Nat. Prod. Rep.* **2008**, 25, 892–904.
18. López-Pérez JL, Theron R, del Olmo E, Díez D, Vaquero M, Adserias JF. Application of Chemoinformatics to the Structural Elucidation of Natural Compounds. In *Intelligent Data Engineering and Automated Learning – IDEAL 2006*; Springer Berlin Heidelberg, **2006**; 1150–1157.
19. Lai Z, Tsugawa H, Wohlgemuth G, Mehta S, Mueller M, Zheng Y, Ogiwara A, Meissen J, Showalter M, Takeuchi K, et al. Identifying Metabolites by Integrating Metabolome Databases with Mass Spectrometry Cheminformatics. *Nat. Methods.* **2018**, 15, 53–56.
20. Cragg GM, Newman DJ. Biodiversity: A Continuing Source of Novel Drug Leads. *Pure and Applied Chemistry.* **2005**, 7–24.
21. Prachayasittikul V, Worachartcheewan A, Shoombuatong W, Songtawee N, Simeon S, Prachayasittikul V, Nantasenamat C. Computer-Aided Drug Design of Bioactive Natural Products. *Curr. Top. Med. Chem.* **2015**, 15, 1780–1800.
22. Rodrigues T, Reker D, Schneider P, Schneider G. Counting on Natural Products for Drug Design. *Nat. Chem.* **2016**, 8, 531–541.
23. Hert J, Irwin JJ, Laggner C, Keiser MJ, Shoichet BK. Quantifying Biogenic Bias in Screening Libraries. *Nat. Chem. Biol.* **2009**, 5, 479–483.
24. Perron F, Albizati KF. Chemistry of Spiroketal. *Chem. Rev.* **1989**, 89, 1617–1661.
25. Kiyota H. Synthesis of Marine Natural Products with Bicyclic And/or Spirocyclic Acetals. In *Marine Natural Products*; Kiyota H, Ed.; Springer Berlin Heidelberg: Berlin, **2006**, 65–95.
26. Reymond JL. The Chemical Space Project. *Acc. Chem. Res.* **2015**, 48, 722–730.
27. Visini R, Awale M, Reymond JL. Fragment Database FDB-17. *J. Chem. Inf. Model.* **2017**, 57, 700–709.
28. Chevillard F, Kolb P. SCUBIDOO: A Large yet Screenable and Easily Searchable Database of Computationally Created Chemical Compounds Optimized toward High Likelihood of Synthetic Tractability. *J. Chem. Inf. Model.* **2015**, 55, 1824–1835.
29. Walters, W. P.; Patrick Walters, W. Virtual Chemical Libraries. *J Med Chem.* **2019**, 62, 1116–1124.
30. Pye CR, Bertin MJ, Lokey RS, Gerwick WH, Linington RG. Retrospective Analysis of Natural Products Provides Insights for Future Discovery Trends. *Proc. Natl. Acad. Sci. U. S. A.* **2017**, 114, 5601–5606.

31. Hoffmann T, Gastreich M. The next Level in Chemical Space Navigation: Going Far beyond Enumerable Compound Libraries. *Drug Discov. Today*. **2019**. En prensa. Doi: 10.1016/j.drudis.2019.02.013.
32. O'Connell KMG, Warren RJ, Spring DR. The Basics of Diversity-Oriented Synthesis. In *Diversity-Oriented Synthesis*. Trabocchi A(Ed.). **2013**. <https://doi.org/10.1002/9781118618110.ch1>.
33. Nielsen TE, Schreiber SL. Towards the Optimal Screening Collection: A Synthesis Strategy. *Angew. Chem. Int. Ed Engl*. **2008**, *47*, 48–56.
34. Welsch ME, Snyder SA, Stockwell, BR. Privileged Scaffolds for Library Design and Drug Discovery. *Curr. Opin. Chem. Biol*. **2010**, *14*, 347–361.
35. 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.
36. González-Medina M, Prieto-Martínez FD, Naveja JJ, Méndez-Lucio O, El-Elimat T, Pearce CJ, Oberlies NH, Figueroa M, Medina-Franco JL. Chemoinformatic Expedition of the Chemical Space of Fungal Products. *Future Med. Chem*. **2016**, *8*, 1399–1412.
37. Pilon AC, Valli M, Dametto AC, Pinto MEF, Freire RT, Castro-Gamboa I, Andricopulo AD, Bolzani VS. NuBBEDB: An Updated Database to Uncover Chemical and Biological Information from Brazilian Biodiversity. *Sci. Rep*. **2017**, *7*, 7215.
38. 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.
39. Gu J, Gui Y, Chen L, Yuan G, Lu HZ, Xu X. Use of Natural Products as Chemical Library for Drug Discovery and Network Pharmacology. *PLoS One*. **2013**, *8*, e62839.
40. Lipinski CA, Lombardo F, Dominy BW, Feeney PJ. Experimental and Computational Approaches to Estimate Solubility and Permeability in Drug Discovery and Development Settings. *Adv. Drug Deliv. Rev*. **2001**, *46*, 3–26.
41. Veber DF, Johnson SR, Cheng HY, Smith BR, Ward KW, Kopple KD. Molecular Properties That Influence the Oral Bioavailability of Drug Candidates. *J. Med. Chem*. **2002**, *45*, 2615–2623.
42. Ripley BD. The R Project in Statistical Computing. MSOR Connections. *The newsletter of the LTSN Maths, Stats OR Network*. **2001**, *1*, 23–25.
43. 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.
44. KNIME - Open for Innovation <https://www.knime.com/> (accessed Mar 26, 2019).
45. Bemis GW, Murcko MA. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem*. **1996**, *39*, 2887–2893.
46. 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.
47. Medina-Franco JL, Maggiora GM. MOLECULAR SIMILARITY ANALYSIS. In *Chemoinformatics for Drug Discovery*. **2013**, 343–399. <https://doi.org/10.1002/9781118742785.ch15>.

48. Daylight Theory: SMARTS - A Language for Describing Molecular Patterns <http://www.daylight.com/dayhtml/doc/theory/theory.smarts.html> (accessed Mar 28, 2019).
49. Welcome to ZINC Is Not Commercial - A database of commercially-available compounds <http://zinc.docking.org/> (accessed Mar 28, 2019).
50. Wishart DS, Feunang YD, Guo AC, Lo EJ, Marcu A, Grant JR, Sajed T, Johnson D, Li C, Sayeeda Z, et al. DrugBank 5.0: A Major Update to the DrugBank Database for 2018. *Nucleic Acids Res.* **2018**, 46, D1074–D1082.
51. Gaulton A, Hersey A, Nowotka M, Bento AP, Chambers J, Mendez D, Motow P, Atkinson F, Bellis LJ, Cibrián-Uhalte E, et al. The ChEMBL Database in 2017. *Nucleic Acids Res.* **2017**, 45, D945–D954.
52. Lenci E, Menchi G, Saldívar-Gonzalez FI, Medina-Franco JL, Trabocchi A. Bicyclic Acetals: Biological Relevance, Scaffold Analysis, and Applications in Diversity-Oriented Synthesis. *Org. Biomol. Chem.* **2019**, 17, 1037–1052.
53. 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.
54. Uchida T, Rodriguez M, Schreiber SL. Skeletally Diverse Small Molecules Using a Build/Couple/ Pair Strategy. *Organic Letters.* **2009**, 11, 1559–1562.
55. Denis. Building Blocks - Enamine <https://enamine.net/building-blocks> (accessed Mar 15, 2019).

Chemical Space and Diversity of the NuBBE Database: A Chemoinformatic Characterization

Fernanda I. Saldívar-González,[†] Marília Valli,[‡] Adriano D. Andricopulo,[§] Vanderlan da Silva Bolzani,[‡] and José L. Medina-Franco^{*,†}

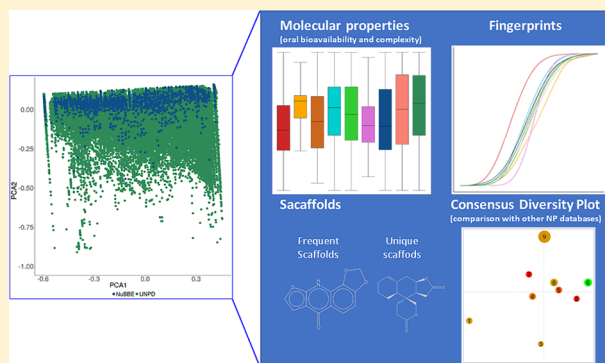
[†]School of Chemistry, Department of Pharmacy, Universidad Nacional Autónoma de México, Avenida Universidad 3000, Mexico City 04510, Mexico

[‡]Nuclei of Bioassays, Biosynthesis and Ecophysiology of Natural Products (NuBBE), Department of Organic Chemistry, Institute of Chemistry, Sao Paulo State University - UNESP, 14800-060 Araraquara, Sao Paulo, Brazil

[§]Laboratório de Química Medicinal e Computacional (LQMC), Centro de Pesquisa e Inovação em Biodiversidade e Fármacos, Institute of Physics of Sao Carlos, University of Sao Paulo - USP, 13563-120 Sao Carlos, Sao Paulo, Brazil

Supporting Information

ABSTRACT: NuBBE_{DB} is the first library of natural products of Brazilian biodiversity. It includes a large variety of classes of compounds and structural types of secondary metabolites of plants, fungi, insects, marine organisms, and bacteria. So far the chemical diversity and complexity of NuBBE_{DB} have not been characterized in a systematic and detailed manner. Herein, we report a comprehensive chemoinformatic analysis of the most current version of NuBBE_{DB}. As part of the characterization, NuBBE_{DB} was compared with several databases of natural products in terms of structural diversity and complexity. Results of the analysis showed that NuBBE_{DB} is diverse in terms of structural fingerprints, distribution of chemical scaffolds, and molecular properties. In addition, the results of the visualization of chemical space support quantitatively that NuBBE_{DB} is a promising source of molecules for drug discovery and medicinal chemistry.



INTRODUCTION

The growth and evolution of our civilization have been based on the use of biodiversity. Human survival was much sustained by the use of plant species, especially for nutrition and medicinal uses. Natural products are the most traditional source for the design and development of new drugs.^{1–3} Taking into account all drugs approved worldwide, 67% are either a natural product, a semisynthetic derivative, or a macromolecule isolated from an organism or have a pharmacophore group inspired by a natural product.⁴

Natural products researchers have been studying the medicinal properties of plants, and great advances regarding biosynthesis, ecology, and biological properties were achieved in the last century.⁵ The biodiversity of tropical and equatorial environments is plentiful and could offer a particularly rich potential in the search for biologically active compounds to be used as models for drug discovery and medicinal chemistry.⁶

Brazil is the country with one of the greatest biodiversities in the world, accounting for more than ca. 10% of all living species.⁷ It comprises six terrestrial biomes: the Amazonian rainforest, the Caatinga, the Cerrado (Savanna), the Atlantic forest, the Pantanal, and the Pampas. This extraordinary biodiversity remains underexplored, and the chemical diversity

could be used for the development of bioproducts, including pharmaceuticals, cosmetics, food supplements, and agricultural pesticides.⁸ The Atlantic Forest and the Cerrado are regarded as hotspots of biodiversity due to an enhanced loss of biodiversity caused by urbanization, agriculture, and livestock. Many species from these biomes are endangered and near extinction, and the chemistry, genes, and biological properties of them would also be lost.^{8–10} The chemical diversity of the flora and fauna of the Brazilian Biomes is revealed by a diversity of compound classes and structural types of secondary metabolites from plants, fungi, insects, marine organisms, and bacteria.¹¹

The scientific information published in more than 50 years of studies on Brazilian biodiversity becomes easier to access when standardized, certified, and organized in a database. As an initial effort, the NuBBE_{DB} was created, a database of compounds from Brazilian biodiversity with the objective of organizing its chemical, biological, and pharmacological information.¹² NuBBE_{DB} presently contains data of 2218 compounds, an estimative of 5% of the published information

Received: September 11, 2018

Published: December 3, 2018

on natural products isolated/identified from species collected in the Brazilian territory. NuBBE_{DB} is an ongoing project and already has good coverage because it comprises compounds from plants, marine organisms, and fungi species from all six Brazilian biomes, and the chemical space is notably diverse and rich, with compounds from several molecular classes, such as flavonoids, alkaloids, terpenes, iridoids, lignans, etc.

Chemical (metabolic class, chemical structure, physicochemical properties, common and IUPAC name, and molecular mass), biological (species, geographic location, and biological activities), pharmacological, and spectroscopic data (molar mass and nuclear magnetic resonance) are provided.¹³ NuBBE_{DB} is accessible online for free, and a search can be filtered by any properties and a combination of criteria.¹⁴ The organization and mapping of this information on a systematized and correlated system allow the access and use of the benefits of biodiversity. This information center significantly reduces the time spent in scientific studies and processes involving technological research. Consequently, due to the simplified access of this molecular heritage the industrial segment is more encouraged to invest in the technological development and research of products. In this sense, NuBBE_{DB} may assist in the development of different fields of science, technological development of biodiversity products with high added value, and public policies, that is, to bring benefits both to science and to strengthen the bioeconomy. The scientific community is aware of the significance of NuBBE_{DB}, and several studies were recently published either reporting its use or its importance.^{13,15–21} However, a comprehensive chemoinformatic analysis in terms of diversity and chemical complexity of the NuBBE_{DB} has not been carried out. This type of chemoinformatics analysis plays an important role as a guide for the acquisition of compounds and in the selection of databases for the detection and optimization of leads.¹⁶

The goal of this paper was to perform a quantitative characterization of the chemical diversity of NuBBE_{DB} and compare it with other databases of natural products. The analysis was carried out using multiple criteria including physicochemical properties of pharmaceutical relevance, diversity of scaffolds, diversity based on fingerprints, structural complexity, and visual representation of the chemical space. The global diversity of each database was assessed using the Consensus Diversity Plots developed recently.²²

METHODS

The chemical diversity and chemical space coverage of NuBBE_{DB} were analyzed using descriptors and quantification approaches used recently to study the diversity of other natural products data sets, for example, natural products from Panama,³⁷ metabolites from fungi,³⁸ and natural products from plants.³⁹ The diversity and chemical space of NuBBE_{DB} were compared to other compound databases that were used as reference. NuBBE_{DB} was compared with metabolites from fungi,³⁸ metabolites from cyanobacteria, natural products commercially available (MEGx), and a data set of marine compounds reported recently⁴⁰ and with the compounds in the Universal Natural Products Database (UNPD) and the Traditional Chinese Medicine Database@Taiwan (TCM) (*vide infra*). Other reference collections were a database of semisynthesis compounds (NATx) and a database of drugs approved by the United States Food and Drug Administration (FDA). Of note, similar to NuBBE_{DB}, the reference data set of marine compounds analyzed in this work had not been

analyzed in terms of molecular properties, structural diversity, and complexity.

All calculations in this work were done with KNIME²³ and R. An open Web-based implementation of several diversity analyses used in this work is implemented in the server Platform for Unified Molecular Analysis (PUMA)²⁴ freely available at D-Tools (<https://www.difacqum.com/d-tools/>).²⁵

Compound Data Sets. The compound data sets used in this study were summarized in Table 1. Prior to analysis, each

Table 1. NuBBE_{DB} and Compound Databases Used as Reference

database	size ^a	reference
cyanobacteria metabolites	473	in-house
fungi NP	206	38
marine NP	6253	40
purified natural product screening compounds (MEGx)	4103	ac-discovery.com
semisynthetics (NATx)	26318	ac-discovery.com
approved drugs	1806	www.drugbank.ca
NuBBE _{DB}	2214	13
Traditional Chinese Medicine Database@Taiwan (TCM)	17986	52
Universal Natural Products Database (UNPD)	209574	53

^aNumber of unique compounds after data curation.

molecule was prepared using the node Wash provided by Molecular Operating Environment (MOE).²⁶ This node disconnects salts of metals, removes simple components, recalculates states of protonation, determines wedge bonds for bonds from chiral centers, and calculates missing chiral parities from existing wedge bond. With this same program, inorganic compounds were eliminated, as well as repeated compounds. Tautomeric forms were not considered in this study.

Molecular Properties of Pharmaceutical Relevance and Chemical Space. Seven molecular properties of pharmaceutical interest were computed namely, hydrogen-bond donors (HBD), hydrogen-bond acceptors (HBA), partition coefficient octanol/water (xlogP), molecular weight (MW), number of rotatable bonds (RB), topological polar surface area (TPSA), and the fraction of carbon atoms with sp³ hybridization (FCsp³). The statistical comparison of the descriptors was carried out in RStudio. To facilitate the visual representation of the seven molecular descriptors and generate a visual representation of the chemical space, a principal components analysis (PCA)^{27,28} was carried out in the KNIME program. The representation of the chemical space was made in RStudio.

Drug-, Extended Drug-, Lead-, Fragment-, PPI-like, and PAINS Profiling. Compounds in NuBBE_{DB} and the reference databases were analyzed in terms of the number of compounds that falls within each of the six categories: drug-, extended drug-, lead-, fragment-, PPI-like, and PAINS. These categories are characterized by the definition of ranges of physicochemical properties as follows: drug-like (150 < = MW < = 500 Da, xlogP < = 5, Num HBD < = 5, Num HBA < = 10); “extended drug-like” (this is drug-like with additional constraints: drug-like AND rotatable bonds < = 7, TPSA < 150); lead-like (250 < = MW < = 350 Da, xlogP < = 3.5, rotatable bonds < = 7); fragment-like (MW < 300 Da, Num HBD < = 3, Num HBA < = 3, xlogP < = 3); PPI-like (MW > 400 Da, Num Rings > = 4, Num HA > 4, xlogP > 4); PAINS

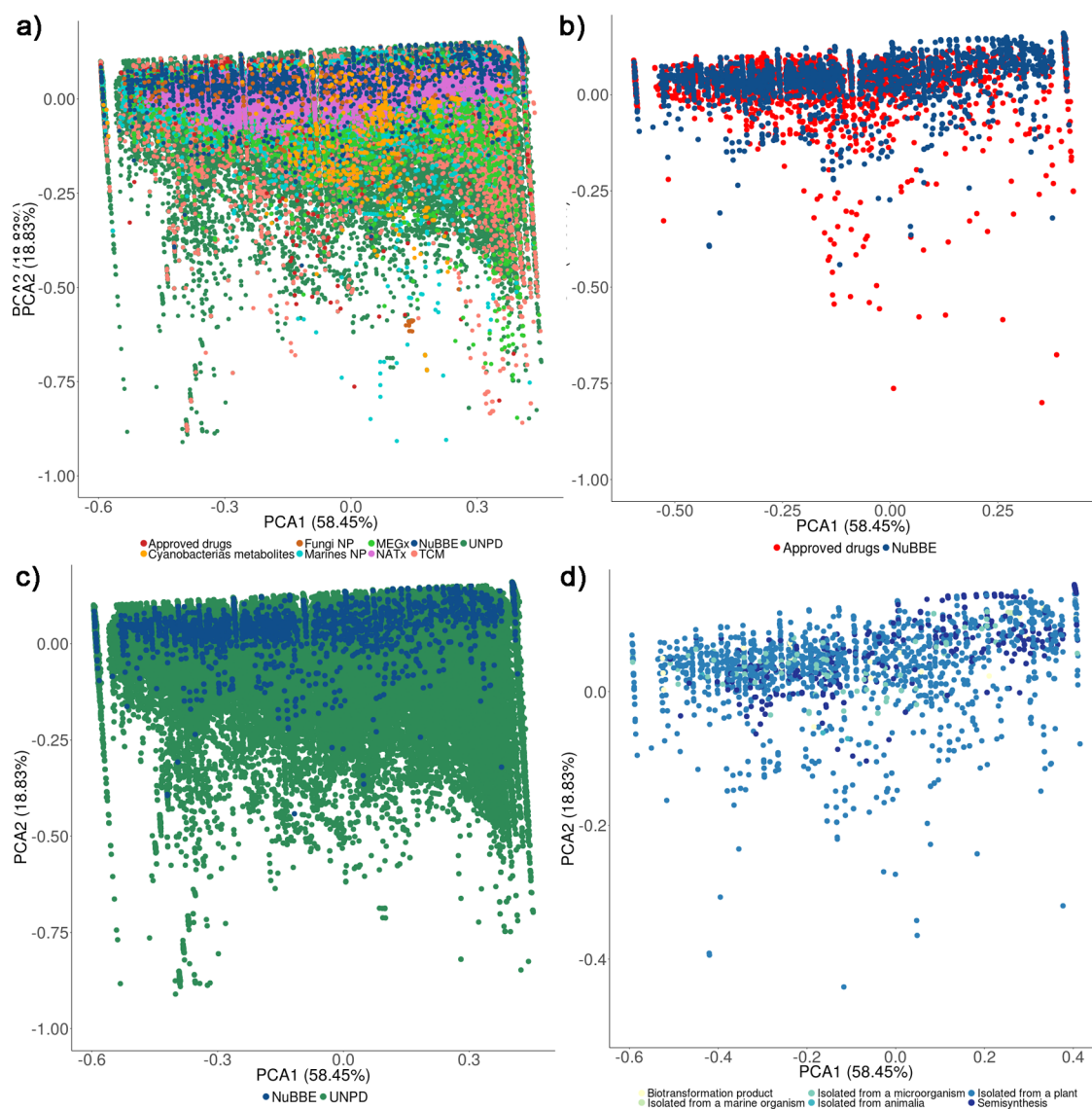


Figure 1. Visual representation of the chemical space: a) Full space and b) NuBBE_{DB} vs Approved drugs c) NuBBE_{DB} vs UNPD and d) NuBBE_{DB} subspace by origin. The visual representation was generated with a principal component analysis of seven physicochemical properties: molecular weight, hydrogen bond donors, hydrogen bond acceptors, the octanol and/or water partition coefficient, topological polar surface area, number of rotatable bonds, and the fraction of carbon atoms with sp³ hybridization (FCsp³).

(based on structural alerts, as defined by Guha et al.).^{29,30} Compound data sets were further analyzed and clustered based on the number of drug-, fragment-, PPI-like, and PAINS compounds.

Molecular Complexity. In addition to MW, the fraction of carbon atoms with sp³ hybridization (FCsp³) and the fraction of chiral carbons (FCC) were computed as metrics of molecular complexity. Overall, large values for these descriptors are associated with larger, more three-dimensional, and greater stereochemical complexity, respectively. Also, the Scaffold Complexity index (*S_i*) defined by Xu³¹ was calculated in KNIME using the node “Scaffold Classification Approach” provided by MOE. *S_i* is composed of four structural descriptors: (1) the maximum number of the smallest set of smallest rings (ssrs), (2) the maximum number of heavy atoms, (3) the maximum number of bonds, where covalent bonds between hydrogen atoms and other atoms are excluded, and (4) the maximum sum of heavy atomic numbers.

