UNIVERSIDAD NACIONAL AUTÓNOMA DE MEXICO

PROGRAMA DE MAESTRÍA Y DOCTORADO EN CIENCIAS QUÍMICAS

DESARROLLO Y ANÁLISIS DE BIBLIOTECAS DE FRAGMENTOS BASADOS EN PRODUCTOS NATURALES

TESIS

PARA OPTAR POR EL GRADO DE

DOCTORA EN CIENCIAS

PRESENTA

M. en C. ANA LUISA CHÁVEZ HERNÁNDEZ

TUTOR
DR. JOSÉ LUIS MEDINA FRANCO
FACULTAD DE QUÍMICA, UNAM

FACULTAD DE QUÍMICA, MAYO 2024

UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

PROGRAMA DE MAESTRÍA Y DOCTORADO EN CIENCIAS QUÍMICAS

DESARROLLO Y ANÁLISIS DE BIBLIOTECAS DE FRAGMENTOS BASADOS EN PRODUCTOS NATURALES

TESIS
PARA OPTAR POR EL GRADO DE

DOCTORA EN CIENCIAS

PRESENTA

M. en C. ANA LUISA CHÁVEZ HERNÁNDEZ

TUTOR
DR. JOSÉ LUIS MEDINA FRANCO
FACULTAD DE QUÍMICA, UNAM

Ciudad de México, mayo 2024

JURADO

Presidente	Dr. Eduardo Guillermo Delgado Lamas	Instituto de Química, UNAM
Vocal	Dr. Francisco Hernández Luis	Facultad de Química, UNAM
Vocal	Dr. Juan Gabriel Navarrete Vázquez	Universidad Autónoma del Estado de Morelos
Vocal	Dra. Carmina Montiel Pacheco	Facultad de Química, UNAM
Secretario	Dr. José Alberto Rivera Chávez	Instituto de Química, UNAM

COMITÉ TUTOR DE EVALUACIÓN

Tutor	Dr. José Luis Medina Franco	Facultad de Química, UNAM
Miembro del comité	Dr. Eduardo Guillermo Delgado Lamas	Instituto de Química, UNAM
Miembro del comité	Dr. Jaime Pérez Villanueva	UAM-Xochimilco

Esta tesis se realizó en los cubículos 309 y 117 del grupo de investigación de Diseño de Fármacos Asistido por Computadora de la Facultad de Química (DIFACQUIM) de la UNAM, Ciudad Universitaria, México, bajo la dirección del Dr. José Luis Medina Franco. La tesis contó con el apoyo económico de la beca de doctorado del Consejo Nacional de Humanidades, Ciencia y Tecnología, CONAHCYT (CVU:847870).

AGRADECIMIENTOS

Con mucho cariño a mi maestro y mentor, el Dr. José Luis Medina Franco, un gran ejemplo a seguir como persona y científico. Gracias por compartir sus conocimientos conmigo, por su apoyo incondicional en la realización de este proyecto de investigación. Gracias por brindarme su amistad y creer en mí.

Con mucho cariño y amor, a mis padres Rosa y Rolando; mis hermanos Rosa, Juan, Ángel y David por ser un pilar en mi formación académica y brindarme su apoyo incondicional. A mis amigos Mabel Enríquez, Carmen Torres, Alejandra Pineda, Kari Salomón, Yesenia Cruz, Enrique Soto y Roberto Márquez por brindarme su apoyo incondicional.

Con mucho cariño y amor a la memoria de mi tío José Juan, mi abuelita María y mi hijo Oliver Alejandro, un gran abrazo hasta el cielo.

Muchas gracias a los miembros de mi comité tutor, el Dr. Guillermo Delgado Lamas y el Dr. Jaime Pérez Villanueva, por sus comentarios, sugerencias y el apoyo durante el desarrollo de esta tesis. A los miembros del jurado, Dr. Eduardo Guillermo Delgado Lamas, Dr. Francisco Hernández Luis, Dr. Juan Gabriel Navarrete Vázquez, Dra. Carmina Montiel Pacheco y Dr. José Alberto Rivera Chávez, muchas gracias por sus comentarios y sugerencias.