Molecular Scaffolds. In this work, scaffolds were generated under the Bemis-Murcko definition using the RDKit nodes available in KNIME.³² The Shannon entropy (SE)³³ of a population of *P* compounds distributed in *n* systems is defined as

$$SE = - \sum_{i=1}^n p_i \log_2 p_i$$

$$p_i = \frac{c_i}{P}$$

where *p_i* is the estimated probability of the occurrence of a specific chemotype *i* in a population of *P* compounds containing a total of *n* acyclic and cyclic systems, and *c_i* is the number of molecules that contain a particular chemotype *c*. To normalize SE to the different *n*, Scaled Shannon Entropy (SSE) is defined as

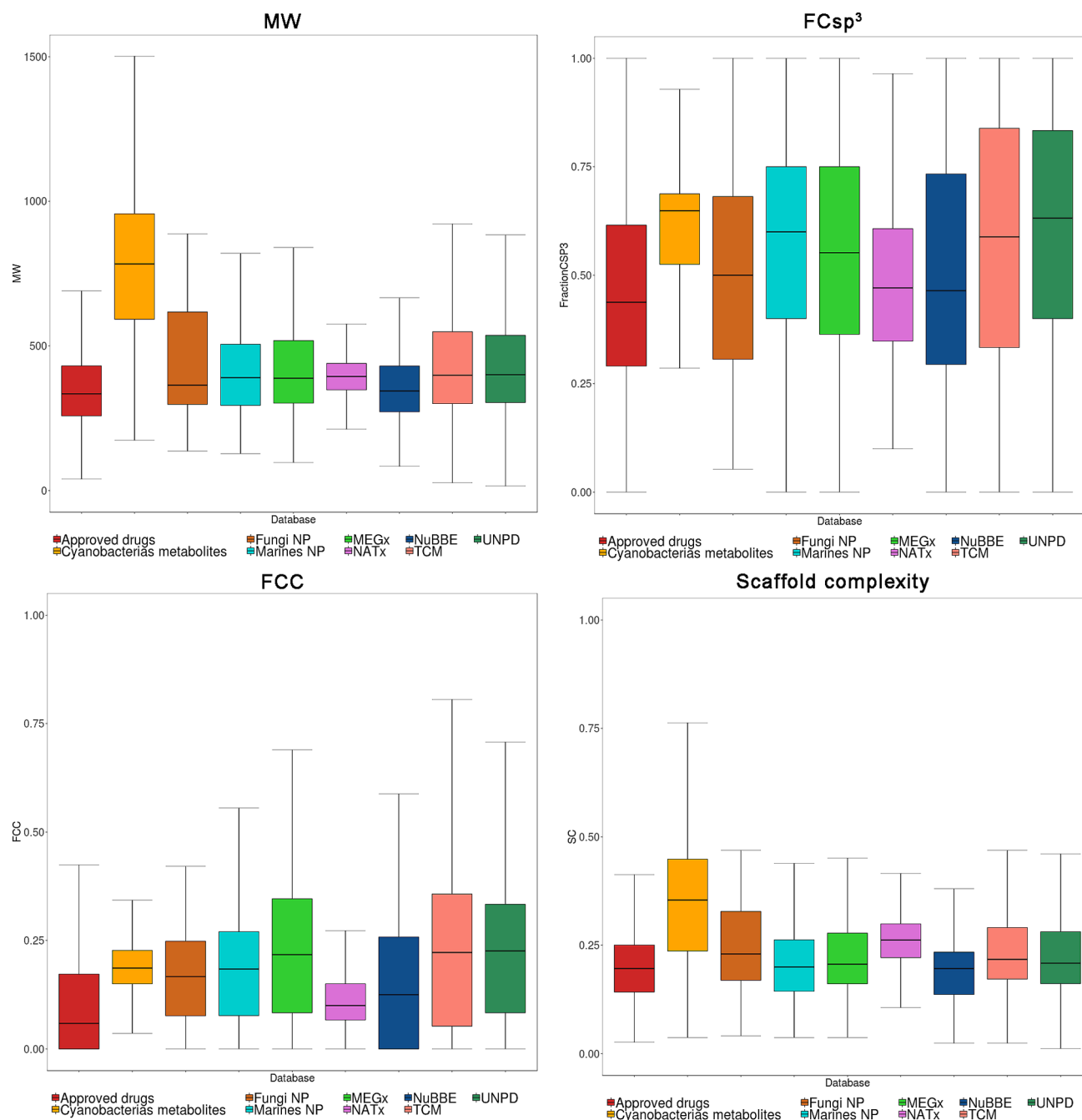


Figure 2. Box plots of the distribution of molecular weight, fraction with sp^3 hybridization (FC_{sp^3}), fraction of chiral carbons (FCC), and Scaffold Complexity index (S_i). Values were computed for NuBBE_{DB} and the eight reference databases.

$$SSE = \frac{SE}{\log_2 n}$$

The SSE value oscillates between 0, when all the compounds have the same chemotype (minimum diversity), and 1.0, when all the compounds are evenly distributed among the n acyclic and/or cyclic systems (maximum diversity).

In order to identify possible unique scaffolds, the N_{sing} of each database was compared with each other. The unique scaffolds were searched in the scaffolds of the ChEMBL database, version 24 (Gaulton et al. 2017) (1,727,112 compounds). In addition, unique scaffolds in NuBBE_{DB} were searched in the Dictionary of Natural Products database (Dictionary of Natural Products 27.1) (274,478 compounds)

to identify scaffolds that were not reported within natural products.

Structural Fingerprints. Extended Connectivity fingerprints radius two (ECFP4) and MACCS keys (166-bits) were calculated for all molecules in KNIME. Based on these two fingerprints, the similarity matrix was calculated with the Tanimoto coefficient.³⁶ Values outside the diagonal of the similarity matrix were used to graph the cumulative distribution function for each of the databases.

Global or Total Diversity. CDPlots²² are two-dimensional graphs where a representative measure of the fingerprint diversity of the data set is plotted on one axis (e.g., X), and the measure of the scaffold diversity is plotted on the second axis (e.g., Y). Each data point in the plot represents a compound

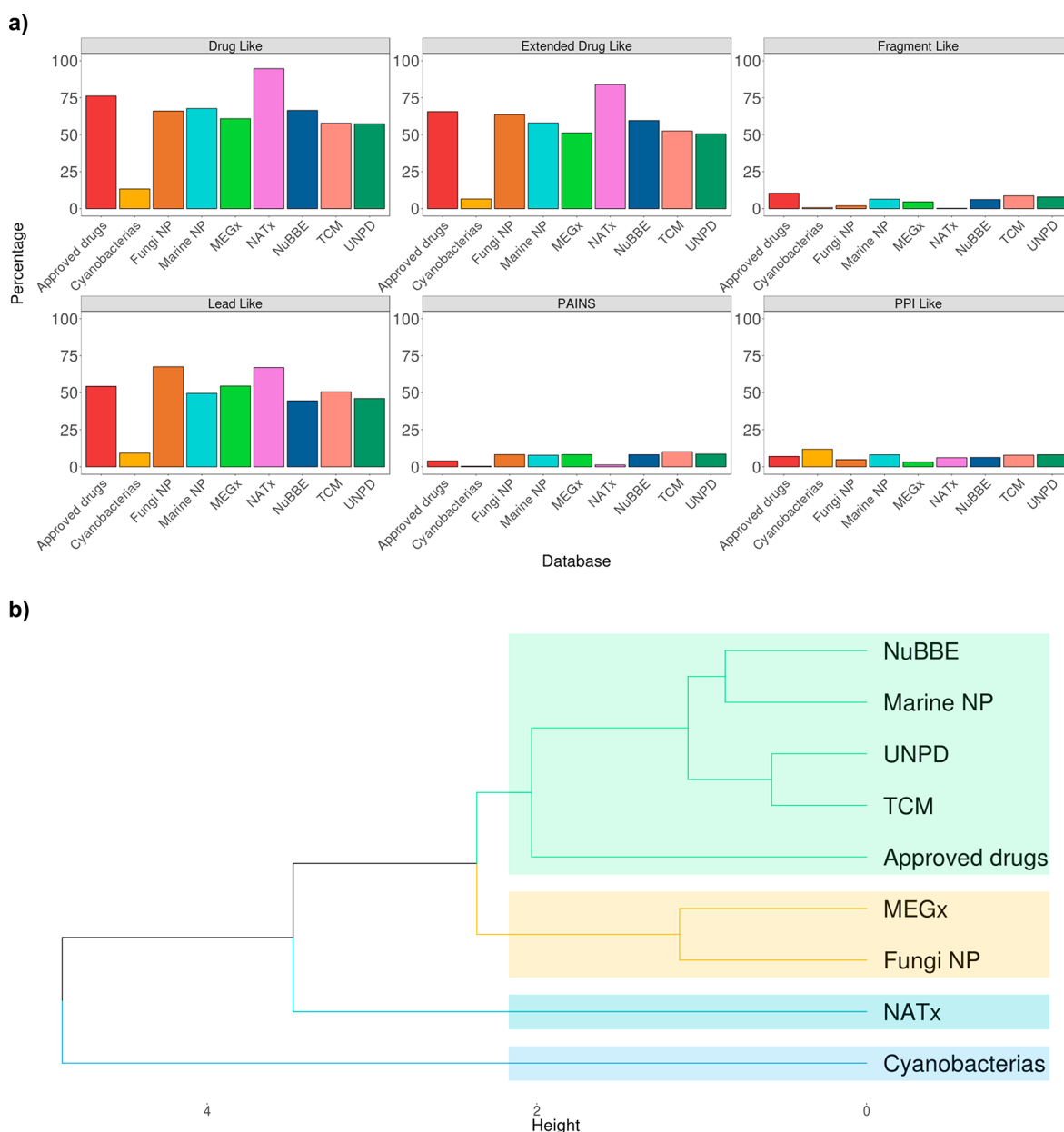


Figure 3. a) Percentage of compounds in each of the six subsets (drug-like, extended drug-like, fragment-like, lead-like, PPI-like, and Pan Assay Interference Compounds - PAINS). b) Hierarchical clustering (average linkage/Euclidean distance) according to the percentage of compounds of uncorrelated variables (drug-like, fragment-like, PPI-like, and PAINS).

data set. Data points are further differentiated by size and color that are used to represent the relative number of compounds in the data set and a different property, respectively. Additional properties of the compound data sets that can be associated with different colors of the data points are, for instance, diversity computed based on physicochemical properties or molecular complexity.³⁷ As described in the [Results and Discussion](#), the CDPlots to compare NuBBE_{DB} with other data sets were generated using the metrics of fingerprint and scaffold diversity computed as described above.

RESULTS AND DISCUSSION

Molecular Properties, Chemical Space. For evaluation of the chemical space of NuBBE_{DB}, seven properties were computed as is it described in the [Methods section](#). These properties are often associated with factors that contribute to a

good oral bioavailability of drugs, as described by Lipinski⁴¹ and Veber.⁴² [Figure 1a](#) shows a visual representation of the chemical space of NuBBE_{DB} and the reference data set based on the seven physicochemical properties. The figure shows that the database of UNPD covers most of the chemical space and is also the database with the greatest diversity in physicochemical properties. In contrast, NuBBE_{DB} occupies a more focused region of the property space, which in turn is included within the space of approved drugs ([Figure 1b](#)) and UNPD ([Figure 1c](#)). The distribution of the seven physicochemical properties for NuBBE_{DB} and the reference data sets is shown in [Figure S1](#) of the Supporting Information. Of note, the box plots show that cyanobacteria metabolites have distinct properties with an average value and interquartile variation for all the calculated descriptors above that of the other databases.

Table 2. Scaffold Diversity Analyses on the Data Sets^a

database	M	N	N/M	N _{sing}	N _{sing} /M	N _{sing} /N	AUC	F ₅₀	SSE10	unique scaffolds ^b
approved drugs	1806	986	0.546	794	0.440	0.805	0.712	0.139	0.961	319 (17.66%)
cyanobacteria	473	178	0.376	95	0.201	0.534	0.750	0.124	0.753	76 (16.06%)
fungi NP	206	116	0.563	74	0.359	0.638	0.672	0.233	0.964	27 (13.10%)
marine NP	6253	2496	0.399	1493	0.239	0.598	0.754	0.117	0.898	1661 (26.56%)
MEGx	4103	1632	0.398	1085	0.264	0.665	0.765	0.095	0.894	857 (20.89%)
NATx	26318	8256	0.314	4455	0.169	0.540	0.779	0.106	0.992	8472 (32.19%)
NuBBE _{DB}	2214	644	0.291	338	0.152	0.524	0.792	0.087	0.930	272 (12.28%)
TCM	17986	17219	0.957	6716	0.373	0.390	0.778	0.079	0.878	377 (2.09%)
UNPD	209574	196728	0.939	40412	0.193	0.205	0.857	0.025	0.812	28753 (13.72%)

^aM: number of molecules, N: number of scaffolds, N_{sing}: number of singletons, AUC: area under the curve, F₅₀: fraction of chemotypes that contains 50% of the data set; SEE10: Scaled Shannon Entropy at the 10 most frequent scaffolds. ^bUnique scaffolds compared to ChEMBL

In a recent study comparing the chemical space of NuBBE_{DB} and other libraries of NPs using the Dictionary of Natural Products (DNP) and approved drugs as reference,⁴³ it can be seen that NuBBE_{DB} is within the DNP and approved drugs space; however, with a difference of other libraries of NPs such as TCM, TCMID (Traditional Chinese Medicine Integrated Database), UNPD, and StreptomeDB (NPs produced by streptomycetes), the chemical space of NuBBE_{DB} is not greater than that of approved drugs, similar to what is observed in this work.

Figure 1d shows only the space covered by compounds in NuBBE_{DB} but further distinguishing the molecules by the source. In other words, Figure 1d depicts the so-called subspace of NuBBE_{DB} visualizing the compounds according to their origin. From this visualization it can be concluded that the largest property diversity in the current version of NuBBE_{DB} is given by the compounds isolated from plants. Another representation of the NuBBE_{DB} subspace is shown in Figure S2 in the Supporting Information, where the compounds are distinguished by the geographical area in which they were isolated. In this representation are the Southeast compounds that present a larger diversity.

From the distribution of the physicochemical properties (Figure S1) and the visual representation of the chemical space (Figure 1), it can be observed that, in general, NuBBE_{DB} and other natural products compared in this work occupy a similar property space similar to drugs, so that their study can lead to identifying new compounds with possible therapeutic activity. Similarly, it is also observed that some of the compounds present in natural products collections occupy regions of the chemical space not yet covered by current drugs and may be useful in virtual screening for therapeutic targets in which molecules with biological activity have not yet been found.

Molecular Complexity. Molecular complexity is becoming an important property to characterize databases, especially natural product databases, which are generally assumed to contain “complex” structures. This concept represents a crucial component in the design of drugs, where it has been associated with selectivity⁴⁴ and safety³⁸ and with the success of compounds in the progress toward clinical development.⁴⁵ Also, molecular complexity is implicated in the design of chemical libraries for virtual screening.⁴⁶ Different approaches have been proposed to evaluate molecular complexity.^{47,48} However, there is still no single or universal method. The molecular complexity of NuBBE_{DB} and reference data sets was quantified by calculating descriptors such as molecular weight, carbon fraction with sp³ hybridization (FCsp³), and fraction of chiral carbons (FCC). The results of the distributions of FCC,

FCsp³, and Si that were used as metrics to quantify complexity are summarized in the box plots of Figure 2 (the summary statistics are in Table S1 of the Supporting Information). Results indicated that NuBBE_{DB} has, in general, a comparable molecular complexity as approved drugs. Notably, of the other natural products data sets analyzed, metabolites from cyanobacteria have the largest complexity according to the molecular weight, FCsp³, and FCC metrics. As discussed above, this is related to tridimensional complex molecules and greater stereochemical complexity, respectively. Likewise, metabolites from cyanobacteria is the data set that has the greatest structural complexity of scaffolds.

Drug-, Extended Drug-, Lead-, Fragment-, PPI-like, and PAINS Profiling. Based on the distribution of molecular properties, several metrics have been proposed to categorize chemical libraries in different subsets with specific characteristics⁴⁹ such as drug-like, extended drug-like, lead-like, fragment-like, or protein–protein interaction (PPI)-like molecules. In addition, the use of filters to eliminate compounds containing PAINS (Pan Assay Interference Compounds) has also been introduced recently: small molecules that are reactive under test conditions and produce false positive signals²⁹ (e.g., the relevance of these alerts in natural products has been discussed in ref 50). Compounds in NuBBE_{DB} and the reference collections were classified into the six subsets: drug-like, extended drug-like, fragment-like, lead-like, PPI-like, and PAINS, Figure 3a. According to this subset classification, all but cyanobacteria metabolites have a similar profile. Notably, NuBBE_{DB} has a large percentage of drug- and extended drug-like compounds and roughly 50% of lead-like molecules. In contrast, NuBBE_{DB} has a small percentage of PAINS molecules. In sharp contrast, cyanobacteria metabolites have a small fraction of drug-, extended drug-, and lead-like molecules with an increased fraction of PPI-like compounds. This can be associated with the overall increased size (as measured by molecular weight) of these compounds. Also, cyanobacteria present few compounds with PAINS alerts. These results, coupled with molecular complexity results, suggest that cyanobacteria compounds have the potential to show objective selectivity in biological assays. It is also noticeable in Figure 3a the high percentage of drug-like, extended drug-like fraction in NATx that is consistent with the assembly of this commercial collection of natural products and semisynthesis compounds.

A hierarchical clustering (average linkage/Euclidean distance) of the nine compound data sets based on the percentage of compounds of each data set in the drug-like, fragment-like, PPI-like, and PAINS categories is shown in Figure 3b. For this

clustering these four categories were selected because they have low correlations. The clustering clearly revealed the overall large profiling of all compared natural product data sets, including NuBBE_{DB}, with the clear exception of cyanobacteria metabolites that has a unique profile among all compared data sets. Also, the hierarchical clustering in Figure 3b shows the close relationship between approved drugs and natural products such as NuBBE_{DB}, Marine NP, TCM, and UNPD.

Scaffold Diversity. The scaffold term is used to describe the central or core structure of a molecule. This criterion was also used to quantify and compare the diversity of NuBBE_{DB} and reference compound collections. Table 2 summarizes the results of the scaffold diversity computed with different metrics. The scaffold count analysis reveals that among the natural product data sets, fungi metabolites are the most diverse, with a scaffold diversity comparable to approved drugs; both databases contain the highest proportion of scaffolds relative to the number of compounds (N/M) and the highest proportion of singletons (unique scaffolds) relative to the number of cyclic systems (N_{sing}/N). In contrast, NuBBE_{DB}, the semisynthetic compounds (NATx), and UNPD present a high redundancy of structures (low number of singletons in relation to the number of different cyclic systems (N_{sing}/N)).

Based on the results of the scaffolds count, the fraction of cyclic systems was plotted against the cumulative fraction of the database (*Cyclic System Retrieval (CSR) graphs*). In this manner, a direct comparison was made with the content and scaffold diversity of the entire databases. Figure 4 shows the

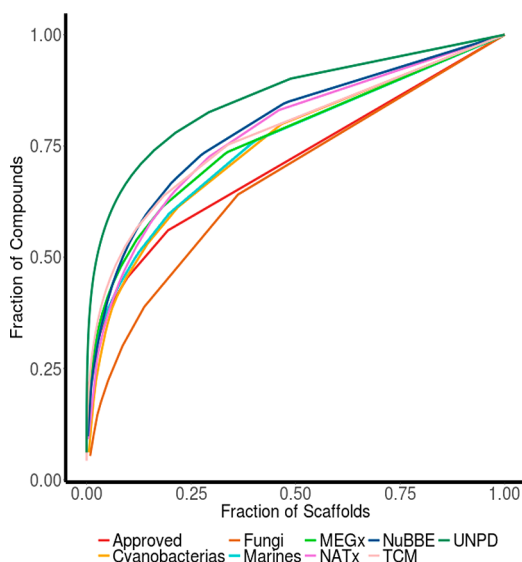


Figure 4. Cyclic system retrieval (CSR) curves for the data sets studied in this work.

CSR curves of all the databases analyzed in this study. The graph indicates that the fungi metabolite database being the closest to a diagonal is the most diverse. The curves for UNPD and NuBBE_{DB} have a rapid increase in its slope, indicating that these data sets have the lowest scaffold diversity. Quantitatively, the CSR curves can be compared using the metrics: area under the curve (AUC) and the fraction of chemotypes that recover 50% of the molecules in the data set (F_{50}). Based on these metrics, the diversity order decreases in the following order: fungal metabolites > approved drugs > metabolites of

cyanobacteria > marine products > MEGx > TCM > NATx > NuBBE_{DB} > UNPD (see Table 3).

For a more comprehensive analysis of the scaffold diversity, a commonly used metric is Shannon entropy (SE).³³ Unlike the CSR curves that quantify the diversity of the entire data sets, SE has been employed to measure the scaffold diversity of the most populated scaffolds.⁵¹ The SSE value oscillates between 0, when all the compounds have the same chemotype (minimum diversity), and 1.0, when all the compounds are evenly distributed among the n acyclic and/or cyclic systems (maximum diversity).

Table S2 in the Supporting Information reports the number and fraction of compounds in the five, ten, and 20 most populated scaffolds in each database. Focusing the diversity analysis on the ten most frequent scaffolds Fungi NP and NATx show a high relative diversity as captured by the SSE10 metric (Table 2). Figure 5 shows the ten most frequent scaffolds in NuBBE_{DB}. The most frequent scaffolds for the reference databases are in Figure S3 in the Supporting Information.

Unique Scaffolds. In order to identify potential unique scaffolds in each data set, we searched for scaffolds that were not included in the entire ChEMBL database, version 24³⁴ (1,727,112 compounds) (see procedure in the Methods section). Marine products and MEGx were the NP databases with the relative larger number of unique scaffolds (20–26%, Table 2). 272 scaffolds in NuBBE_{DB} (12.28%) were not found in ChEMBL.

The unique scaffolds in NuBBE_{DB} were further searched in the database of the Dictionary of Natural Products³⁵ (274,478 compounds) to identify the ones that have not been reported within natural products. Out of the 272 scaffolds not found in ChEMBL, 36 scaffolds are not included in the current version of the Dictionary of Natural Products. The list of the 36 unique scaffolds is in Figure S4 of the Supporting Information. These scaffolds can be used as a starting point for the identification of new agents with therapeutic activity.

Structural Fingerprints. The structural diversity of NuBBE_{DB} was assessed using two structural fingerprints, namely Extended Connectivity fingerprints (ECFP4) and MACCS keys (166-bits), which are of different design. The diversity based on molecular fingerprints was useful to quantify the diversity of the compounds considering the entire structures (including not only the core scaffold but also side chains).

Figure 6 shows the cumulative distribution function of the pairwise intraset similarity values calculated with ECFP4/Tanimoto. The table beneath the figure summarizes the statistics of the distribution.

According to the median of the MACCS keys/Tanimoto the NPs with the greatest diversity were NuBBE_{DB} and the compounds were of marine origin. However, the largest diversity was of the approved drugs. Interestingly, these natural product libraries were the least diverse when studied based solely on the scaffolds, which means that acyclic systems and side chains are the contributing factors to the diversity of these data sets. The least test diverse collections considering molecular fingerprints were cyanobacteria and fungi metabolites. This can be explained by the fact that these databases have a high fraction of molecules contained in the most populated scaffolds; for both databases more than 30% of all compounds are contained in only 10 scaffolds (Table S2 in the Supporting Information).

Table 3. Summary of the Diversity Study^a

data set	code CDP	size	ECFP4	MACCS	N/M	AUC	F ₅₀	SSE10	molecular properties
approved drugs	1	1806	0.066	0.320	0.546	0.712	0.1389	0.961	2.22
cyanobacteria metabolites	2	473	0.096	0.500	0.376	0.750	0.123	0.753	2.96
fungi NP	3	206	0.087	0.440	0.563	0.672	0.233	0.964	2.18
marine NP	4	6253	0.079	0.426	0.399	0.754	0.117	0.898	2.49
MEGx	5	4103	0.085	0.470	0.398	0.765	0.095	0.894	2.73
NATx	6	26318	0.104	0.519	0.314	0.778	0.106	0.992	1.07
NuBBE _{DB}	7	2214	0.081	0.420	0.291	0.792	0.087	0.930	2.98
TCM	8	17986	0.078	0.462	0.957	0.778	0.079	0.878	2.18
UNPD	9	209574	0.076	0.446	0.939	0.857	0.025	0.812	2.25

^aM: number of molecules, N: number of scaffolds, AUC: area under the curve, F₅₀: fraction of chemotypes that contains 50% of the data set, SSE10: Scaled Shannon Entropy at the 10 most frequent scaffolds.

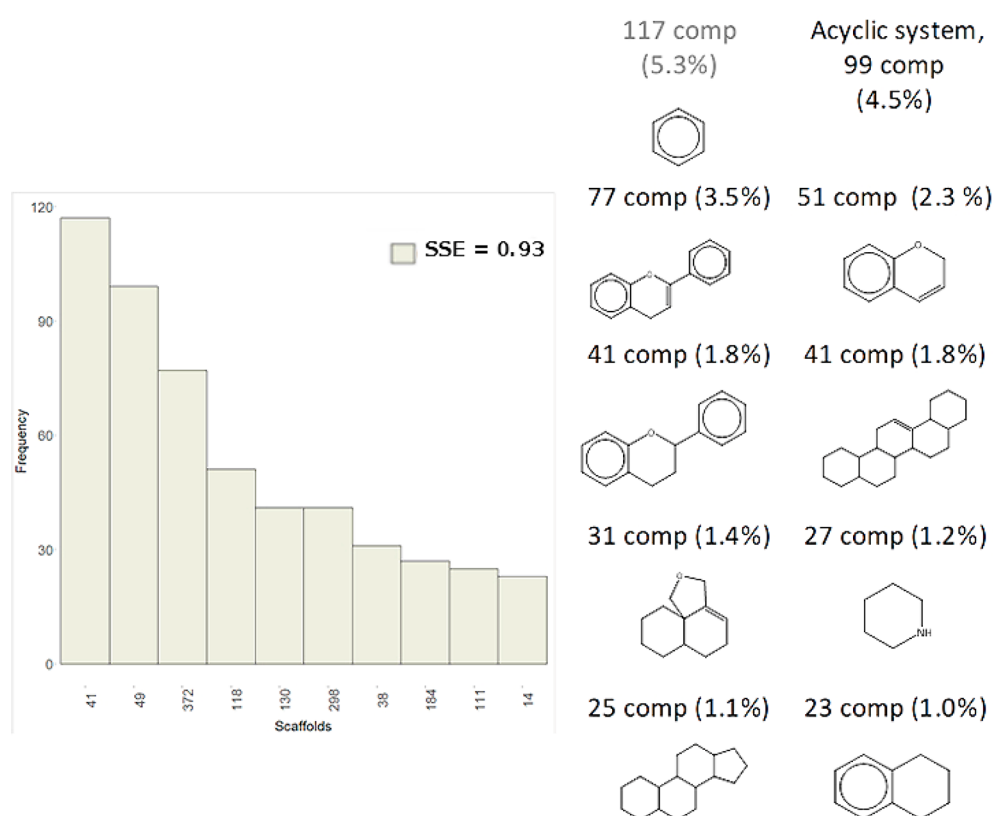


Figure 5. Distribution and Shannon entropy for the 10 most frequent chemotypes in NuBBE_{DB}.

Consensus Diversity Plot. In order to compare the diversity of NuBBE_{DB} with other compound collections considering multiple criteria simultaneously, we used consensus diversity plots (CDPs).²² A CDP comparing the overall (e.g., global) structural diversity of all data sets considering four criteria is presented in Figure 7. In this plot each point represents one data set. The median MACCS keys/Tanimoto of the data set is plotted on the X axis, and the AUC of the scaffold recovery curves is plotted on the Y axis. The size of the data points represents the relative size of each data set, and the color of each data point represents the diversity of molecular properties. The CDPlot indicates that the set of approved drugs is, overall, the most diverse, while the set of semisynthesis compounds (NATx) is the least diverse under all these four metrics.

Regarding the natural product databases, these are, in general, very diverse in terms of molecular properties but have different scaffold and fingerprint diversities. Marine NP as well

as approved drugs is diverse in terms of scaffolds and fingerprints. NuBBE_{DB} has low scaffold diversity but large fingerprint diversity. Among these two, NuBBE_{DB} stands out for having a greater diversity of molecular properties. Fungi (in the lower right quadrant) have a large scaffold diversity but low structural diversity based on fingerprints. MEGx, TCM, and UNPD have low scaffold diversity and low structural diversity but, unlike NATx, have a high diversity by molecular properties.

CONCLUSIONS

The herein reported chemoinformatic study of the most recent version of NuBBE_{DB} reveals that this database is characterized by having a focused chemical space, within the space of traditional and drug-like physicochemical properties. The results support quantitatively that NuBBE_{DB} is a promising source of molecules for drug discovery and medicinal chemistry. In NuBBE_{DB} the larger source of diversity of the

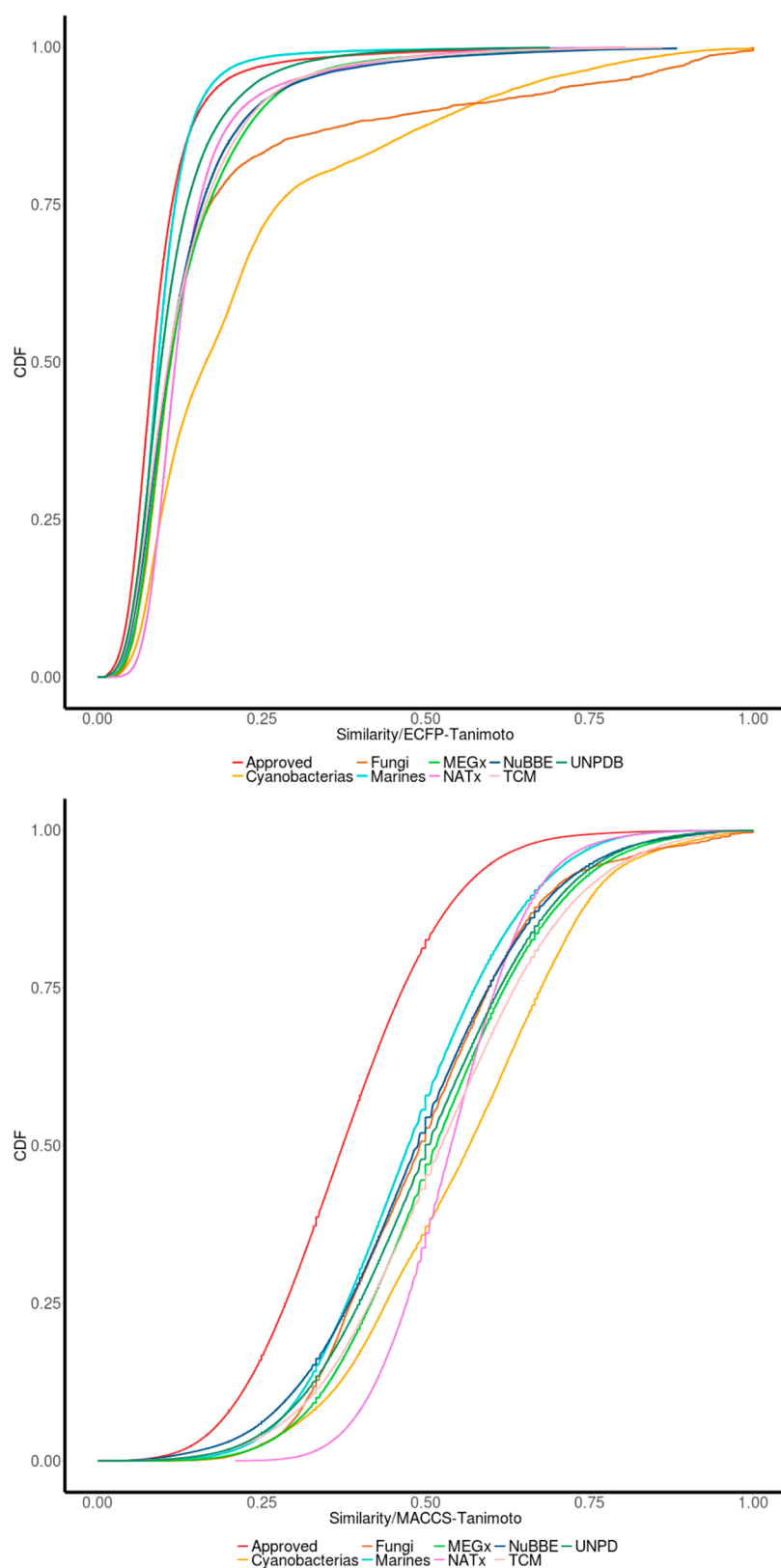


Figure 6. Cumulative distribution functions of all pairwise similarity comparisons using the ECFP4/Tanimoto coefficient and MACCS keys/Tanimoto. The table summarizes the statistics of the cumulative distribution functions.

compounds is driven by the side chains. Interestingly, a significant number of scaffolds in NuBBE_{DB} (12.28%, 272 compounds) are not present in ChEMBL. Furthermore, 36 of these scaffolds are not reported in the Dictionary of Natural

Products, which can serve as a starting point for the design of chemical libraries with novel scaffolds. When comparing NuBBE_{DB} with other collections of natural products, it was concluded that the diversity and complexity of the natural

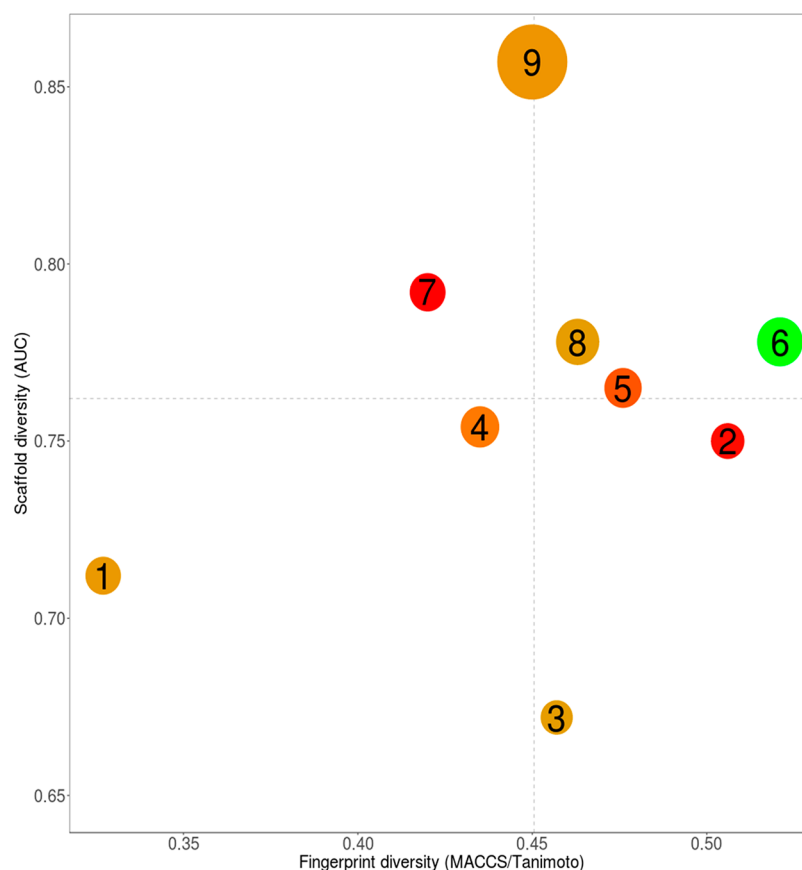


Figure 7. Consensus Diversity Plot comparing the diversity of NuBBE_{DB} with the reference data sets. Each data point represents a compound database: approved drugs (1), cyanobacteria metabolites (2), fungi NP (3), marine NP (4), MEGx (5), NATX (6), NuBBE_{DB} (7), TCM (8), and UNPD (9). The median MACCS/Tanimoto of the data set is plotted on the X axis, and the AUC of the scaffold recovery curves is plotted on the Y axis. 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.

products varies according to the origin of the compounds. The metabolites from fungi and cyanobacteria have the largest scaffold diversity, while NuBBE_{DB} and Marine NP are the most diverse considering structural fingerprints. During the course of this work it was also concluded that cyanobacteria metabolites are remarkable for their high structural complexity and distinct profile based on molecular properties and substructural alerts that are different from other natural products.

■ ASSOCIATED CONTENT

📄 Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: [10.1021/acs.jcim.8b00619](https://doi.org/10.1021/acs.jcim.8b00619).

Table S1, summary statistics; Table S2, number/fraction of compounds in the five, ten, and 20 most populated scaffolds in each database; Figure S1, box plots and summary statistics of the six physicochemical properties; Figure S2, NuBBE_{DB} subspace by geographical region; Figure S3, most frequent scaffolds for the reference databases; Figure S4, 36 unique scaffolds (PDF)

■ AUTHOR INFORMATION

Corresponding Author

*Phone: +5255-5622-3899, Ext. 44458. E-mail: medinajl@unam.mx, jose.medina.franco@gmail.com.

ORCID

Marilia Valli: [0000-0003-1106-183X](https://orcid.org/0000-0003-1106-183X)

José L. Medina-Franco: [0000-0003-4940-1107](https://orcid.org/0000-0003-4940-1107)

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was supported by the National Council of Science and Technology (CONACyT, Mexico) grant number 282785. F.I.S.-G. is thankful to CONACyT for the granted scholarship number 629458. The authors wish to acknowledge Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) grants #2013/07600-3 (CIBFar-CEPID), #2014/50926-0 (INCT BioNat CNPq/FAPESP), Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), and Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) for grant support and research fellowships. M.V. acknowledges scholarships #167874/2014-4 and #152243/2016-0 from CNPq and #120/2017 from Finatéc.

■ REFERENCES

(1) Newman, D. J. Natural Products as Leads to Potential Drugs: An Old Process or the New Hope for Drug Discovery? *J. Med. Chem.* **2008**, *51*, 2589–2599.

- (2) Kinghorn, A. D.; Pan, L.; Fletcher, J. N.; Chai, H. The Relevance of Higher Plants in Lead Compound Discovery Programs. *J. Nat. Prod.* **2011**, *74*, 1539–1555.
- (3) Koehn, F. E.; Carter, G. T. The Evolving Role of Natural Products in Drug Discovery. *Nat. Rev. Drug Discovery* **2005**, *4*, 206–220.
- (4) Newman, D. J.; Cragg, G. M. Natural Products as Sources of New Drugs from 1981 to 2014. *J. Nat. Prod.* **2016**, *79*, 629–661.
- (5) Bolzani, V. da S.; Valli, M.; Pivatto, M.; Viegas, C. Natural Products from Brazilian Biodiversity as a Source of New Models for Medicinal Chemistry. *Pure Appl. Chem.* **2012**, *84*, 1837–1846.
- (6) Valli, M.; Pivatto, M.; Danuello, A.; Castro-Gamboa, I.; Silva, D. H. S.; Cavalheiro, A. J.; Araújo, Â. R.; Furlan, M.; Lopes, M. N.; Bolzani, V. da S. Tropical Biodiversity: Has It Been a Potential Source of Secondary Metabolites Useful for Medicinal Chemistry? *Quim. Nova* **2012**, *35*, 2278–2287.
- (7) Lewinsohn, T. M.; Prado, P. I. How Many Species Are There in Brazil? *Conserv. Biol.* **2005**, *19*, 619–624.
- (8) Valli, M.; Russo, H. M.; Bolzani, V. S. The Potential Contribution of the Natural Products from Brazilian Biodiversity to Bioeconomy. *An. Acad. Bras. Cienc.* **2018**, *90*, 763–778.
- (9) Myers, N.; Mittermeier, R. A.; Mittermeier, C. G.; da Fonseca, G. A.; Kent, J. Biodiversity Hotspots for Conservation Priorities. *Nature* **2000**, *403*, 853–858.
- (10) Biodiversidade. <http://www.mma.gov.br/biodiversidade.html> (accessed Sep 1, 2018).
- (11) Barreiro, E. J.; Bolzani, V. da S. Biodiversidade: Fonte Potencial Para a Descoberta de Fármacos. *Quim. Nova* **2009**, *32*, 679–688.
- (12) Valli, M.; dos Santos, R. N.; Figueira, L. D.; Nakajima, C. H.; Castro-Gamboa, I.; Andricopulo, A. D.; Bolzani, V. S. Development of a Natural Products Database from the Biodiversity of Brazil. *J. Nat. Prod.* **2013**, *76*, 439–444.
- (13) Pilon, A. C.; Valli, M.; Dametto, A. C.; Pinto, M. E. F.; Freire, R. T.; Castro-Gamboa, I.; Andricopulo, A. D.; Bolzani, V. S. NuBBEDB: An Updated Database to Uncover Chemical and Biological Information from Brazilian Biodiversity. *Sci. Rep.* **2017**, *7*, 7215.
- (14) NuBBE - Núcleo de Bioensaios, Biossíntese e Ecofisiologia de Produtos Naturais. <http://nubbe.iq.unesp.br/portal/index.html> (accessed Sep 1, 2018).
- (15) Villoutreix, B. O.; Lagorce, D.; Labbé, C. M.; Sperandio, O.; Miteva, M. A. One Hundred Thousand Mouse Clicks down the Road: Selected Online Resources Supporting Drug Discovery Collected over a Decade. *Drug Discovery Today* **2013**, *18*, 1081–1089.
- (16) 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: 2015; pp 455–475, DOI: 10.1016/B978-0-12-800874-4.00021-0.
- (17) Harvey, A. L.; Edrada-Ebel, R.; Quinn, R. J. The Re-Emergence of Natural Products for Drug Discovery in the Genomics Era. *Nat. Rev. Drug Discovery* **2015**, *14*, 111–129.
- (18) Neves, B. J.; Andrade, C. H.; Cravo, P. V. L. Natural Products as Leads in Schistosome Drug Discovery. *Molecules* **2015**, *20*, 1872–1903.
- (19) Kuenemann, M. A.; Labbé, C. M.; Cerdan, A. H.; Sperandio, O. Imbalance in Chemical Space: How to Facilitate the Identification of Protein-Protein Interaction Inhibitors. *Sci. Rep.* **2016**, *6*, 23815.
- (20) Tietz, J. I.; Mitchell, D. A. Using Genomics for Natural Product Structure Elucidation. *Curr. Top. Med. Chem.* **2016**, *16*, 1645–1694.
- (21) Mohamed, A.; Nguyen, C. H.; Mamitsuka, H. Current Status and Prospects of Computational Resources for Natural Product Dereplication: A Review. *Briefings Bioinf.* **2016**, *17*, 309–321.
- (22) González-Medina, M.; Prieto-Martínez, F. D.; Owen, J. R.; Medina-Franco, J. L. Consensus Diversity Plots: A Global Diversity Analysis of Chemical Libraries. *J. Cheminf.* **2016**, *8*, 63.
- (23) KNIME - Open for Innovation. <https://www.knime.com/> (accessed Sep 1, 2018).
- (24) Gonzalez-Medina, M.; Medina-Franco, J. L. Platform for Unified Molecular Analysis: PUMA. *J. Chem. Inf. Model.* **2017**, *57*, 1735–1740.
- (25) DIFACQUIM Tools for Chemoinformatics. <https://www.difacquim.com/d-tools/> (accessed Oct 25, 2018).
- (26) *Molecular Modeling Environment (MOE)* version 2018; Chemical Computing Group: Montreal, Canada.
- (27) Osolodkin, D. I.; Radchenko, E. V.; Orlov, A. A.; Voronkov, A. E.; Palyulin, V. A.; Zefirov, N. S. Progress in Visual Representations of Chemical Space. *Expert Opin. Drug Discovery* **2015**, *10*, 959–973.
- (28) Opassi, G.; Gesù, A.; Massarotti, A. The Hitchhiker's Guide to the Chemical-Biological Galaxy. *Drug Discovery Today* **2018**, *23*, 565–574.
- (29) Baell, J. B.; Holloway, G. A. New Substructure Filters for Removal of Pan Assay Interference Compounds (PAINS) from Screening Libraries and for Their Exclusion in Bioassays. *J. Med. Chem.* **2010**, *53*, 2719–2740.
- (30) Saubern, S.; Guha, R.; Baell, J. B. KNIME Workflow to Assess PAINS Filters in SMARTS Format. Comparison of RDKit and Indigo Cheminformatics Libraries. *Mol. Inf.* **2011**, *30*, 847–850.
- (31) Xu, J. A. New Approach to Finding Natural Chemical Structure Classes. *J. Med. Chem.* **2002**, *45*, 5311–5320.
- (32) Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. I. Molecular Frameworks. *J. Med. Chem.* **1996**, *39*, 2887–2893.
- (33) Shannon, C. E.; Weaver, W. *The Mathematical Theory of Communication*; University of Illinois Press: 1998.
- (34) Gaulton, A.; Hersey, A.; Nowotka, M.; Bento, A. P.; Chambers, J.; Mendez, D.; Motow, P.; Atkinson, F.; Bellis, L. J.; Cibrián-Uhalte, E.; et al. The ChEMBL Database in 2017. *Nucleic Acids Res.* **2017**, *45*, D945–D954.
- (35) Dictionary of Natural Products 27.1. <http://dnp.chemnetbase.com/faces/chemical/ChemicalSearch.xhtml> (accessed Sep 1, 2018).
- (36) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983–996.
- (37) Olmedo, D. A.; González-Medina, M.; Gupta, M. P.; Medina-Franco, J. L. Cheminformatic Characterization of Natural Products from Panama. *Mol. Diversity* **2017**, *21*, 779–789.
- (38) González-Medina, M.; Prieto-Martínez, F. D.; Naveja, J. J.; Méndez-Lucio, O.; El-Elmat, T.; Pearce, C. J.; Oberlies, N. H.; Figueroa, M.; Medina-Franco, J. L. Chemoinformatic Expedition of the Chemical Space of Fungal Products. *Future Med. Chem.* **2016**, *8*, 1399–1412.
- (39) Ouguéné, P. A.; Simoben, C. V.; Fotso, G. W.; Andrae-Marobela, K.; Khalid, S. A.; Ngadjui, B. T.; Mbaze, L. M.; Ntie-Kang, F. In silico toxicity profiling of natural product compound libraries from African flora with anti-malarial and anti-HIV properties. *Comput. Biol. Chem.* **2018**, *72*, 136–149.
- (40) Pye, C. R.; Bertin, M. J.; Lokey, R. S.; Gerwick, W. H.; Linington, R. G. Retrospective Analysis of Natural Products Provides Insights for Future Discovery Trends. *Proc. Natl. Acad. Sci. U. S. A.* **2017**, *114*, 5601–5606.
- (41) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and Computational Approaches to Estimate Solubility and Permeability in Drug Discovery and Development Settings. *Adv. Drug Delivery Rev.* **1997**, *23*, 3–25.
- (42) Veber, D. F.; Johnson, S. R.; Cheng, H.-Y.; Smith, B. R.; Ward, K. W.; Kopple, K. D. Molecular Properties That Influence the Oral Bioavailability of Drug Candidates. *J. Med. Chem.* **2002**, *45*, 2615–2623.
- (43) 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.
- (44) Clemons, P. A.; Bodycombe, N. E.; Carrinski, H. A.; Wilson, J. A.; Shamji, A. F.; Wagner, B. K.; Koehler, A. N.; Schreiber, S. L. Small Molecules of Different Origins Have Distinct Distributions of Structural Complexity That Correlate with Protein-Binding Profiles. *Proc. Natl. Acad. Sci. U. S. A.* **2010**, *107*, 18787–18792.

(45) Lovering, F.; Bikker, J.; Humblet, C. Escape from Flatland: Increasing Saturation as an Approach to Improving Clinical Success. *J. Med. Chem.* **2009**, *52*, 6752–6756.

(46) Nilar, S. H.; Ma, N. L.; Keller, T. H. The Importance of Molecular Complexity in the Design of Screening Libraries. *J. Comput.-Aided Mol. Des.* **2013**, *27*, 783–792.

(47) Méndez-Lucio, O.; Medina-Franco, J. L. The Many Roles of Molecular Complexity in Drug Discovery. *Drug Discovery Today* **2017**, *22*, 120–126.

(48) Barone, R.; Chanon, M. A New and Simple Approach to Chemical Complexity. Application to the Synthesis of Natural Products. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 269–272.

(49) Gally, J. M.; Bourg, S.; Do, Q. T.; Aci-Sèche, S.; Bonnet, P. VSPrep: A General KNIME Workflow for the Preparation of Molecules for Virtual Screening. *Mol. Inf.* **2017**, *36* (10), 1700023.

(50) Baell, J. B. Feeling Nature's PAINS: Natural Products, Natural Product Drugs, and Pan Assay Interference Compounds (PAINS). *J. Nat. Prod.* **2016**, *79*, 616–628.



(51) Medina-Franco, J.; 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.

(52) Chen, CY-C TCM Database@Taiwan: The World's Largest Traditional Chinese Medicine Database for Drug Screening in Silico. *PLoS One* **2011**, *6*, No. e15939.

(53) 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*, No. e62839.

Article

BIOFACQUIM: A Mexican Compound Database of Natural Products

B. Angélica Pilón-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).

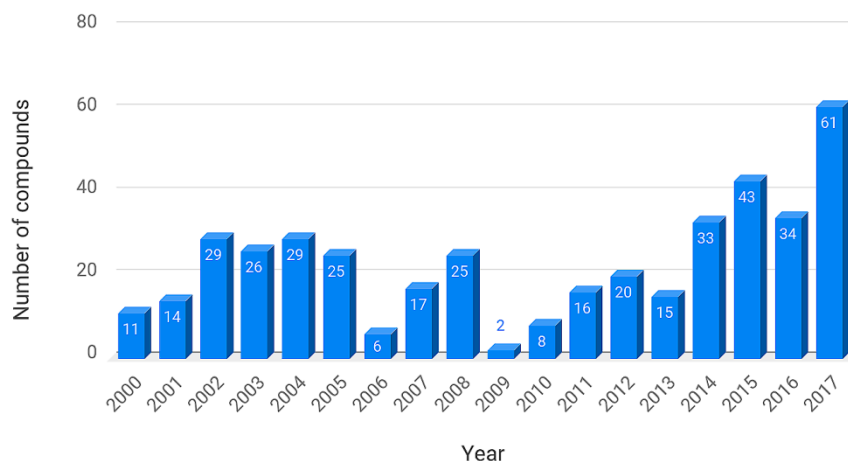


Figure 1. Distribution of compounds reported from 2000 to 2017, as contained in the first version of BIOFACQUIM. Compounds published in 2018 are not shown in this graph.

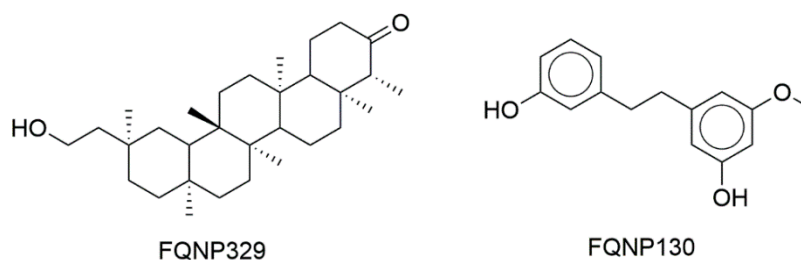


Figure 2. Select compounds contained in BIOFACQUIM.

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 NuBBE_{DB} have the most similar MW profile compared to drugs. BIOFACQUIM has a median of 4 HBA, the same as that of the NuBBE_{DB} and Marine data sets. Furthermore, BIOFACQUIM has a very similar profile of HBA compared to MEGx. Comparing HBD, BIOFACQUIM, NuBBE_{DB}, 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.

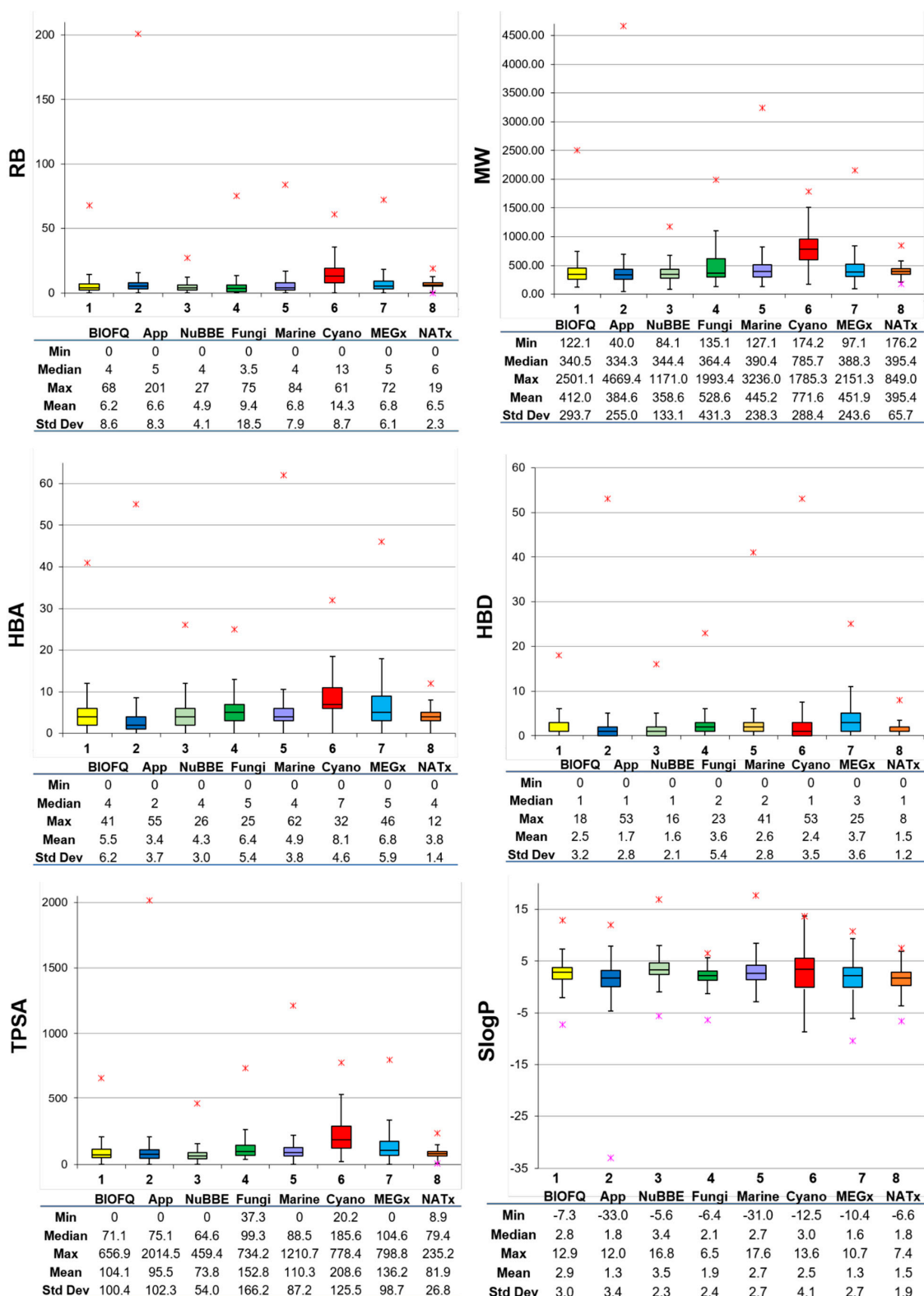


Figure 3. Box plots for the physicochemical properties of BIOFACQUIM (BIOFQ) and reference data sets (Table 1). The boxes enclose data points with values within the first and third quartile. The red asterisks indicate outliers. Summary statistics are included below each plot. RB: number of rotatable 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].

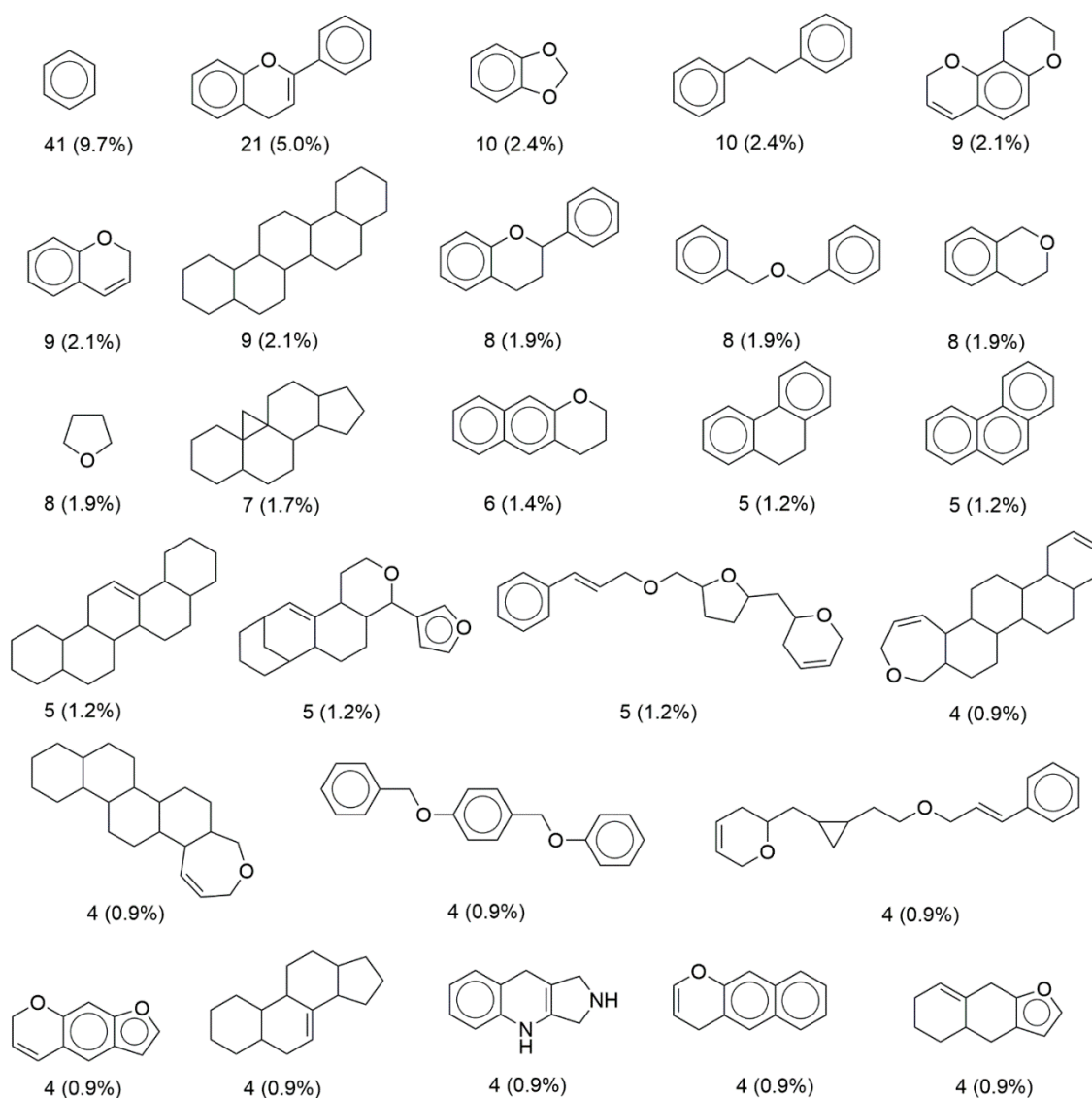


Figure 4. Most frequent scaffolds in BIOFACQUIM. The frequency and percentage are shown. The 27 scaffolds shown in the figure contain half of the total compounds in the database (50.6%).

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.

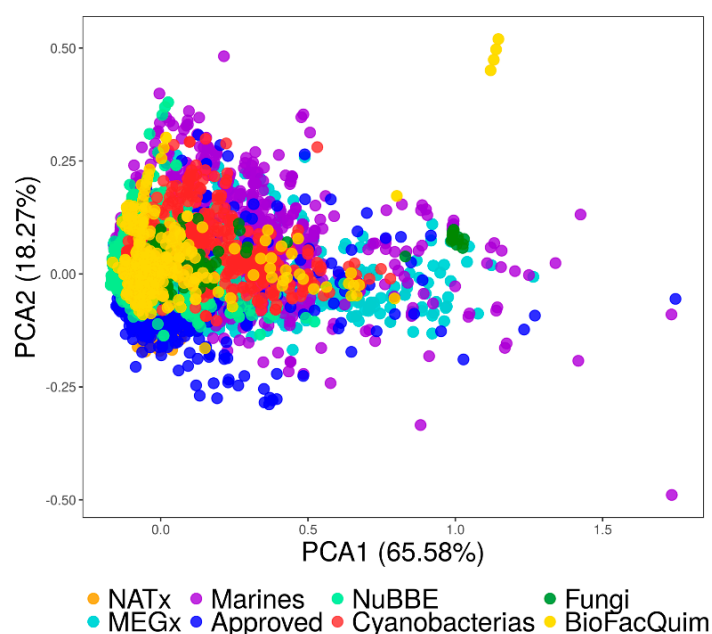


Figure 5. Visual representation of the chemical space based on the physicochemical properties of eight data sets. BIOFACQUIM (423 compounds, yellow); fungi metabolites (206 compounds, green); cyanobacteria metabolites (473 compounds, red); NuBBE_{DB} (2214 compounds, light green); NATx (26318 compounds, 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.

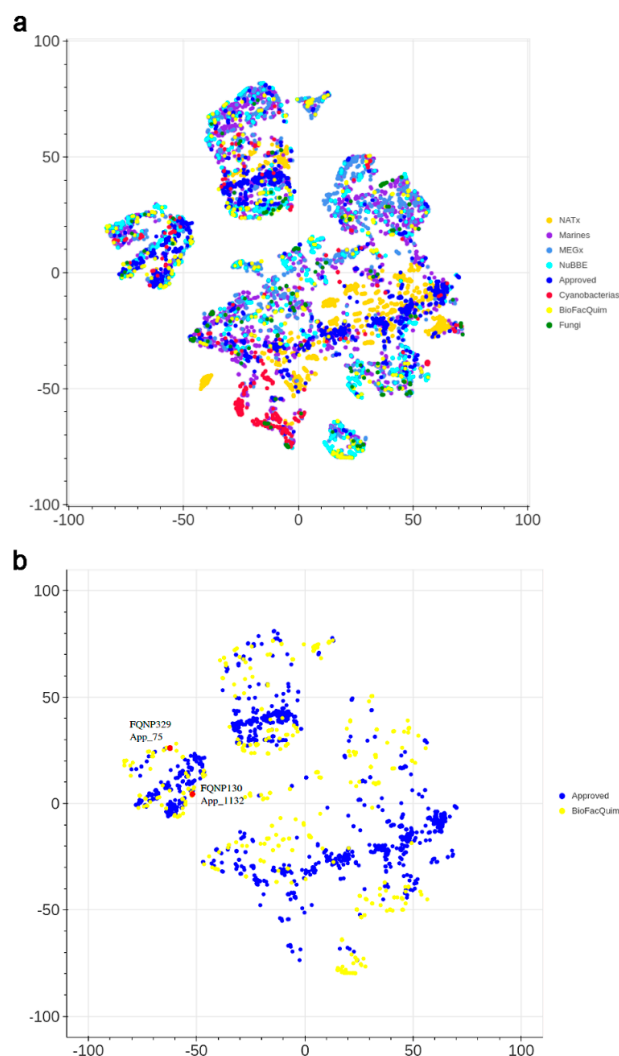


Figure 6. Visual representation of the chemical space of BIOFACQUIM compared with: (a) all reference data sets; and (b) approved drugs. The visualization was generated using *t*-distributed stochastic neighbor embedding (*t*-SNE) based on topological fingerprints. The red dots indicate the position of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

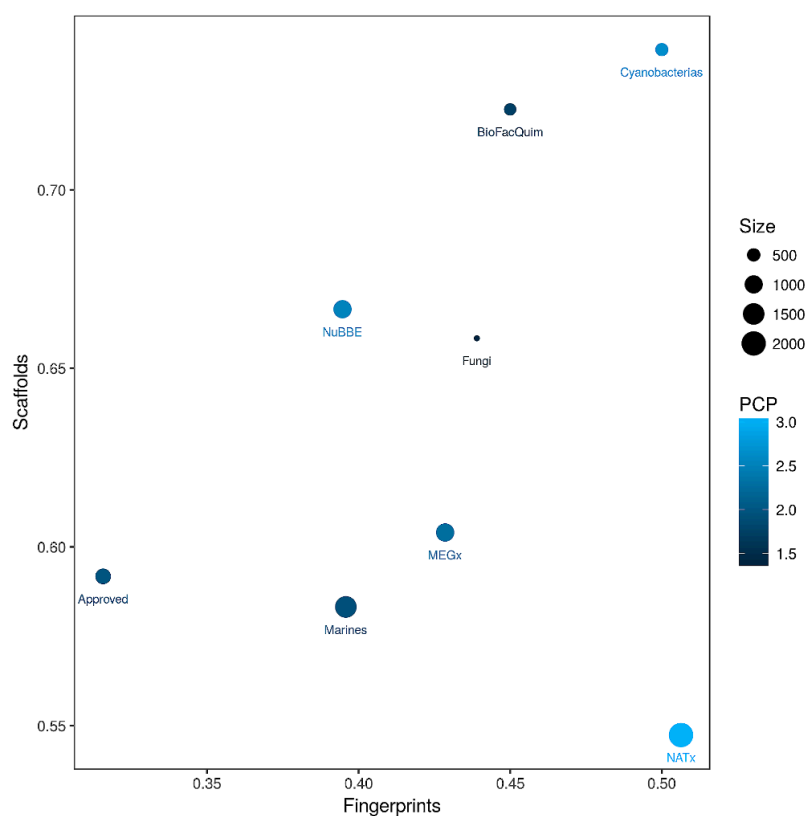


Figure 7. Consensus Diversity Plot comparing the global diversity of BIOFACQUIM with other natural product databases. The structural diversity (fingerprint diversity), calculated with the median Tanimoto coefficient of MACCS keys fingerprints, is plotted on the x -axis. The scaffold diversity of each 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](#)]





Cite this: *Org. Biomol. Chem.*, 2019, **17**, 1037

Bicyclic acetals: biological relevance, scaffold analysis, and applications in diversity-oriented synthesis†

Elena Lenci,  *^a Gloria Menchi, ^{a,b} Fernanda I. Saldívar-Gonzalez,^c José L. Medina-Franco ^c and Andrea Trabocchi  *^{a,b}

Natural products (NPs) have been shown to be an extraordinary source of bioactive compounds and three-dimensional complex frameworks that can be useful to produce high-value molecular collections that are able to address “undruggable” and difficult therapeutic targets. Bicyclic acetals are particularly relevant for these purposes, given their key role in several biological interactions and the structural and stereochemical diversity that comes from the many possible ring combinations. To investigate this topological diversity, we have systematically characterized in a systematic and detailed manner fused, spiro and bridged bicyclic acetals in a large set of NPs, highlighting the great potential of bicyclic acetals in Diversity-Oriented Synthesis (DOS). Accordingly, a summary of some recent efforts on the development of acetal-containing small molecule collections through DOS approaches is herein reported.

Received 10th November 2018,
Accepted 19th December 2018

DOI: 10.1039/c8ob02808g

rsc.li/obc

^aDepartment of Chemistry “Ugo Schiff”, University of Florence, Via della Lastruccia 13, 50019 Sesto Fiorentino, Florence, Italy. E-mail: elena.lenci@unifi.it, andrea.trabocchi@unifi.it; Fax: (+39) 055 4574913

^bInterdepartmental Center for Preclinical Development of Molecular Imaging (CISPIM), University of Florence, Viale Morgagni 85, 50134 Florence, Italy

^cSchool of Chemistry, Department of Pharmacy, Universidad Nacional Autónoma de México, Avenida Universidad 3000, Mexico City 04510, Mexico

† Electronic supplementary information (ESI) available: SMARTS code of all the 68 possible fused, spiro and bridged bicyclic acetals containing rings composed of 3 to 8 atoms. See DOI: 10.1039/c8ob02808g

1. Introduction

Natural products are an extraordinary source of inspiration for the development of high-quality chemical libraries.¹ Many of these compounds possess complex, sp³-rich molecular frameworks, which have been identified as ‘privileged’,² as regularly occurring motifs in many bioactive protein ligands, and have been comprehensively mapped to facilitate the navigation of



Elena Lenci

Dr Elena Lenci is a postdoctoral researcher at the University of Florence, working on the diversity-oriented synthesis of novel molecular scaffolds and their applications to medicinal chemistry projects. She received the PhD title in Chemical Sciences in 2017 under the supervision of Prof. A. Trabocchi, after spending a research period in the group of Prof. Dixon at the University of Oxford, UK. In 2015, she was awarded the

Reaxys – Italian Chemical Society Young Researcher Award and in 2017, she received the International Award for Young Chemists “NATCHEM DRUGS”.



Gloria Menchi

Gloria Menchi graduated in Chemistry in 1977 with full marks at the University of Florence. She is an Assistant Professor of Organic Chemistry at the Department of Chemistry ‘U. Schiff’ of the University of Florence. Her scientific activity regards stereoselective organic synthesis of heterocyclic compounds and their applications in Medicinal Chemistry and Biotechnology. Her recent research studies concern the

development of new molecular scaffolds and their applications for the discovery of new drugs, the synthesis of peptidomimetic libraries and modified peptides, and the labeling of biologically active compounds with stable and radioactive isotopes. She is the author of 104 publications and 4 patent applications.

the chemical space.³ Exploiting unconventional molecular scaffolds to develop chemical libraries can increase the chance of finding compounds that are able to address the so-called “undruggable” targets, such as protein–protein interactions.⁴ Natural products are characterized by a notable structural and stereochemical complexity that increases the binding selectivity and frequency towards biological targets.⁵ In this context, Diversity-Oriented Synthesis (DOS)⁶ and Biology-Oriented Synthesis (BIOS)⁷ represent two leading approaches for the synthesis of natural product-inspired compound collections. Specifically, DOS operates in a way to generate the maximum diversity and complexity from simple starting materials using divergent synthetic strategies, such as the use of complexity-generating reactions and the build/couple/pair approach.⁶

Among all the different molecular frameworks present in Nature, bicyclic acetals are particularly relevant for medicinal chemistry purposes, given their key role in several biological interactions and the chemical diversity that comes from the many possible ring combinations.^{8–10} Structurally, a bicyclic acetal moiety consists of at least two rings where two oxygen atoms belonging to different rings are linked through a common carbon atom. These moieties are widespread in a variety of natural products from different sources, spanning from the insect world, marine organisms, fungi, plants and microbes.^{8–10} In several natural products, such chemotypes are found as spiroacetals (Fig. 1a),⁸ fused bicyclic acetals (Fig. 1b),⁹ or bridged bicyclic acetals (Fig. 1c);¹⁰ also in some natural products, a combination of them is found (Fig. 1d).

This review provides a perspective of the utility of bicyclic acetal chemotypes for the development of small molecule chemical libraries. Specifically, we herein present (a) insights into the biological potency of compounds containing bicyclic acetal chemotypes, (b) a systematic and quantitative evaluation of the diversity and complexity of bicyclic acetals in Nature, also concerning their distribution regarding products of

different origins and (c) selected case studies on the diversity-oriented synthesis of acetal-containing compound collections and their biological output.

2. Biological relevance of bicyclic acetals

Bicyclic acetals, both when they are a substructure in a highly functionalized system or when they are relatively unsubstituted, play a key role in the biological outcome of natural products, by directly interacting with the biological target as a pharmacophore, or acting as a rigid scaffold to direct side chains into specific directions. For example, the spiroacetal moiety in ivermectin B_{1a} is the key element responsible for the binding with glutamate-gated chloride channels (GluCl) through a hydrogen bonding interaction with Thr285, which results in anthelmintic and insecticidal properties of the entire molecule (Fig. 2, left).¹¹ The spiroacetal moiety of bistramide A (Fig. 2, center) has a scaffolding function acting as a saddle-like turn element, which directs the long hydrophobic chains into a deep cleft between subdomains 1 and 3 of G-actin, resulting in its depolymerization and subsequent anti-proliferative action.¹²

Recently, Milroy and coworkers have reported the first example of a complex formed by a novel bis-benzannulated spiroacetal and the retinoid X receptor (RXR), a potential therapeutic target for Alzheimer's disease (Fig. 2, right).¹³

The conformation of bicyclic acetals plays an important role in assisting the interaction with the biomacromolecule. Detailed analytical studies have been conducted in complex polyketide natural products, such as bryostatins and monensin,¹⁴ concluding that steric, electronic and electrostatic interactions,¹⁵ as well as internal hydrogen bonds between ring-acetal oxygen and hydroxyl groups,¹⁶ play an important role in



Fernanda I. Saldívar-Gonzalez

Fernanda Saldívar received her BSc degree in Chemistry Pharmacy and Biology (2017) from the National Autonomous University of Mexico (UNAM). Since 2016, she has been working in the DIFACQUIM research group. She is currently studying a Master's degree in Chemistry in the area of pharmacy where she develops her project focused on the study of natural products as a resource for virtual screening and identification of bioactive compounds.



José L. Medina-Franco

José L. Medina-Franco received his Ph.D. degree from the National Autonomous University of Mexico (UNAM). He was a postdoctoral fellow at the University of Arizona and joined Torrey Pines Institute for Molecular Studies in Florida in 2007. In 2013, he moved to the Mayo Clinic and later joined UNAM as a Full Time Research Professor. He currently leads the DIFACQUIM research group. In 2017, he was named Fellow of the Royal Society of Chemistry. His research interests include the development and application of chemoinformatics and molecular modeling methods for bioactive compounds with emphasis on drug discovery.

adopting the three-dimensional conformation of the molecules necessary for the binding, while retaining the backbone flexibility beneficial for transport and solubility. These remarkable features of bicyclic acetals can also explain why natural products possessing these moieties have a broad variety of biological outcomes.^{8e} For example, nine different avermectins coming from the same biosynthetic intermediate show different biological activities depending on their different substituents and stereochemical features.¹⁷

Even though the use of bicyclic acetals in medicinal chemistry and drug discovery is still underexplored, acetal containing-drugs and synthetic inhibitors have appeared in the literature (Fig. 3).^{18–21} For example, a novel Neurokinin 1 receptor (NK₁) antagonist (compound 1) containing a rigid 1,6-dioxo-9-azaspiro[4.5]decane framework was developed by Williams and coworkers.¹⁸ The structure of the antibiotic spectinomycin (2), used for the treatment of gonorrhoea, is characterized by the presence of a [6,6]-fused bicyclic acetal.¹⁹

Tofogliflozin (3), designed using a 3D-pharmacophore modelling²⁰ and based on the *O*-spiroacetal *C*-arylglucoside scaffold, is an active Sodium Glucose Cotransporter 2 (SGLT2) inhibitor and a potential therapeutic agent for the treatment of type 2 diabetes. Another example is the novel potent cytotoxic compound 4 active in primary B-cell chronic lymphocytic leukaemia cells that contains the 6,8-dioxabicyclo[3.2.1]octane scaffold.²¹

3. Scaffold analysis and classification

3.1 Scaffold content

Scaffold content analysis is gaining broad attention both in chemical biology and drug design, because it helps to quantify the structural diversity of compound collections and have a better understanding of the coverage of such a collection in the chemical space.²² Likewise, it is broadly used to identify

novel scaffolds in a compound library, and to analyse the structure–activity relationships of sets of molecules with the measured activity.²³ Different methods have been developed for a systematic and consistent scaffold analysis,²⁴ such as the graph framework methodology,²⁵ the atomic frameworks of Bemis and Murcko,²⁶ and Scaffold Hunter.²⁷

To investigate the topological diversity of bicyclic acetal frameworks present in NPs, a large set of 466 238 NPs was analysed, combining different databases available in the public domain including the Universal Natural Product Database (UNPD) (Table 1).^{28–33} The datasets used in this work differ in size, origin and compound features and allow us to survey in a quantitative manner the presence of different bicyclic acetal-containing molecular frameworks on NPs from different sources, spanning from fungi metabolites to marine compounds.

All the possible frameworks of bicyclic acetals, containing both rings composed of 3–8 atoms, were defined and classified into three main categories: fused (Fig. 4a), spiro (Fig. 4b), and bridged bicyclic acetals (Fig. 4c). The full list of SMARTS codes can be found in the ESI.† An example of the extrapolation of the [5,6]-spiroacetal ring combination from the structure of allamandicin is reported in Fig. 4d.

In total, of the 466 328 NPs analysed, 4699 (1%) containing at least one bicyclic acetal in their structure were identified. These compounds were found to contain 45 different molecular frameworks, distributed among the three categories of fused, spiro and bridged bicyclic acetals, with a significant prevalence of spiroacetals.

1578 fused bicyclic acetals were identified in the data set of NPs surveyed in this work, with a different frequency distribution for the 21 possible ring combinations (Fig. 4a). As expected, given the dominance of five- and six-membered heterocycles in Nature, the three most frequent ring combinations were found to be the [5,5]-fused bicyclic acetal, which is present for example in the cytotoxic macrocalyxoforin A³⁴ and mycotoxin aflatoxins,³⁵ the [6,5]-fused bicyclic acetal, contained in the alkaloid core of the insecticidal compounds of the stemofoline family,³⁶ and the [6,6]-fused bicyclic acetal, as this framework is embedded in the structure of colubrin³⁷ and erinacine B, a xylose-conjugated terpenoid, is able to stimulate the synthesis of the nerve growth factor (NGF).³⁸

Spiroacetals are by far the most frequent *O,O*-containing scaffolds, as 2633 spiroacetal moieties were counted in the NP database (Fig. 4b). The ring combinations with a consistently higher frequency were [6,5]-, [6,6]- and [5,5]-spiroacetals. These moieties are in fact widespread in natural products, including compounds of the families of marine macrolide spongistatins,³⁹ antitumoral steroidal cephalostatins⁴⁰ and insecticidal avermectins.⁴¹ On the other hand, bridged bicyclic acetals with three to eight atoms in the ring can exist in 26 topologically different combinations that differ in terms of bridge length, ring size, and distribution of oxygen atoms (Fig. 4c). The analysis of the NPs uncovered 2159 bridged bicyclic acetals, which can be clustered in 17 different bridged bicyclic acetal molecular frameworks. The most frequent molecular



Andrea Trabocchi

Andrea Trabocchi is an Associate Professor of Organic Chemistry at the University of Florence, Italy, where his research is focused in the area of diversity-oriented synthesis and peptidomimetic chemistry. He is also involved in PET and MRI molecular imaging research, and medicinal chemistry concerning enzyme inhibitors in the fields of oncology and Alzheimer's disease. He obtained a PhD in Chemical Sciences in 2003 from

the University of Florence, and received training on peptide chemistry at Imperial College, UK from Prof. Leatherbarrow. He recently edited a Wiley book on diversity-oriented synthesis and authored a Wiley book on Peptidomimetics.

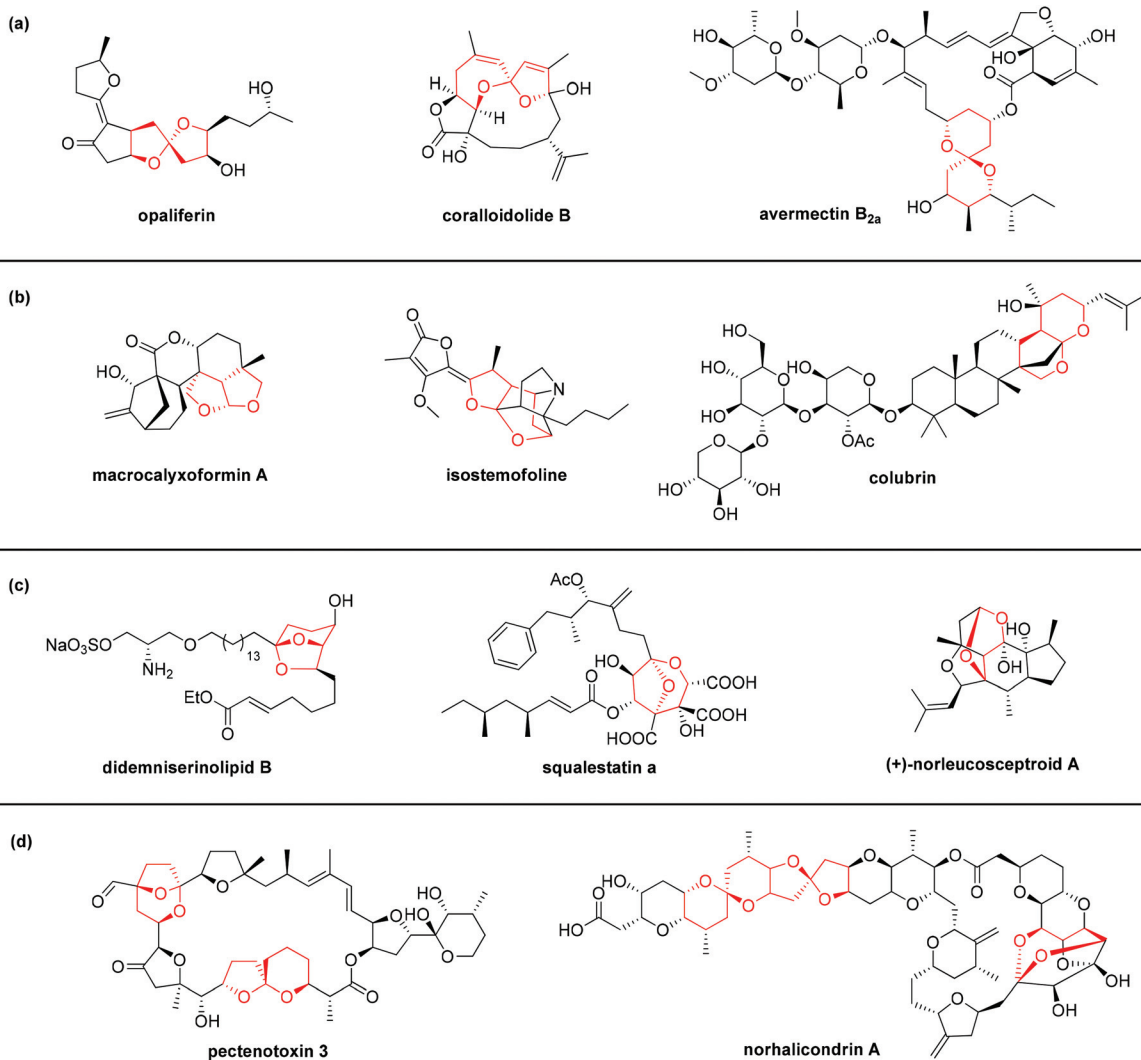


Fig. 1 Representative example of natural products containing a spiroacetal (a), a fused bicyclic acetal (b), a bridged bicyclic acetal (c), or a combination of them (d).

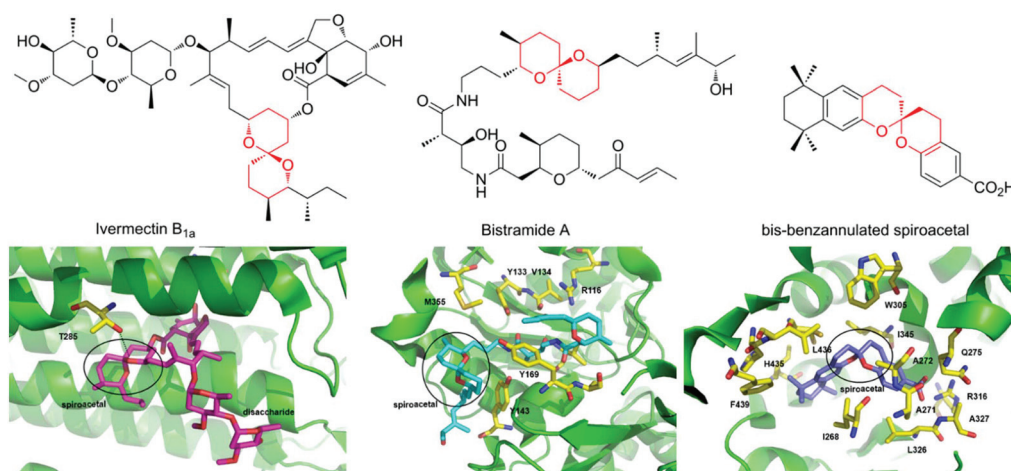


Fig. 2 Left: View from the extracellular site of ivermectin B_{1a} (fuchsia) into the binding site of the GluCl subunit (PDB ID: 3RHW).¹¹ Center: Selected amino acid side chains of actin subdomains 1 and 3 interacting with bistramide A (cyan, PDB ID: 2FXU).¹² Right: Selected amino acids of the RXR protein interacting with the bis-benzannulated spiroacetal (blue) synthesized by Milroy *et al.* (PDB ID: 5LYQ).¹³

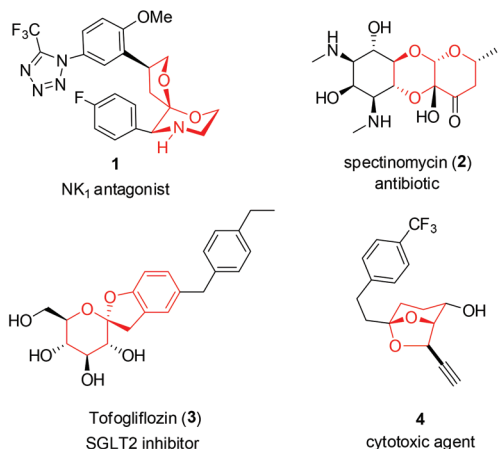


Fig. 3 Representative examples of acetal containing inhibitors and drugs.

Table 1 Databases of natural products considered in this work

Database	Size	Ref.
Fungi NP	206	29
Marine NP	6253	30
Purified Natural Product Screening Compounds (MEGx)	4103	http://ac-discovery.com
Brazilian NP (NuBBE _{DB})	2214	31
Traditional Chinese medicine (TCM)	18281	http://tcm.cmu.edu.tw 32
ZINC natural products (NP)	225667	http://tcm.cmu.edu.tw 32
Universal Natural Product Database (UNPD)	209514	33

framework was found to be the 6,8-dioxabicyclo[3.2.1]octane,⁴² contained for example in the structure of rubrobramide⁴³ and didemniserinolipid B,⁴⁴ as well as in the neurotoxic compounds of the families of pinnatoxins⁴⁵ and pterioxins.⁴⁶ The analogues 2,8-dioxabicyclo[3.2.1]octane and 2,7-dioxabicyclo[3.2.1]octane are less represented, but they can be found in the structure of zaragozic acid,⁴⁷ pectenotoxins,⁴⁸ sinduro,⁴⁹ and antimicrobial colomitides.⁵⁰

To further investigate the distribution of bicyclic acetals in NPs of different origins, for the seven data sets analysed in this work, we recorded the number of natural products containing each molecular scaffold. The percentage of these counts *versus* the total number of natural products identified (6369) is reported in Fig. 5, where each dataset is shown in a different colour.

This analysis revealed that datasets of Universal Natural Products Database (UNPD), Traditional Chinese Medicine (TCM) and the compounds from marine sources had the largest proportion of bicyclic acetal frameworks, with a prevalence of [6,5]-spiroacetals in the UNPD and TCM databases and [6,6]-spiroacetals in the marine database. Also, in the category of bridged bicyclic acetals, the presence of the 6,8-dioxabicyclo[3.2.1]octane framework stands out in the ZINC NP

database, whereas the MEGx database, which is a collection of natural products derived from plants and microorganisms, was found to be abundant of fused bicyclic acetal scaffolds.

3.2 Scaffold diversity

The scaffold diversity of the 4699 acetal-containing natural products was then quantified following Bemis and Murcko's definition of scaffolds (see an example in Fig. 4d)²⁶ and using different established metrics. The scaffold count analysis reveals that among the bicyclic acetals, fused bicyclic acetals are the most diverse. These compounds contain the highest proportion of singletons (unique scaffolds) relative to the number of cyclic systems (N_{sing}/N). In contrast, bridged bicyclic acetals present a high redundancy of structures (a low number of singletons in relation to the number of different cyclic systems (N_{sing}/N)).

The fraction of compounds was plotted against the cumulative fraction of scaffolds for the three main categories of bicyclic acetals to obtain a direct comparison between the scaffold content and the diversity of acetal-containing natural products.

The Cyclic System Retrieval (CSR) graph⁵¹ thus obtained (Fig. 6) shows that fused bicyclic acetals (orange line), being the closest to a diagonal, have the largest diversity of the three data sets, while bridged bicyclic acetals (blue line) and spiroacetals compounds (green line) are less diverse. A similar conclusion can be obtained quantitatively by comparing the area under the curve (AUC) and the fraction of chemotypes that recover 50% of the molecules in the data set (F_{50}), which shows a diversity decrease by moving from fused bicyclic acetals, to spiroacetals and bridged bicyclic acetals (Table 2).

To measure the scaffold diversity of the three groups of compounds focusing only on the most populated scaffolds,⁵² we used the concept of Shannon Entropy (SE). When applied to quantify the scaffold diversity,⁵³ SE is defined as:

$$SE = - \sum_{i=1}^n p_i \log_2 p_i \text{ where } p_i = \frac{c_i}{P}$$

where p_i is the estimated probability of the occurrence of a specific chemotype i in a population of P compounds containing a total of n acyclic and cyclic systems, and c_i is the number of molecules that contain a particular chemotype. SE was normalized for the number of cyclic systems of each class (n):

$$SSE = \frac{SE}{\log_2 n}$$

The scaled SE (SSE) value range between 0 (minimum diversity, *i.e.*, all the compounds have the same chemotype) and 1 (maximum diversity, *i.e.*, compounds are evenly distributed among the n chemotypes). The SSE was used to measure the scaffold diversity of the 10 most populated scaffolds (SSE10) of each category of bicyclic acetals. The results, summarized in Table 2, show that considering the 10 most populated scaffolds, fused bicyclic acetal compounds are the most diverse and the bridged bicyclic acetals are the least diverse.

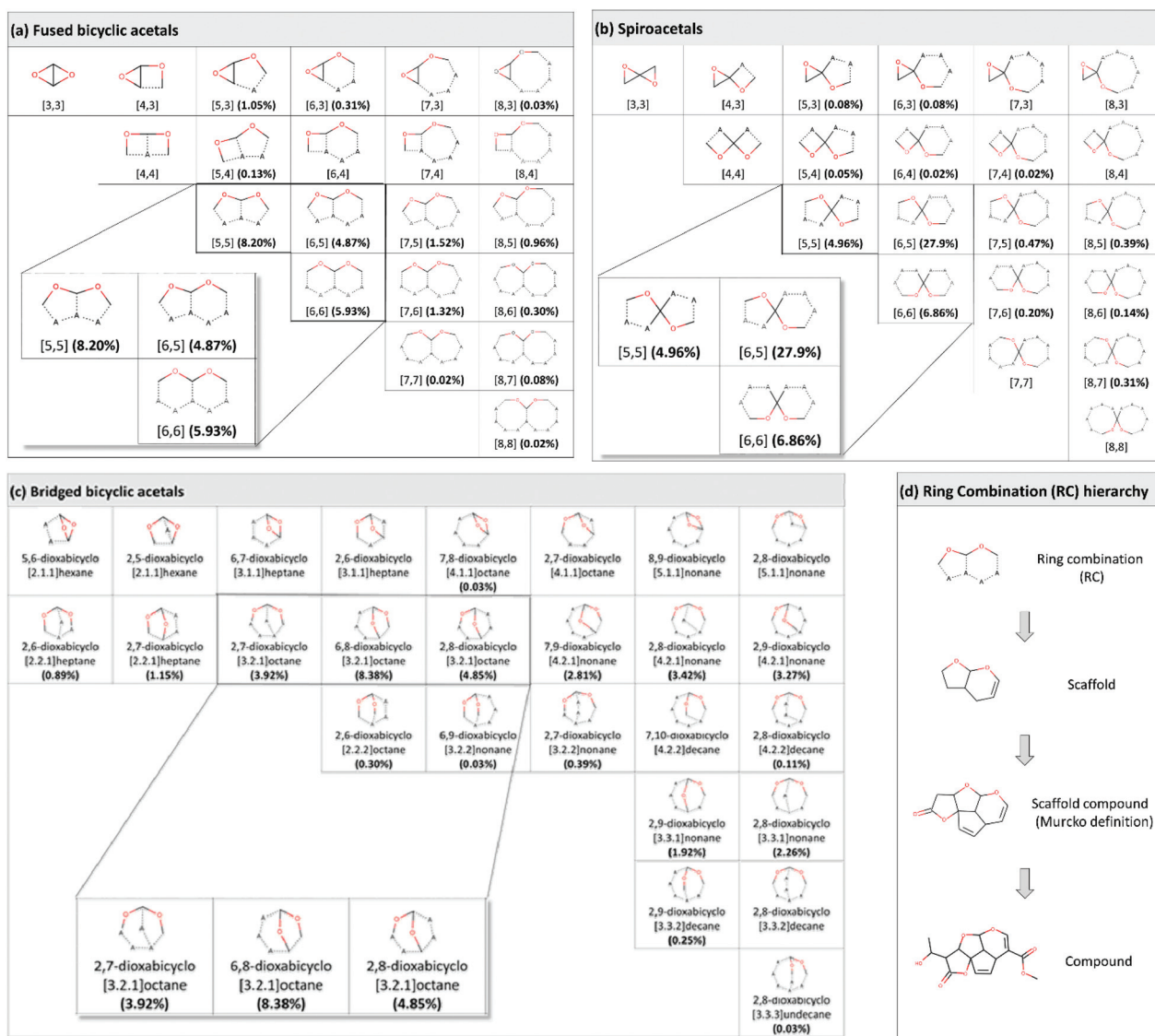


Fig. 4 Classification and percent frequency of fused bicyclic acetals (a), spiroacetals (b) and bridged bicyclic acetals (c) in a dataset of 466 238 NPs available in the public domain. The three most frequent ring combinations for each class of bicyclic acetals are shown in zoomed boxes. The percentage shown is relative to the total number of bicyclic acetal frameworks identified in the database (6369) and not to the total number of acetal-containing natural products (4699), as some NPs contain more than one framework in their structure. (d) Example of the identification of the [5,6]-spiroacetal ring combination in the structure of allamandicin.

These results mirror the relative diversity of the data sets considering all the scaffolds (Table 2).

3.3 Structural complexity

To quantify the three-dimensional complexity of the acetal-containing scaffolds, we calculated the Saturation Index (F_{sp^3}) as the ratio between the number of sp^3 hybridized carbons versus the total carbon count.⁵⁴ As shown in Fig. 7, spiroacetals proved to be the class of compounds with the highest F_{sp^3} ratio. This distinct feature, together with the typical “twisted” architecture of spirocycles, makes these molecular frameworks particularly relevant for the development of compound collections for drug discovery programs. Sp^3 -rich structures possess

several out-of-plane substituents, which at least in principle, would favour ligand–target complementarity,⁵⁵ providing greater selectivity⁵⁶ and higher success in the progress towards clinical development.⁵⁴

Putting together the results of the chemoinformatic analysis of 466 238 NPs, 4699 acetal-containing compounds were identified, which were clustered in 14 fused, 17 bridged and 14 spiroacetal topologically different molecular frameworks. For the large set of natural products considered in this work, compounds from marine sources and TCM were found to possess a higher number of bicyclic acetals, with a prevalence of [6,5]- and [6,6]-spiroacetals, respectively. Among the family of bridged bicyclic acetals, 6,8-dioxabicyclo[3.2.1]octane

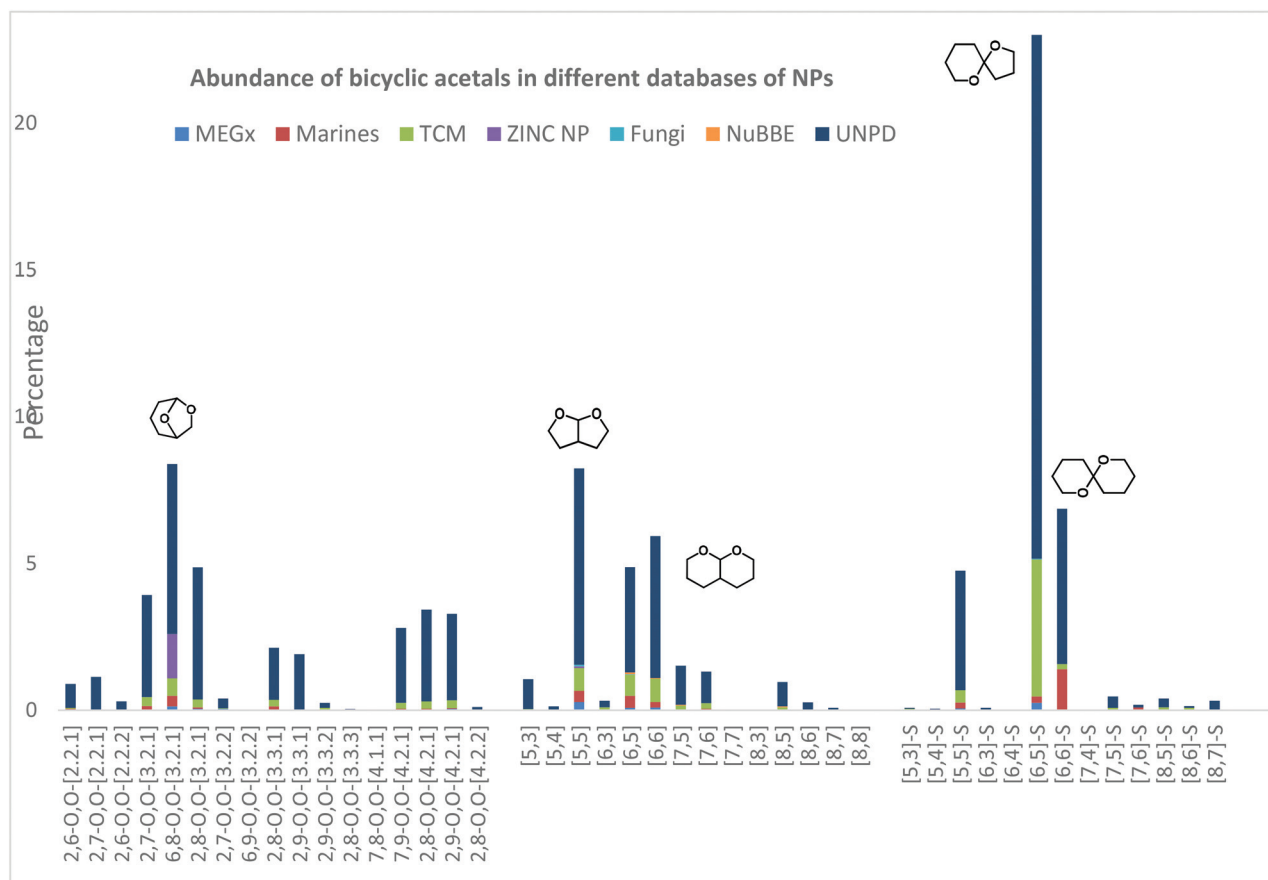


Fig. 5 Abundance of bridged, fused and spiro bicyclic acetals in seven different databases of natural products available in the public domain (466 238 compounds in total). For each molecular framework, the relative percentage of natural products containing this scaffold *versus* the total number of acetal-containing compound (4699). Each data set from different origin is identified with a different color: MEGx in light blue, marine compounds in red, Traditional Chinese Medicine (TCM) in green, ZINC Natural Product (ZINC NP) in purple, Fungi metabolites in cyan, NuBBE compounds in orange and Universal Natural Products Database (UNPD) in dark blue.

proved to be the most abundant framework and its presence stands out in the ZINC NP data set. The quantification of the scaffold diversity of the 4699 acetal-containing natural products, based both on CSR and on the SSE values, showed that fused bicyclic acetals have the largest diversity, while the quantification of their three-dimensional complexity revealed that spiroacetals possess the highest F_{sp}^3 index and a high structural complexity.

All in all, the results of this analysis provide a tool for correlating the bicyclic acetal chemotypes with the natural product domain, and for selecting which scaffolds may be optimal for tuning both the chemical diversity and complexity of novel small molecules containing bicyclic acetals.

5. Acetal chemistry in diversity-oriented synthesis

Taking into account the considerations about biological relevance and the structural features assessed by the chemoinformatics analysis of bicyclic acetals in NPs, the conformational

and configurational flexibility of these moieties provides a wide stereochemical and skeletal diversity that can be exploited in the divergent synthesis of chemical libraries. For example, a spiroacetalisation reaction can potentially give access to four different stereoisomers (Fig. 8a), depending on the contribution of hydrogen bonding networks, steric hindrance, anomeric effect and chelation interactions, which can preferentially stabilize one configuration more than others. Moreover, the acid-catalysed interconversion between different spiroacetal systems can lead to skeletally different arrangements of molecular architecture, providing topologically different frameworks (Fig. 8b).

Bicyclic acetal skeletons are also attractive for their high density of polar functional groups, which offer many possibilities in the scaffold decoration and further chemical manipulations. In particular, they are oxygen-rich moieties, and considering that oxygen atoms play an important role in hydrogen-bonding interactions, their introduction into small molecules can increase the binding capabilities with biomacromolecules.⁵⁷ Indeed, several review papers are present in the literature, describing strategies to construct bicyclic acetal moi-

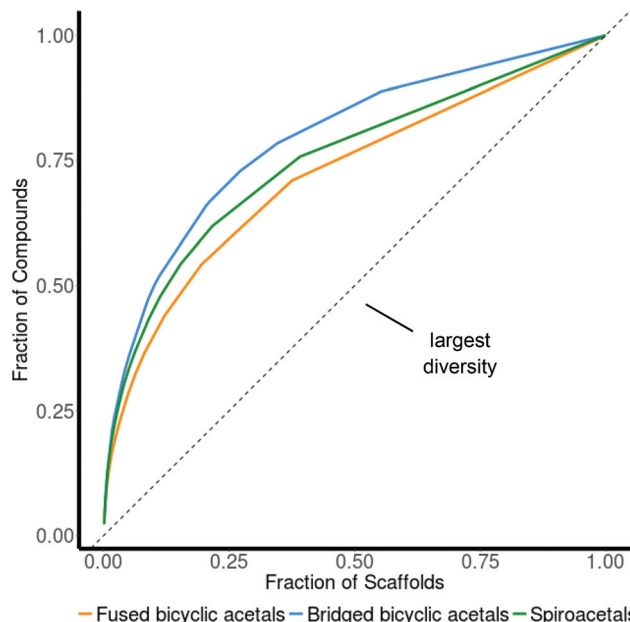


Fig. 6 Cyclic system retrieval (CSR) curves for the data sets studied in this work. The curves can be characterized quantitatively by the area under the curve (AUC) and the fraction of chemotypes required to retrieve 50% of the compounds in the data sets F_{50} (see Table 2).

Table 2 Summary scaffold diversity measures of three major classes of bicyclic acetals^a

Class	N	N_{sing}	N_{sing}/N	AUC	F_{50}	SSE10
Fused	1578	733	0.465	0.7195	0.1655	0.927
Bridged	2159	544	0.252	0.7927	0.101	0.882
Spiroacetals	2633	1049	0.398	0.749	0.126	0.913

^a N : number of scaffolds, N_{sing} : number of singletons, AUC: area under the curve, F_{50} : Fraction of chemotypes that contain 50% of the dataset, SSE10 Scaled Shannon Entropy of the 10 most frequent scaffolds.

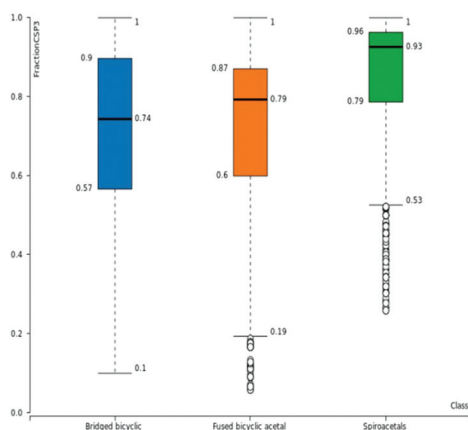


Fig. 7 Box plots of the distribution of the Saturation Index (F_{sp}^3) for bridged (blue), fused (orange) and spiro (green) bicyclic acetal compounds. For each box, summary statistics (Minimum, Lower Whisker, Lower Quartile, Median, Upper Quartile, Upper Whisker) are indicated.

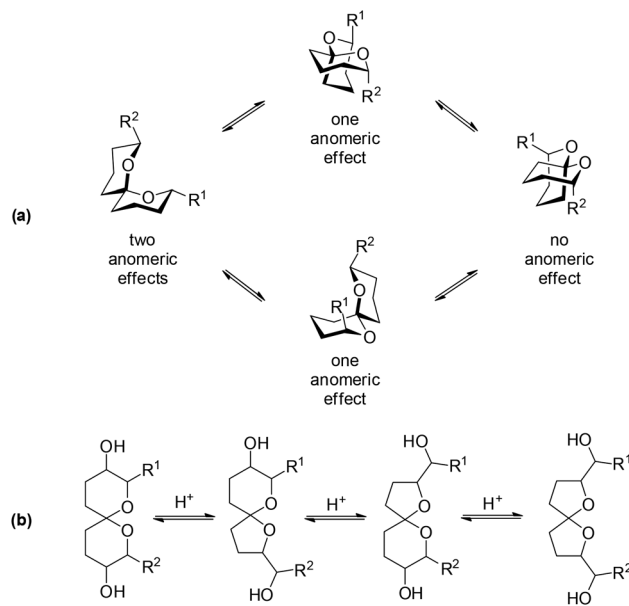


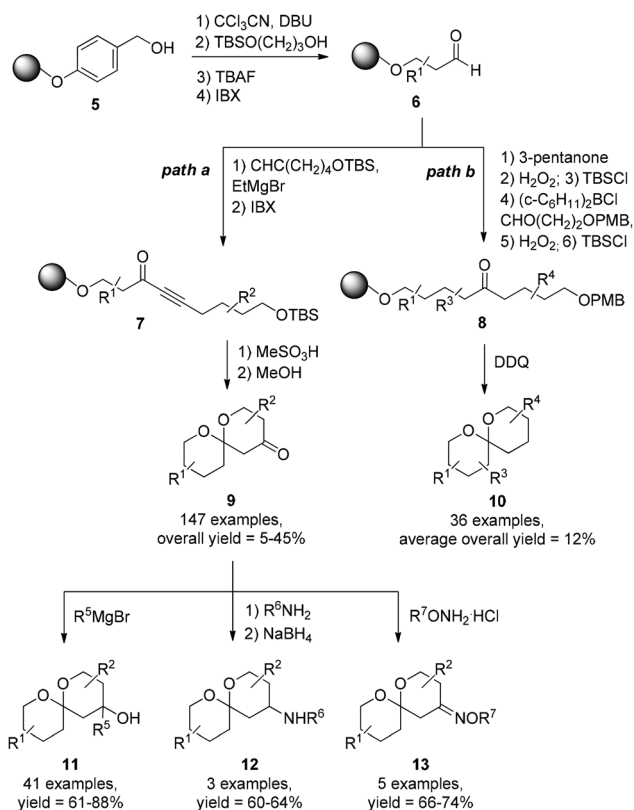
Fig. 8 (a) Stereochemical diversity coming from different configurational stereoisomers of a [6,6]spiroacetal and (b) skeletal diversity resulting from the interconversion between four different acetal-containing ring systems.

eties, especially in the context of the total synthesis of natural products.^{8a,h,9,42,58}

Despite the fact there is progress in the field, there is still much room left for the exploitation of this chemistry in the DOS of novel chemical libraries, although some recent applications of acetal-containing derivatives in drug discovery programmes have appeared in the literature.⁵⁹ In this section, we highlight selected synthetic efforts to access the three different classes of *O,O*-acetal molecular frameworks on a diversity-oriented synthesis⁶ perspective, particularly referring to how these studies have led to the discovery of novel biologically active compounds against different targets.

5.1 Diversity-oriented synthesis of bicyclic spiroacetals

In the context of solid phase synthesis,⁶⁰ the preparation of spiroacetal collections has received great improvement by Waldmann's research group.⁶¹ Although they require several synthetic steps, more than 250 [6,6]-spiroacetals were prepared starting from commercial polystyrene resin **5**, exploiting the double intramolecular hetero-Michael addition of alkynone **7** (Scheme 1, path a),⁶² or taking advantage of the acid-catalyzed spiroacetalization of immobilized aldol intermediate **8** (Scheme 1, path b).⁶³ Further chemical diversification was then obtained from ketone containing spiroacetals **9** by using Grignard additions, reductive aminations and oxime formations in the solution phase. Some selected components of this [6,6]-spiroacetal collection were tested for their biological activity as phosphatase inhibitors and as tubulin cytoskeleton formation modulators (Fig. 9). The biological results showed that spiroacetal **14** is an active inhibitor of the vaccinia virus VH1-related phosphatase (VHR), whereas spiroacetal **15** is able



Scheme 1 Solid phase synthesis of a [6,6]-spiroacetal collection reported by Waldmann and coworkers.

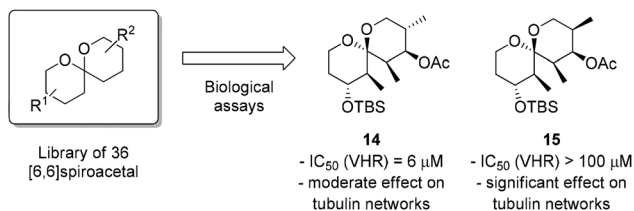
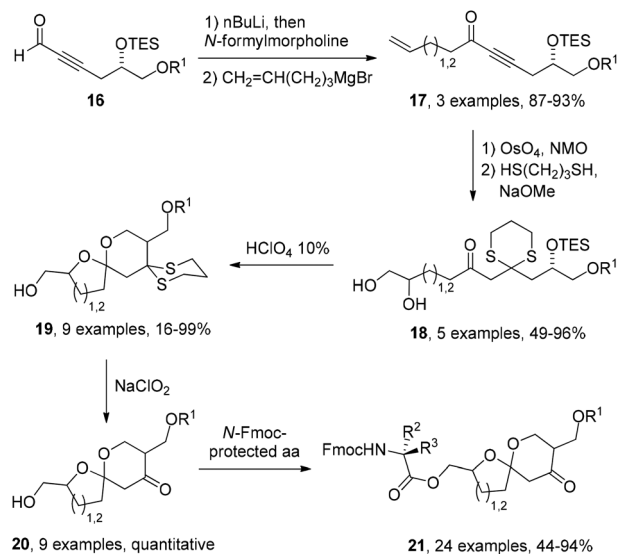


Fig. 9 Phosphatase inhibition and tubulin modulation activity of [6,6]-spiroacetals **14** and **15**.

to affect the tubulin and actin networks in MDA-MB-231 cells, by targeting a molecular process that is different from the usual microtubule polymerization one.⁶⁴

Also, Milroy and Ley reported the synthesis of a small molecule collection based on [6,6]- and [6,5]-spiroacetal frameworks starting from ynal **16** (Scheme 2).⁶⁵ Briefly, different stereoisomers of spiroacetal derivatives **19** were obtained from the β -keto-1,3-dithiane precursor **18** in specific isomeric ratios, influenced by the presence of the lone pair orbitals of the sulfur atoms of the dithiane group.⁶⁶ Then, the dithiane group was removed by a chemoselective oxidation with NaClO_4 to achieve ketone derivatives **20** in quantitative yields. Finally, the free hydroxyl group of spiroacetals **20** was coupled with a variety of *N*-Fmoc-protected α -amino acids, obtaining a small library of bicyclic spiroacetals **21** (Scheme 2).



Scheme 2 Synthesis of a library of [6,6]- and [6,5]-spiroacetals **20** and **21**.

To assess the biological activity of these novel compounds, a small collection of 15 spiroacetals was subjected to a cytotoxicity assay against B-cell chronic lymphocytic leukaemia cells (CLL), a type of leukaemia that has currently no treatments.⁶⁷

Compound **22** was selected for its submicromolar activity and used to develop a second generation library around this structure (Fig. 10).²¹ The screening of this second generation library showed an increase in activity, thanks to structure-activity relationship indications. In particular, the most potent spiroacetal compound was found to be **23**, the enantiomer of **22**, with the ketone functionality replaced by the bulkier dithiane group. Preliminary biological investigation indicated that these spiroacetals are able to induce cell death by an apoptotic pathway, which was found to be selective for CLL cell lines, as these compounds were inactive in MCF7 and A549 cells.²¹

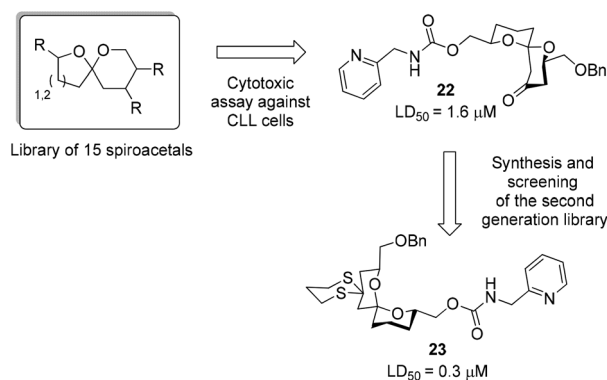
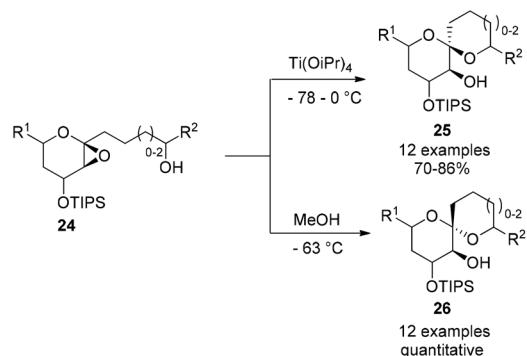


Fig. 10 First generation and second generation library of spiroacetals and biological activity towards CLL cells.



Scheme 3 Reagent-based approach for the synthesis of stereochemically different spiroacetals starting from glycol epoxide **24**.

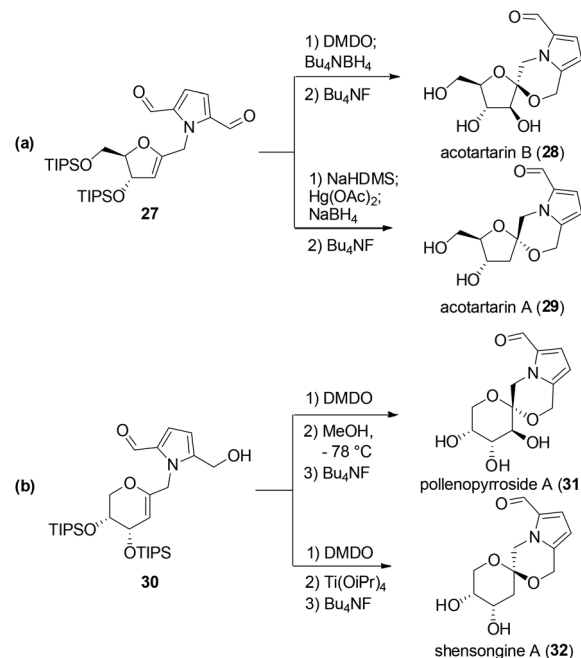
Finally, in the context of diversity-oriented synthesis of spiroacetals, Tan and coworkers⁶⁸ applied a reagent-based approach⁶⁹ for the synthesis of stereochemically different spiroacetals. This approach consists of applying different reaction conditions on the same substrates and differs from the substrate-based strategy⁷⁰ in which different compounds are obtained by using the same reaction conditions on various starting materials. In particular, these authors exploit kinetically controlled spirocyclization that is independent of the thermodynamic stability of the resulting products.⁶⁸ For example, anomeric spiroacetal **25** was obtained starting from glycol epoxide **24** with inversion of configuration, when the spirocyclization was performed under methanol mediated conditions, whereas spiroacetal **26**, with retention of configuration, was obtained by using the $\text{Ti}(\text{OiPr})_4$ control (Scheme 3).⁷¹

Similarly, the same approach was applied by the authors in the preparation of pyrrolomorpholine natural products acotararins (**28–29**), pollenopyrroside (**31**) and shensongine A (**32**), starting from pyrrole-functionalized D -furanose glycol substrates **27** (Scheme 4a)⁷² and D -pyranose glycol substrates **30** (Scheme 4b).⁷³

Thermodynamically less favoured acotararin A (**29**) and shensongine A (**32**) were obtained by using other methods employing $\text{Hg}(\text{OAc})_2$ and $\text{Ti}(\text{OiPr})_4$, respectively. These natural products exhibit antioxidant activity and may have therapeutic potential in the treatment of oxidative-stress related diseases, such as diabetes. Thus, to investigate the potential of these compounds and to improve their biological activity, the authors assayed some naturally occurring pyrrolomorpholine spiroacetals, as well as some corresponding C_2 -hydroxy analogues, in a glucose-induced oxidative stress test, revealing that compounds **33** and **34** were more active than their parent natural products, pollenopyrroside A (**31**) and shensongine A (**32**) (Fig. 11).⁷³

5.2 Diversity-oriented synthesis of bridged bicyclic acetals

In addition to the synthesis of spiroacetals, Milroy and Ley took advantage of their expertise in acetal chemistry,⁷⁴ and extended their studies towards the synthesis of bridged bicyclic



Scheme 4 Diversity Oriented Synthesis of the pyrrolomorpholine spiroacetal family of natural products starting from pyrrole functionalized glycol substrates **27** and **30**.

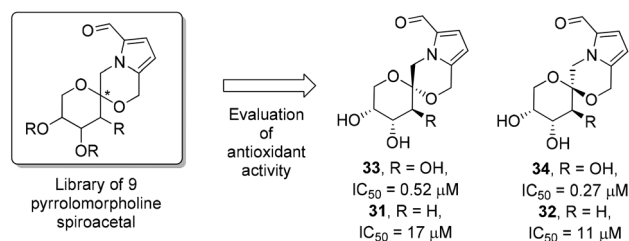
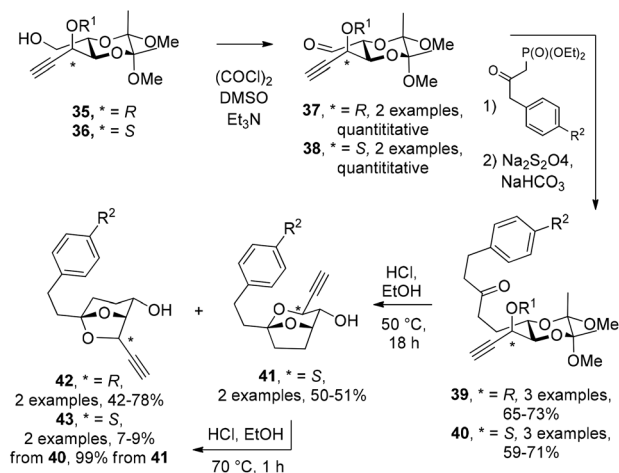


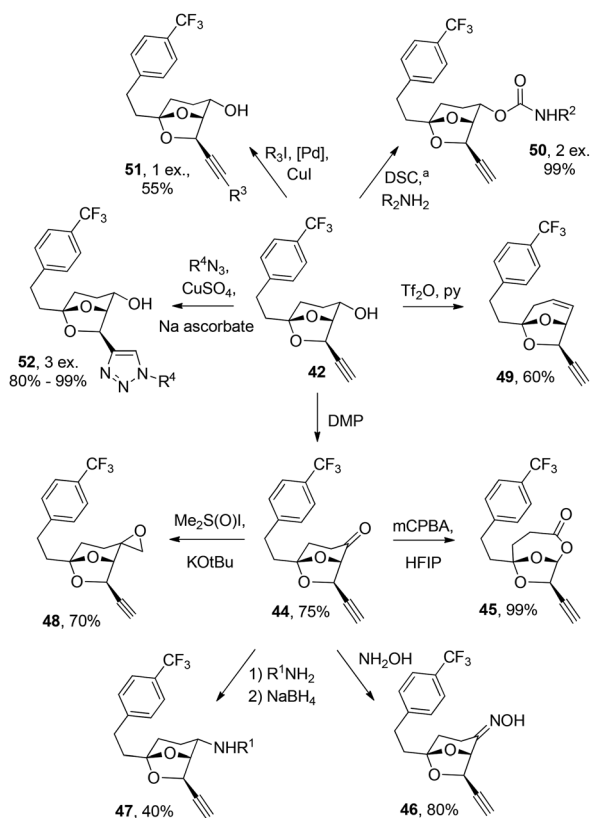
Fig. 11 Inhibition of high glucose-induced ROS production by pyrrolomorpholine spiroketal natural products **31** and **32** and analogues **33** and **34**.

acetal scaffolds.⁷⁵ In particular, a small collection of 2,8- and 6,8-dioxabicyclo[3.2.1]octanes **41–43** was obtained starting from butane-2,3-diacetal-protected tartrate derivatives **35** (and its diastereoisomer **36**), as shown in Scheme 5. Swern oxidation of **35/36** gave aldehydes **37/38**, which, after reaction with the phosphonate diesters and the subsequent reduction of the resulting enones, allowed the acid-catalyzed acetalisation of **39/40** into 6,8-dioxabicyclo[3.2.1]octanes **42** and **43**.

Compound **40** cyclized preferentially into 2,8-dioxabicyclo[3.2.1]octane **41**; however, after chromatographic purification, the more thermodynamically stable 6,8-dioxabicyclo[3.2.1]octane **43** was isolated as the major product. Starting from bicyclic acetal **42**, a second generation library was obtained with further synthetic elaborations (Scheme 6). The alcohol functionality of **42** gave access to the ketone derivative **44** by oxidation with DMP, which was then transformed into lactone **45**, oxime **46**, amine **47** and epoxide **48**. The alcohol function-



Scheme 5 Synthesis of 2,8-dioxabicyclo[3.2.1]octane **41** and 6,8-dioxabicyclo[3.2.1]octane **42-43**.



Scheme 6 Skeletal diversity from 6,8-dioxabicyclo[3.2.1]octane **42**. ^aDSC: disuccinimidyl carbonate.

ality of **42** was also exploited to obtain the olefin-containing compound **49**, by base-mediated elimination, and the carbamate derivatives **50** by reaction with disuccinimidyl carbonate (DSC) and different amines. Finally, acetylene reactivity was involved for introducing appendages, by a Sonogashira coupling (compound **51**) and different copper(I)-catalyzed azide alkyne cycloaddition (CuAAC) reactions (compound **52**).

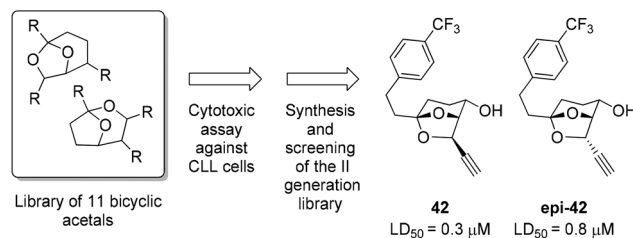


Fig. 12 First generation and second generation library of bicyclic acetals and biological activity towards CLL cells.

As previously mentioned for spiroacetals, also in this case, a small collection of 11 bicyclic acetals was evaluated for biological activity against B-cell chronic lymphocytic leukaemia cells (CLL).²¹ However, in this case, no interesting results were found in the second generation library developed around compound **42**, as its diastereomer *epi-42* showed comparable activity (Fig. 12), proving that the appendages, such as the acetylene functionality, were more important than the three-dimensional structure of the scaffold.

Our contribution in this field involved the development of aza-bicyclic acetals, particularly consisting of the 6,8-dioxabicyclo[3.2.1]octane molecular framework, through two-step strategies based on a coupling reaction between an amino carbonyl derivative and a diol species, followed by acid-catalyzed acetalization of the resulting coupling intermediate (Fig. 13).

All the compounds synthesized by our group in these years were achieved in no more than 4 steps,⁷⁶ including the bicyclic acetals obtained from tartaric acid and more complex sugar derivatives, such as mannose.^{76c} A representative synthesis, fol-

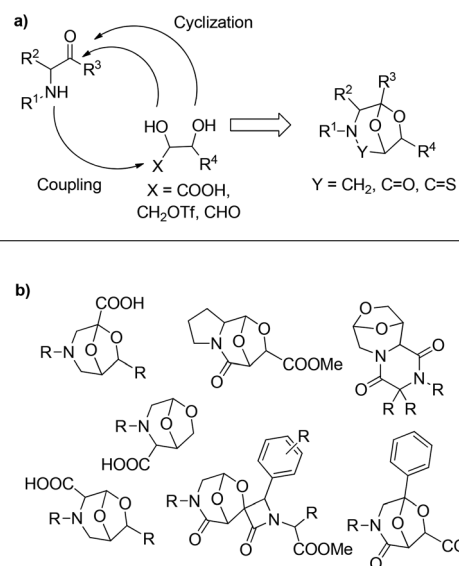
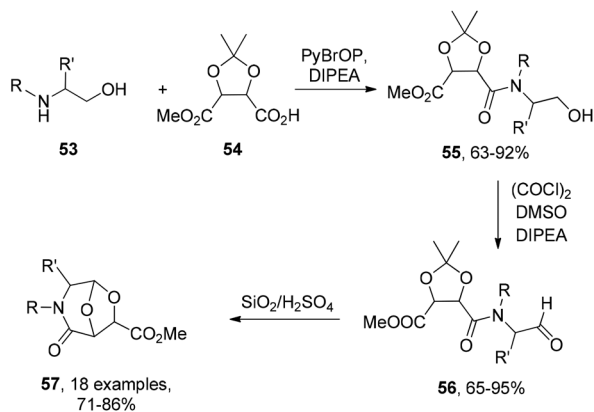


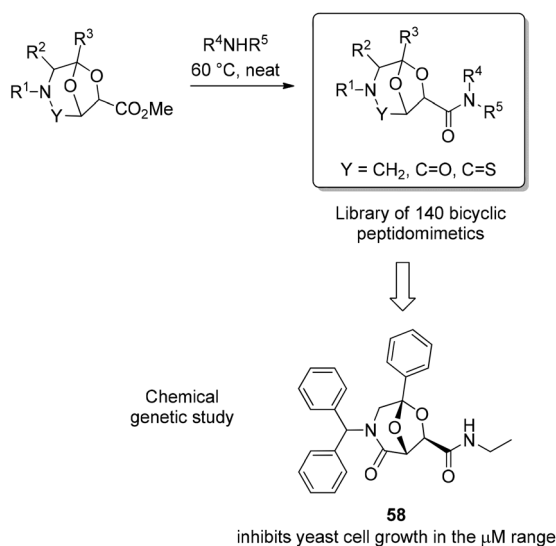
Fig. 13 (a) Strategic approach for the synthesis of peptidomimetic scaffolds starting from sugar and amino acids derivatives. (b) Representative examples of bicyclic rigid scaffolds obtained with this strategy.



Scheme 7 Synthesis of the 2-oxo-6,8-dioxa-3-azabicyclo[3.2.1]octane library.

lowing the principle of the build/couple/pair DOS strategy,⁷⁷ is reported in Scheme 7. Amino alcohol **53**, derived from an amino acid, was coupled to the tartaric acid derivative **54** to give **55**, which was subjected to oxidation to the corresponding aldehyde **56**. The acetalization process consisted of a diol deprotection and concomitant *trans*-acetalization to give the corresponding 2-oxo-6,8-dioxa-3-azabicyclo[3.2.1]octane **57**.⁷⁶

This bicyclic scaffold proved to act as a dipeptide isostere and revealed peculiar biological features against different targets, such as aspartyl proteases (HIV protease, *C. albicans* SAP2, BACE-1).⁷⁸ Also, a focused library of 140 bicyclic dipeptide isosteres was obtained by the slow addition of the amine to the reactive carbomethoxy group of **57**,^{76d} and assayed in a phenotypic screening for the ability to inhibit cell growth in *S. cerevisiae* as a model system,⁷⁹ thus identifying compound **58** as a cell growth inhibitor that interferes with a molecular target localized at the cell wall level (Fig. 14).⁸⁰



Scheme 8 Substrate-based diversity-oriented synthesis of 5,5-fused bicyclic acetals **61** and 5,6-bridged bicyclic acetals **62** starting from hydroxyarenes **60** and different γ -keto-enals **59**.

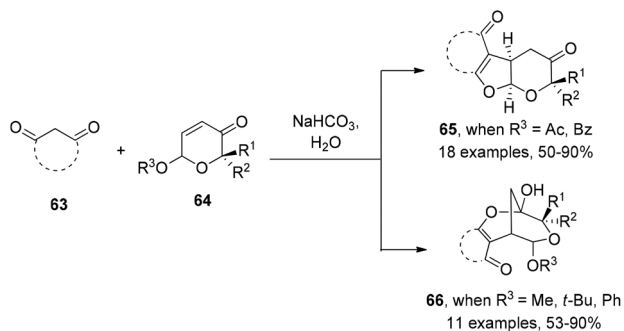
5.3 Diversity-oriented synthesis of fused bicyclic acetals

Although there is still much room left for the development of small molecule collections containing fused bicyclic acetals, some substrate-based divergent strategies to assess these molecular frameworks have been reported both by Jørgensen and Ramasastry groups. Specifically, Jørgensen and coworkers⁸¹ obtained 5,5-fused bicyclic acetals **61** and 5,6-bridged bicyclic acetals **62** taking advantage of an iminium ion organocatalyzed cascade approach on hydroxyarenes **60** and γ -keto-enals **59** (Scheme 8), depending on the nature of the R^1 substituent in the γ -keto- α,β -unsaturated aldehyde **59**. Hence, the process applied to alkyl substituted **59** afforded the tetrahydrofurobenzofurans **61**, while aryl substituents favoured the formation of methanobenzodioxepines **62**. Although no biological activity data are reported for compounds **61** and **62**, these benzofused acetal scaffolds are found in many members of the aflatoxin and bullaketals families of natural products, as well as in butyrylcholinesterase inhibitors for the treatment of Alzheimer's disease,⁸² highlighting the potential of this strategy for the development of small molecule collections around these frameworks.

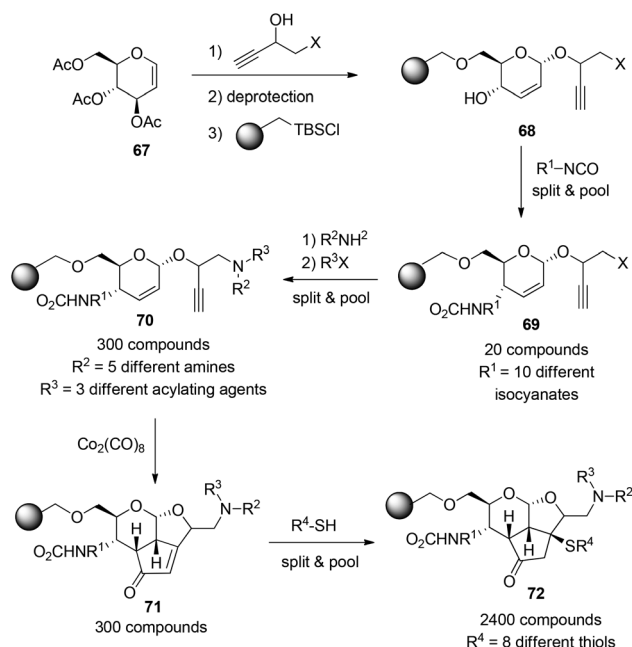
Similarly, the substrate-based divergent synthesis of [5,6]-fused bicyclic acetals **65** and 2,7-dioxabicyclo[3.3.1]nonenones **66** has been performed by Ramasastry and coworkers, starting from 1,3-dicarbonyls **63** and different pyranones **64** (Scheme 9).⁸³ In particular, the cascade process involving Michael addition and cycloacetalization led to [6,5]-fused bicyclic acetals **65** when acetoxy and benzoyloxy pyranones **64** have been used, whereas the same process applied to alkyloxy and aryloxy substituted **64** afforded the bisacetals **66**.

Pyranosides have been also exploited by Schreiber and coworkers to generate appendage diversity around a tricyclic scaffold containing a [6,5]-fused bicyclic acetal (structure **72**, Scheme 10).⁸⁴ More than 2400 tricyclic compounds have been achieved following a solid-phase *split-and-pool* technique, starting from 3,4,6-tri-*O*-acetyl-D-glucal **67** and exploiting, subsequently, a Ferrier reaction with 2 different propargyl alcohols, the 4-hydroxy functionalization with 10 different isocyanates, the installation and further functionalization of a sec-

Fig. 14 Identification of bicyclic acetal **58** as an active antifungal agent against *S. cerevisiae*.



Scheme 9 Substrate-based diversity-oriented synthesis of 5,6-fused bicyclic acetals **65** and 2,7-dioxabicyclo[3.3.1]nonenones **66** starting from 1,3-dicarbonyls **63** and different pyranones **64**.



Scheme 10 Split-and-pool diversity oriented synthesis of a library of 2400 tricyclic compounds containing the [6,5]-fused bicyclic acetal molecular framework.

ondary amine group, and the derivatization of the enone functionality of compounds **71** with 8 different thiols. This huge library of 2400 compounds can represent a powerful platform for high-throughput screening, phenotypic assays and chemical genetics studies.

6. Conclusions and perspectives

Bicyclic acetals are attractive molecular frameworks for the development of high quality compound collections for drug discovery projects. Firstly, these moieties play a key role in the biological outcome of natural products, whether directly interacting with the biological target or acting as a scaffold to direct side chains into specific directions. Secondly, the distinct conformational

flexibility of these frameworks assists the molecule in assuming the three-dimensional profile required for optimal binding to a biomacromolecule, resulting in a variety of different structural arrangements and subsequent biological outcomes. Bicyclic acetals can have a broad variety of topologically different ring combinations, spanning from fused, bridged and spiro ring architectures. To study this structural diversity, the presence of bicyclic acetals has been systematically surveyed in a pool of 466 238 NPs, obtained by combining the database of different sources and accessible in the public domain. This analysis resulted in the identification of a large variety of molecular scaffolds, with a particular abundance of spiroacetals, and allowed assessment of the occurrence of bicyclic acetals in natural products, depending on their origins. Marine compounds were found to be characterized by a high presence of [6,6]-spiroacetals, whereas in natural products from UNPD and TCM, a higher frequency of [6,5]-spiroacetals was observed. Also, among the family of bridged bicyclic acetals, 6,8-dioxabicyclo[3.2.1]octane proved to be the most abundant framework, especially in the ZINC NP database. The scaffold diversity and complexity of bicyclic acetal-containing compounds were also assessed by using different chemoinformatic approaches, such as the CSR graph, the SSE value and the saturation index, revealing that fused bicyclic acetals present a higher diversity, while spiroacetals are characterized by the highest three-dimensional complexity. With this notable scaffold diversity and complexity, bicyclic acetals represent a particularly relevant class of molecular scaffolds for the development of high quality small molecule collections.

The feasibility and versatility of the chemistry of bicyclic acetals allow us to easily generate large collections of such compounds, following DOS approaches. Although there is still much room left in this field, some efforts on the development of bicyclic acetal collections and their application in high-throughput screening and chemical genetics studies allowed for the discovery of novel biologically active compounds. We expect that in the future bicyclic acetals will be taken into account more often for the generation of high value small-molecules useful for drug discovery.

Conflicts of interest

There are no conflicts of interest to declare.

Acknowledgements

Fondazione CR Firenze (grant 2015.0935) and MIUR (PRIN2015, cod. 20157WW5EH) are acknowledged for financial support. FS-G thanks *Consejo Nacional de Ciencia y Tecnologia* (CONACyT, México) for the graduate scholarship.

References

- (a) R. Balamurugan, F. J. Dekker and H. Waldmann, *Mol. BioSyst.*, 2005, **1**, 36; (b) W. Wilk, T. J. Zimmermann,

- M. Kaiser and H. Waldmann, *Biol. Chem.*, 2010, **391**, 491; (c) K. C. Morrison and P. J. Hergenrother, *Nat. Prod. Rep.*, 2014, **31**, 6; (d) A. L. Harvey, R. Edrada-Ebel and R. J. Quinn, *Nat. Rev. Drug Discovery*, 2015, **14**, 111.
- 2 B. E. Evans, K. E. Rittle, M. G. Bock, R. M. DiPardo, R. M. Freidinger, W. L. Whitter, G. F. Lundell, D. F. Veber, P. S. Anderson, R. S. L. Chang, V. J. Lotti, D. J. Cerino, T. B. Chen, P. J. Kling, K. A. Kunkel, J. P. Springer and J. Hirshfield, *J. Med. Chem.*, 1988, **31**, 2235.
 - 3 R. Kombarov, A. Altieri, D. Genis, M. Kirpichenok, V. Kochubey, N. Rakitina and Z. Titarenko, *Mol. Diversity*, 2010, **14**, 193.
 - 4 J. A. Wells and C. L. McClendon, *Nature*, 2007, **450**, 1001.
 - 5 (a) G. L. Thomas and C. W. Johannes, *Curr. Opin. Chem. Biol.*, 2011, **15**, 516; (b) F. Lopez-Vallejo, M. A. Giulianotti, R. A. Houghten and J. L. Medina-Franco, *Drug Discovery Today*, 2012, **17**, 718; (c) Y.-J. Zheng and C. M. Tice, *Expert Opin. Drug Discovery*, 2016, **11**, 831.
 - 6 (a) C. Cordier, D. Morton, S. Murrison, A. Nelson and C. O'Leary-Steele, *Nat. Prod. Rep.*, 2008, **25**, 719; (b) L. A. Marcaurelle and C. W. Johannes, *Prog. Drug Res.*, 2008, **66**, 189; (c) *Diversity-Oriented Synthesis: Basics and Applications in Organic Synthesis, Drug Discovery, and Chemical Biology*, ed. A. Trabocchi, John Wiley and Sons, Hoboken, NJ, 2013; (d) B. R. Balthaser, M. C. Maloney, A. B. Beeler, J. A. Porco Jr. and J. K. Snyder, *Nat. Chem.*, 2011, **3**, 969.
 - 7 (a) S. Wetzel, R. S. Bon, K. Kumar and H. Waldmann, *Angew. Chem., Int. Ed.*, 2011, **50**, 10800; (b) H. van Hattum and H. Waldmann, *J. Am. Chem. Soc.*, 2014, **136**, 11853; (c) S. K. Maurya, M. Dow, S. Warriner and A. Nelson, *Beilstein J. Org. Chem.*, 2013, **9**, 775.
 - 8 (a) F. Perron and K. F. Albizati, *Chem. Rev.*, 1989, **89**, 1617; (b) W. Francke and W. Kitching, *Curr. Org. Chem.*, 2001, **5**, 233; (c) M. A. Brimble and D. P. Furkert, *Curr. Org. Chem.*, 2003, **7**, 1461; (d) J. E. Aho, P. M. Pihko and T. K. Rissa, *Chem. Rev.*, 2005, **105**, 4406; (e) Y. K. Booth, W. Kitching and J. J. De Voss, *Nat. Prod. Rep.*, 2009, **26**, 490; (f) R. Quach, D. F. Chorley and M. A. Brimble, *Org. Biomol. Chem.*, 2014, **12**, 7423; (g) F.-M. Zhang, S.-Y. Zhang and Y.-Q. Tu, *Nat. Prod. Rep.*, 2018, **35**, 75; (h) S. Favre, P. Vogel and S. Gerber-Lemaire, *Molecules*, 2008, **13**, 2570; (i) A. F. Kluge, *Heterocycles*, 1986, **24**, 1699; (j) J. Sperry, Z. E. Wilson, D. C. K. Rathwell and M. A. Brimble, *Nat. Prod. Rep.*, 2010, **27**, 1117.
 - 9 K. Mori, *Eur. J. Org. Chem.*, 1998, 1479.
 - 10 (a) H. Kiyota, *Top. Heterocycl. Chem.*, 2006, **5**, 65; (b) D. J. Faulkner, *Nat. Prod. Rep.*, 2002, **1**, 1.
 - 11 R. E. Hibbs and E. Gouaux, *Nature*, 2011, **474**, 54.
 - 12 (a) M. P. Sauviat, D. Gouiffes-Barbin, E. Ecault and J. F. Verbist, *Biochim. Biophys. Acta*, 1992, **1103**, 109; (b) D. Riou, C. Roussakis, N. Robillard, J. F. Biard and J. F. Verbist, *Biol. Cell.*, 1993, **77**, 261; (c) S. A. Rizvi, V. Tereshko, A. A. Kossiakoff and S. A. Kozmin, *J. Am. Chem. Soc.*, 2006, **128**, 3882.
 - 13 M. Scheepstra, S. A. Andrei, M. Y. Unver, A. K. H. Hirsch, S. Leysen, C. Ottmann, L. Brunsveld and L.-G. Milroy, *Angew. Chem., Int. Ed.*, 2017, **56**, 5480.
 - 14 J. Young and R. E. Taylor, *Nat. Prod. Rep.*, 2008, **25**, 651.
 - 15 (a) R. W. Hoffmann, *Angew. Chem., Int. Ed. Engl.*, 1992, **31**, 1124; (b) R. W. Hoffmann, *Angew. Chem., Int. Ed.*, 2000, **39**, 2054; (c) E. M. Larsen, M. R. Wilson and R. E. Taylor, *Nat. Prod. Rep.*, 2015, **32**, 1183.
 - 16 P. A. Wender, J. De Brabander, P. G. Harran, J. M. Jimenez, M. F. T. Koehler, B. Lipka, C. M. Park, C. Siedenbiedel and G. R. Pettit, *Proc. Natl. Acad. Sci. U. S. A.*, 1998, **95**, 6624.
 - 17 P. Sun, Q. Zhao, H. Zhang, J. Wu and W. Liu, *ChemBioChem*, 2014, **15**, 660.
 - 18 E. M. Seward, E. Carlson, T. Harrison, K. E. Haworth, R. Herbert, F. J. Kelleher, M. M. Kurtz, J. Moseley, S. N. Owen, A. P. Owens, S. J. Sadowski, C. J. Swain and B. J. Williams, *Bioorg. Med. Chem. Lett.*, 2002, **12**, 2515.
 - 19 (a) W. T. R. Linton, B. Hamilton-Smit and R. L. Persad, *Can. Fam. Physician*, 1975, **21**, 107; (b) A. Bryskie, Spectinomycin, in *Antimicrobial Agents: Antibacterials and Antifungals*, ed. A. Bryskier, ASM Press, Washington, USA, 2005.
 - 20 Y. Ohtake, T. Sato, T. Kobayashi, M. Nishimoto, N. Taka, K. Takano, K. Yamamoto, M. Ohmori, M. Yamaguchi, K. Takami, S.-Y. Yeu, K.-H. Ahn, H. Matsuoka, K. Morikawa, M. Suzuki, H. Hagita, K. Ozawa, K. Yamaguchi, M. Kato and S. Ikeda, *J. Med. Chem.*, 2012, **55**, 7828.
 - 21 L.-G. Milroy, G. Zinzalla, F. Loiseau, Z. Qian, G. Prencipe, C. Pepper, C. Fegan and S. V. Ley, *ChemMedChem*, 2008, **3**, 1922.
 - 22 (a) N. Singh, R. Guha, M. A. Giulianotti, C. Pinilla, R. A. Houghten and J. L. Medina-Franco, *J. Chem. Inf. Model.*, 2009, **49**, 1010; (b) Y. Hu, D. Stumpfe and J. Bajorath, *J. Chem. Inf. Model.*, 2011, **51**, 1742.
 - 23 P. Ertl, S. Roggo and A. Schuffenhauer, *J. Chem. Inf. Model.*, 2008, **48**, 68.
 - 24 (a) N. Brown and E. Jacoby, *Mini-Rev. Med. Chem.*, 2006, **6**, 1217; (b) M. Krier, G. Bret and D. Rognan, *J. Chem. Inf. Model.*, 2006, **46**, 512.
 - 25 A. H. Lipkus, Q. Yuan, K. A. Lucas, S. A. Funk, W. F. Bartelt, R. J. Schenck and A. J. Trippe, *J. Org. Chem.*, 2008, **73**, 4443.
 - 26 G. W. Bemis and M. A. Murcko, *J. Med. Chem.*, 1996, **39**, 2887.
 - 27 S. Wetzel, K. Klein, S. Renner, D. Rauh, T. I. Oprea, P. Mutzel and H. Waldmann, *Nat. Chem. Biol.*, 2009, **5**, 581.
 - 28 Prior to analysis, each molecule was prepared in KNIME using the Wash node provided by the program Molecular Operating Environment (MOE). This node disconnects salts of metals, removes simple components and recalculates united of protonation. With this same program, inorganic compounds were eliminated, as well as repeated compounds.
 - 29 M. González-Medina, F. D. Prieto-Martínez, J. Jesús Naveja, O. Méndez-Lucio, T. El-Elmat, C. J. Pearce, N. H. Oberlies,

- M. Figueroa and J. L. Medina-Franco, *Future Med. Chem.*, 2016, **8**, 1399.
- 30 C. R. Pye, M. J. Bertin, R. S. Lokeya, W. H. Gerwick and R. G. Linington, *Proc. Natl. Acad. Sci. U. S. A.*, 2017, **114**, 5601.
- 31 A. C. Pilon, M. Valli, A. C. Dametto, M. E. F. Pinto, R. T. Freire, I. Castro-Gamboa, A. D. Andricopulo and V. S. Bolzani, *Sci. Rep.*, 2017, **7**, 7215.
- 32 C. Y.-C. Chen, *PLoS One*, 2011, **6**, e15939.
- 33 J. Gu, Y. Gui, L. Chen, G. Yuan, H.-Z. Lu and X. Xu, *PLoS One*, 2013, **8**, e62839.
- 34 L. Xian, P. Jian-Xin, W. Zhi-Ying, Z. Yu, Z. Yong, X. Wei-Lie and S. Han-Dong, *Chem. Biodiversity*, 2010, **7**, 2888.
- 35 (a) S. Mohd-Redzwan, R. Jamaluddin, M. S. Abd.-Musalib and Z. Ahmad, *Front. Microbiol.*, 2013, **4**, 334; (b) L. V. Roze, S.-Y. Hong and J. E. Linz, *Annu. Rev. Food Sci. Technol.*, 2013, **4**, 293.
- 36 A. S. Kende, T. L. Smalley and H. Huang, *J. Am. Chem. Soc.*, 1999, **121**, 7431.
- 37 L. Shoei-Sheng, C. Wen-Chuan and C. Chung-Hsiun, *J. Nat. Prod.*, 2000, **63**, 1580.
- 38 (a) H. Watanabe, M. Takano, A. Umino, T. Ito, H. Ishikawa and M. Nakada, *Org. Lett.*, 2007, **9**, 359; (b) H. Kawagishi, S. Furukawa, C. Zhuang and R. Yunoki, *Explore*, 2002, **11**, 46.
- 39 (a) G. R. Pettit, Z. A. Cichacz, F. Gao, C. L. Herald, M. R. Boyd, J. M. Schmidt and J. N. A. Hooper, *J. Org. Chem.*, 1993, **58**, 1302; (b) M. Kobayashi, S. Aoki and I. Kitagawa, *Tetrahedron Lett.*, 1994, **35**, 1243; (c) G. R. Pettit, *Pure Appl. Chem.*, 1994, **66**, 2271.
- 40 (a) S. Hosseini Bai and S. Ogbourne, *Chemosphere*, 2016, **154**, 204; (b) R. W. Burg, *Antimicrob. Agents Chemother.*, 1979, **15**, 361.
- 41 (a) R. Joshi, S. Sood, P. Dogra, M. Mahendru, D. Kumar, S. Bhangalia, H. C. Pal, N. Kumar, S. Bhushan, A. Gulati, A. K. Saxena and A. Gulati, *Med. Chem. Res.*, 2013, **22**, 4030.
- 42 N. Ibrahim, T. Eggimann, E. A. Dixon and H. Wieser, *Tetrahedron*, 1990, **46**, 1503.
- 43 (a) C. Wagner, H. Anke and O. Sterner, *J. Nat. Prod.*, 1998, **61**, 501; (b) S. Mizutani, K. Komori, T. Taniguchi, K. Monde, K. Kuramochi and K. Tsubaki, *Angew. Chem., Int. Ed.*, 2016, **55**, 9553.
- 44 J. Ren and R. Tong, *J. Org. Chem.*, 2014, **79**, 6987.
- 45 C. E. Stivala and A. Zakarian, *J. Am. Chem. Soc.*, 2008, **130**, 3774.
- 46 F. Matsuura, R. Peters, M. Anada, S. S. Harried, J. Hao and Y. Kishi, *J. Am. Chem. Soc.*, 2006, **128**, 7463.
- 47 (a) P. J. Sidebottom, R. M. Highcock, S. J. Lane, P. A. Procopiu and N. S. Watson, *J. Antibiot.*, 1992, **45**, 648; (b) T. Kawamata, M. Nagatomo and M. Inoue, *J. Am. Chem. Soc.*, 2017, **139**, 1814.
- 48 K. Fujiwara, Y. Suzuki, N. Koseki, Y. Aki, Y. Kikuchi, S. Murata, F. Yamamoto, M. Kawamura, T. Norikura, H. Matsue, A. Murai, R. Katoono, H. Kawai and T. Suzuki, *Angew. Chem., Int. Ed.*, 2014, **53**, 780.
- 49 G. Koren-Goldshlager, P. Klein, A. Rudi, Y. Benayahu, M. Schleyer and Y. Kashman, *J. Nat. Prod.*, 1996, **59**, 262.
- 50 J. Y. Dong, L. M. Wang, H. C. Song, K. Z. Shen, Y. P. Zhou, L. Wang and K. Q. Zhang, *Chem. Biodiversity*, 2009, **6**, 1216.
- 51 A. H. Lipkus, Q. Yuan, K. A. Lucas, S. A. Funk, W. F. Bartelt, R. J. Schenck and A. J. Trippe, *J. Org. Chem.*, 2008, **73**, 4443.
- 52 C. E. Shannon and W. Weaver, *The Mathematical Theory of Communication*, University of Illinois Press, Urbana, IL, 1963.
- 53 J. L. Medina-Franco, K. Martínez-Mayorga, A. Bender and T. Scior, *QSAR Comb. Sci.*, 2009, **28**, 1551.
- 54 F. Lovering, J. Bikker and C. Humblet, *J. Med. Chem.*, 2009, **52**, 6752.
- 55 (a) J. A. Haigh, B. T. Pickup, J. A. Grant and A. Nicholls, *J. Chem. Inf. Model.*, 2005, **45**, 673; (b) W. R. J. D. Galloway, A. Isidro-Llobet and D. R. Spring, *Nat. Commun.*, 2010, **1**, 80; (c) T. Flagstad, G. Min, K. Bonnet, R. Morgentin, D. Roche, M. H. Clausen and T. E. Nielsen, *Org. Biomol. Chem.*, 2016, **14**, 4943; (d) S. Stotani, C. Lorenz, M. Winkler, F. Medda, E. Picazo, R. O. Martinez, A. Karawajczyk, J. Sanchez-Quesada and F. Giordanetto, *ACS Comb. Sci.*, 2016, **18**, 330; (e) E. Lenci, A. Rossi, G. Menchi and A. Trabocchi, *Org. Biomol. Chem.*, 2017, **15**, 9710.
- 56 (a) F. Lovering, *MedChemComm*, 2013, **4**, 515; (b) P. A. Clemons, N. E. Bodycombe, H. A. Carrinski, J. A. Wilson, A. F. Shamji, B. K. Wagner, A. N. Koehler and S. L. Schreiber, *Proc. Natl. Acad. Sci. U. S. A.*, 2010, **107**, 18787.
- 57 M. Feher and J. M. Schmidt, *J. Chem. Inf. Comput. Sci.*, 2003, **43**, 218.
- 58 (a) K. T. Mead and B. N. Brewer, *Curr. Org. Chem.*, 2003, **7**, 227; (b) M. F. Jacobs and W. Kitching, *Curr. Org. Chem.*, 1998, **2**, 395; (c) D. J. Faulkner, *Nat. Prod. Rep.*, 2002, **1**, 1; (d) W. Zhang and R. Tong, *J. Org. Chem.*, 2016, **81**, 2203; (e) S. Bera, B. Chatterjee and D. Mondal, *Eur. J. Org. Chem.*, 2018, 5337.
- 59 B. A. Kulkarni, G. P. Roth, E. Lobkovsky and J. A. Porco Jr., *J. Comb. Chem.*, 2002, **4**, 56.
- 60 (a) K. C. Nicolaou, J. A. Pfeifferkorn, A. J. Roecker, G.-Q. Cao, S. Barluenga and H. J. Mitchell, *J. Am. Chem. Soc.*, 2000, **122**, 9939; (b) R. Liu, X. Li and K. S. Lam, *Curr. Opin. Chem. Biol.*, 2017, **38**, 117.
- 61 S. Sommer and H. Waldmann, *Chem. Commun.*, 2005, 5684.
- 62 S. Sommer, M. Kuhn and H. Waldmann, *Adv. Synth. Catal.*, 2008, **350**, 1736.
- 63 O. Barun, S. Sommer and H. Waldmann, *Angew. Chem., Int. Ed.*, 2004, **43**, 3195.
- 64 O. Barun, K. Kumar, S. Sommer, A. Langerak, T. U. Mayer, O. Muller and H. Waldmann, *Eur. J. Org. Chem.*, 2005, 4773.
- 65 G. Zinzalla, L.-G. Milroy and S. V. Ley, *Org. Biomol. Chem.*, 2006, **4**, 1977.
- 66 M. J. Gaunt, D. F. Hook, H. R. Tanner and S. V. Ley, *Org. Lett.*, 2003, **5**, 4815.

- 67 R. L. Auer, J. Gribben and F. E. Cotter, *Br. J. Haematol.*, 2007, **139**, 635.
- 68 A. L. Verano and D. S. Tan, *Isr. J. Chem.*, 2017, **57**, 1.
- 69 (a) E. E. Wyatt, S. Fergus, W. R. J. D. Galloway, A. Bender, D. J. Fox, A. T. Plowright, A. S. Jessiman, M. Welch and D. R. Spring, *Chem. Commun.*, 2006, 3296; (b) G. L. Thomas, R. J. Spandl, F. G. Glansdorp, M. Welch, A. Bender, J. Cockfield, J. A. Lindsey, C. Bryant, D. F. J. Brown, O. Loiseleur, H. Rudyk, M. Ladlow and D. R. Spring, *Angew. Chem., Int. Ed.*, 2008, **47**, 2808.
- 70 (a) M. D. Burke, E. M. Berger and S. L. Schreiber, *Science*, 2003, **302**, 613; (b) M. D. Burke, E. M. Berger and S. L. Schreiber, *J. Am. Chem. Soc.*, 2004, **126**, 14095; (c) H. Oguri and S. L. Schreiber, *Org. Lett.*, 2005, **7**, 47.
- 71 J. S. Potuzak, S. B. Moilanen and D. S. Tan, *J. Am. Chem. Soc.*, 2005, **127**, 13796.
- 72 J. M. Wurst, A. L. Verano and D. S. Tan, *Org. Lett.*, 2012, **14**, 4442.
- 73 A. L. Verano and D. S. Tan, *Chem. Sci.*, 2017, **8**, 3687.
- 74 (a) S. V. Ley and B. Lygo, *Tetrahedron Lett.*, 1982, **23**, 4625; (b) S. V. Ley, N. J. Anthony, A. Armstrong, M. G. Brasca, T. Clarke, C. Greck, P. Grice, A. B. Jones, B. Lygo, A. Madin, R. N. Sheppard, A. M. Z. Slawin and D. J. Williams, *Tetrahedron*, 1989, **45**, 7161; (c) R. Haag, A. G. Leach, S. V. Ley, M. Nettekoven and J. Schnaubelt, *Synth. Commun.*, 2001, **31**, 2965; (d) M. J. Gaunt, A. S. Jessiman, P. Orsini, H. R. Tanner, D. F. Hook and S. V. Ley, *Org. Lett.*, 2003, **25**, 4819; (e) M. Ball, M. J. Gaunt, D. F. Hook, A. S. Jessiman, S. Kawahara, P. Orsini, A. Scolaro, A. C. Talbot, H. R. Tanner, S. Yamanoi and S. V. Ley, *Angew. Chem., Int. Ed.*, 2005, **44**, 5433.
- 75 L.-G. Milroy, G. Zinzalla, G. Prencipe, P. Michel, S. V. Ley, M. Gunaratnam, M. Beltran and S. Neidle, *Angew. Chem., Int. Ed.*, 2007, **46**, 2493.
- 76 (a) A. Guarna, A. Guidi, F. Machetti, G. Menchi, E. G. Occhiato, D. Scarpi, S. Sisi and A. Trabocchi, *J. Org. Chem.*, 1999, **64**, 7347; (b) A. Trabocchi, G. Menchi, F. Guarna, F. Machetti, D. Scarpi and A. Guarna, *Synlett*, 2006, 331; (c) E. Lenci, G. Menchi, A. Guarna and A. Trabocchi, *J. Org. Chem.*, 2015, **80**, 2182; (d) F. Machetti, I. Bucelli, G. Indiani, C. O. Kappe and A. Guarna, *J. Comb. Chem.*, 2007, **9**, 454.
- 77 (a) T. E. Nielsen and S. L. Schreiber, *Angew. Chem., Int. Ed.*, 2008, **47**, 48; (b) N. Kumagai, G. Muncipinto and S. L. Schreiber, *Angew. Chem., Int. Ed.*, 2006, **45**, 3635; (c) J. M. Mitchell and J. T. Shaw, *Angew. Chem., Int. Ed.*, 2006, **45**, 1722; (d) A. Hercouet, F. Berrée, C. H. Lin, L. Toupet and B. Carboni, *Org. Lett.*, 2007, **9**, 1717; (e) A. K. Franz, P. D. Dreyfuss and S. L. Schreiber, *J. Am. Chem. Soc.*, 2007, **129**, 1020.
- 78 (a) A. Trabocchi, C. Mannino, F. Machetti, F. De Bernardis, S. Arancia, R. Cauda, A. Cassone and A. Guarna, *J. Med. Chem.*, 2010, **53**, 2502; (b) C. Calugi, A. Guarna and A. Trabocchi, *Eur. J. Med. Chem.*, 2014, **84**, 444; (c) R. Innocenti, E. Lenci, G. Menchi, A. Pupi and A. Trabocchi, *Bioorg. Med. Chem.*, 2017, **25**, 5077.
- 79 (a) I. Stefanini, A. Trabocchi, E. Marchi, A. Guarna and D. Cavalieri, *J. Biol. Chem.*, 2010, **285**, 23477; (b) A. Trabocchi, D. Cavalieri and A. Guarna, *Pure Appl. Chem.*, 2011, **83**, 687.
- 80 I. Stefanini, L. Rizzetto, D. Rivero, S. Carbonell, M. Gut, S. Heath, I. G. Gut, A. Trabocchi, A. Guarna, N. B. Ghazzi, P. Bowyer, M. Kapushesky and D. Cavalieri, *Sci. Rep.*, 2018, **8**, 5964.
- 81 B. M. Paz, L. Klier, L. Næsborg, V. H. Lauridsen, F. Jensen and K. A. Jørgensen, *Chem. – Eur. J.*, 2016, **22**, 16810.
- 82 M. Pohanka, *Int. J. Mol. Sci.*, 2014, **15**, 9809.
- 83 S. Kasare, S. K. Bankar and S. S. V. Ramasastry, *Org. Lett.*, 2014, **16**, 4284.
- 84 H. Kubota, J. Lim, K. M. Depew and S. L. Schreiber, *Chem. Biol.*, 2002, **9**, 265.



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.

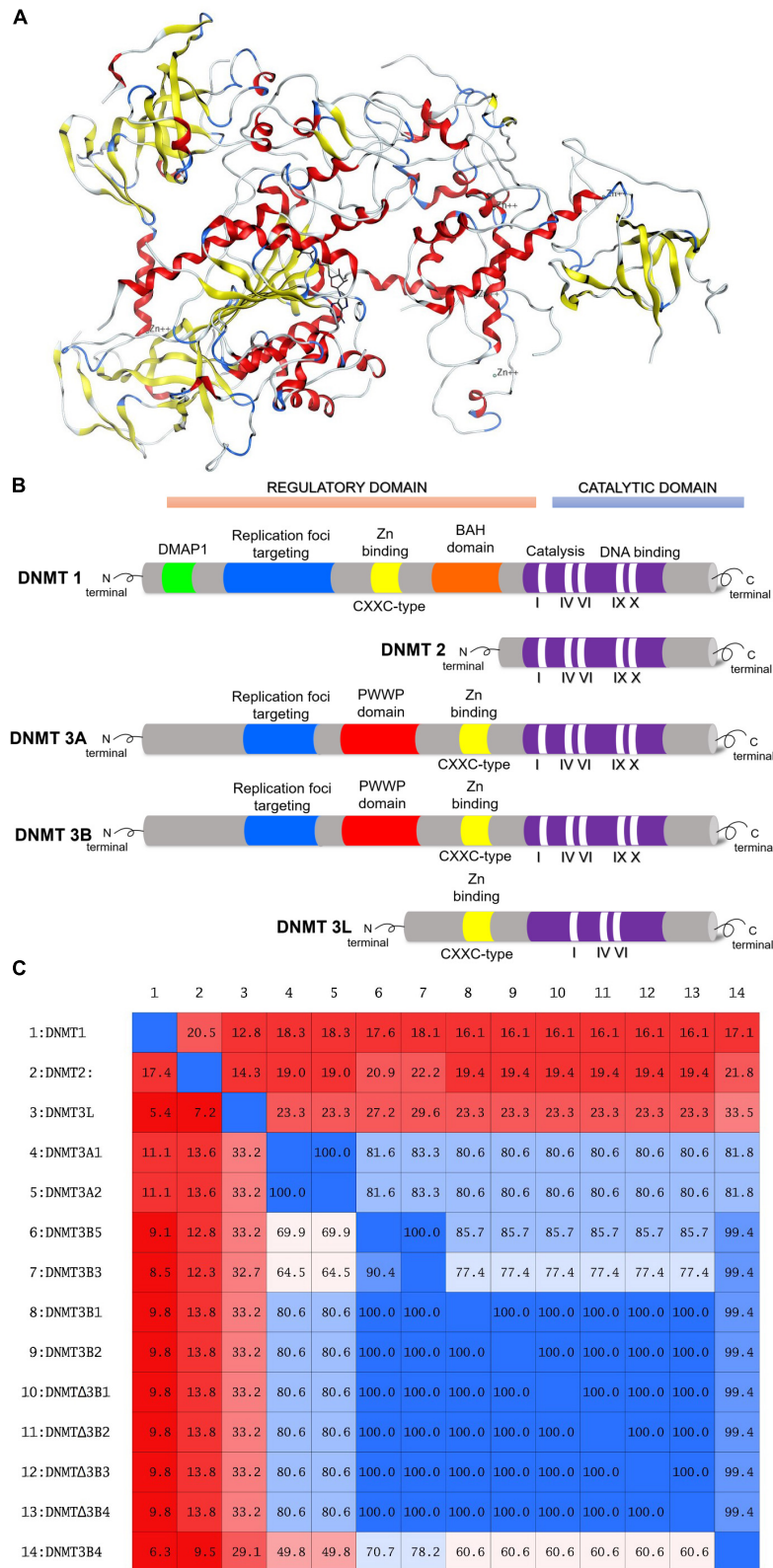


FIGURE 1 | (A) Three-dimensional model of DNMT1, amino acid residues 351–1600. Figure rendered from the Protein Data Bank PDB ID: 4WXX. **(B)** Schematic diagram of the structure of human DNMT1, DNMT2, DNMT3A, DNMT3B, and DNMT3L. **(C)** Identity matrix of the catalytic site of 14 DNMTs isoforms. Note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

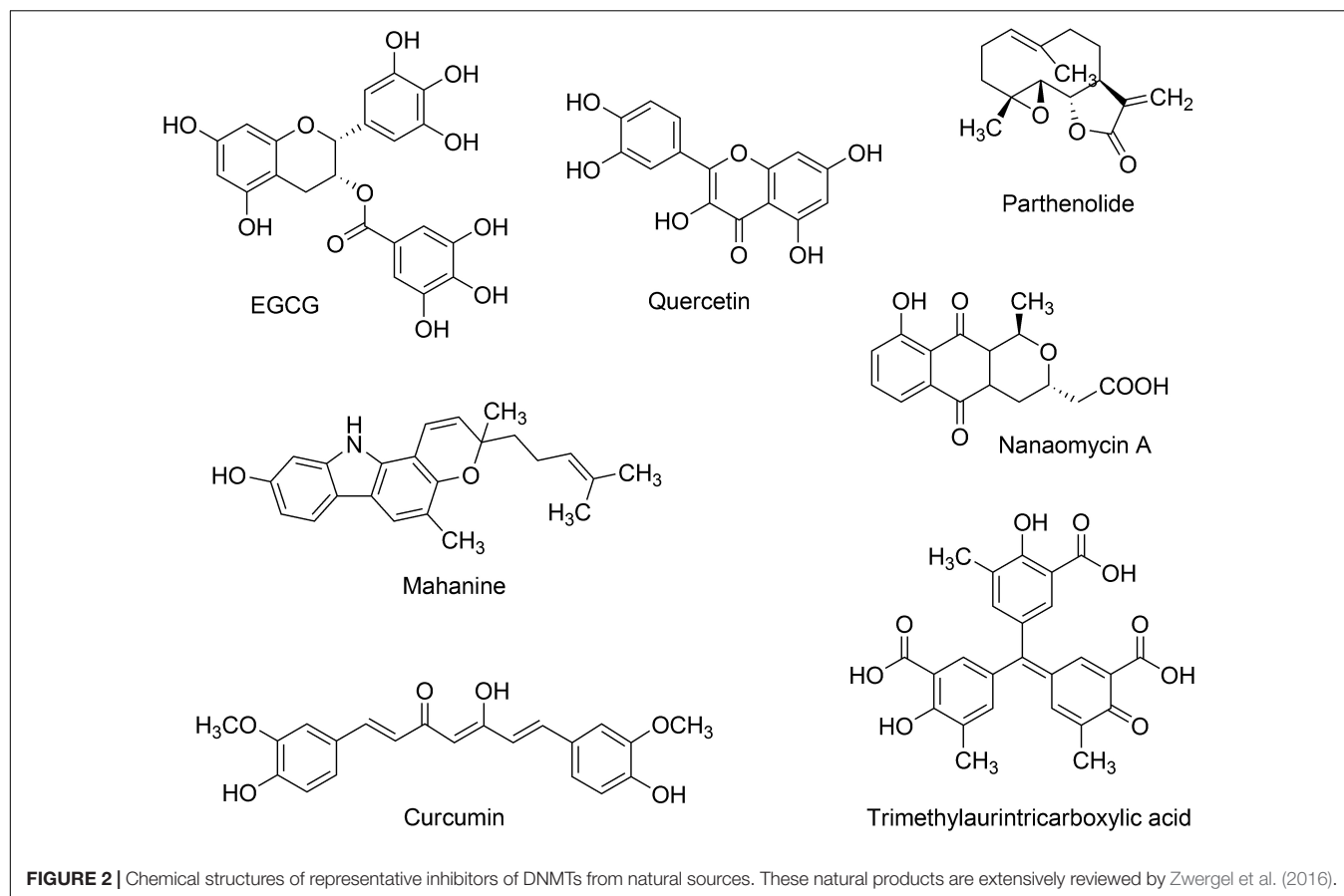
Section 4.1: Natural Products and Food Chemicals

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.

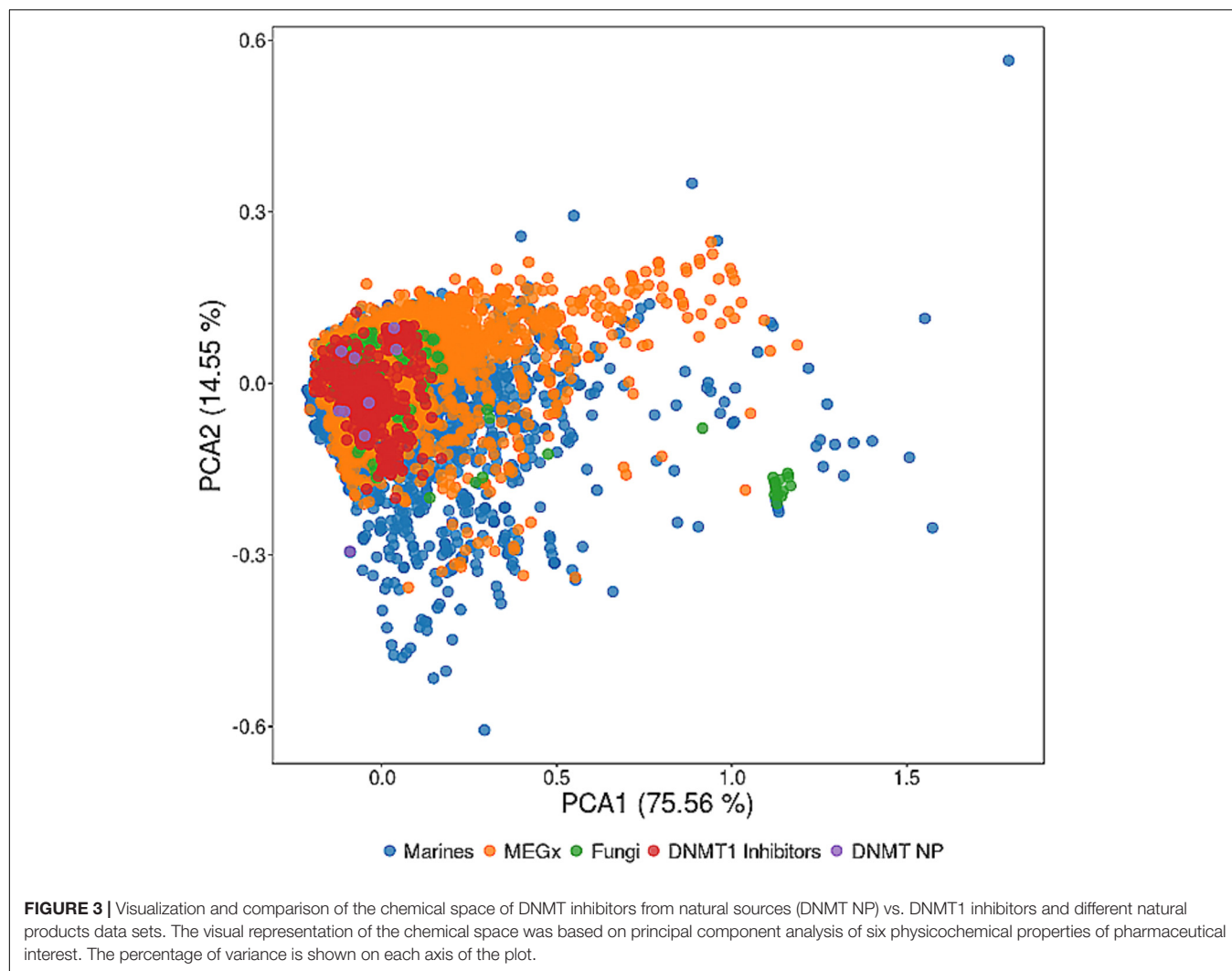
SECTION 6: OPPORTUNITIES FOR SEARCHING FOR NATURAL PRODUCTS AS DNMT INHIBITORS

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 cheminformatic 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 cheminformatic 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
- Maggiara, 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, Pílon-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.