Muchas gracias a mis compañeros del equipo de investigación de Diseño de Fármacos Asistido por Computadora de la Facultad de Química de la UNAM (DIFACQUIM). En especial al Dr. Norberto Sánchez por ser mi maestro en el lenguaje de programación Python y compartir sus conocimientos conmigo. A mis compañeros y colaboradores, los maestros en ciencias Fernanda Saldívar, Diana Prado y Raziel Cedillo, por ser mis mentores en la divulgación científica. A mis compañeros y colaboradores, Edgar López, Eurídice Juárez, Felipe Avellaneda, Jazmín Miranda, Alexis Flores, Hassan Villegas, Alejandro Gómez, Johny Rodríguez y Samuel Homberg, por compartir sus conocimientos conmigo. A mis alumnas Daniela Gaytán y Jessica Román, y a todos los integrantes que conforman el DIFACQUIM, muchas gracias por todo.

Muchas gracias al CONAHCYT por el apoyo económico brindado durante mis estudios de Doctorado (CVU:847870), al Posgrado en Ciencias Químicas de la UNAM, a la Facultad de Química de la UNAM, y al subprograma 121 de formación de profesores de la Facultad de Química de la UNAM.

Muchas gracias al DGTIC, UNAM, proyecto LANCAD-UNAM-DGTIC-335; a DGAPA, UNAM, Programa de Apoyo a Proyectos de Investigación e Innovación Tecnológica (PAPIIT), proyecto No. IN201321, y al espacio de innovación UNAM-HUAWEI, proyecto No.7: “Desarrollo y aplicación de algoritmos de inteligencia artificial para el diseño de fármacos aplicables al tratamiento de diabetes mellitus y cáncer”.

Resumen

El propósito de esta tesis fue generar computacionalmente fragmentos moleculares derivados de productos naturales utilizando el lenguaje de programación Python. Los productos naturales fueron caracterizados en términos de la complejidad estructural mediante el promedio de carbonos con hibridación sp^3 y carbonos quirales. La diversidad estructural de los fragmentos fue caracterizada mediante la mediana de similitud de dos huellas digitales moleculares MACCS keys (166-bits) y ECFP4 (1024-bits). Se generaron y analizaron visualizaciones de espacio químico utilizando los algoritmos PCA, t-SNE y TMAP. Se discute que los fragmentos de productos naturales cubren regiones del espacio químico diferentes a las descritas por los compuestos accesibles sintéticamente y compuestos con actividad biológica. Posteriormente, se utilizaron los fragmentos de productos naturales para generar una biblioteca virtual de compuestos análogos a bevirimat, un inhibidor de la proteasa viral del VIH-1. Se discute una mayor diversidad de estructuras generadas a partir de fragmentos de COCONUT en comparación a dos bibliotecas comerciales de fragmentos de productos naturales (ChemDiv y ENAMINE Real).

Por último, las bibliotecas de fragmentos de productos naturales y las bibliotecas de referencia, el protocolo para generar compuestos análogos a bevirimat y los códigos de Python para generar las visualizaciones de espacio químico, están disponibles en línea para seguir promoviendo la ciencia abierta.

ÍNDICE

I. Resumen de resultados de la investigación y actividades del doctorado

1.1. Artículos científicos

Se generaron bibliotecas de fragmentos de bases de datos de productos naturales. Los resultados se publicaron en:

1. **Chávez-Hernández AL**, Sánchez-Cruz N, Medina-Franco JL. A Fragment Library of Natural Products and Its Comparative Chemoinformatic Characterization. *Mol. Inform.* **2020**, *39*, 2000050. https://doi.org/10.1002/minf.202000050.
2. **Chávez-Hernández AL**, Sánchez-Cruz N, Medina-Franco JL. Fragment Library of Natural Products and Compound Databases for Drug Discovery. *Biomolecules* **2020**, *10*, 1518. https://doi.org/10.3390/biom10111518.

Utilizando los fragmentos moleculares derivados de productos naturales se planteó una metodología para el diseño *de novo*. Los resultados se publicaron en:

3. **Chávez-Hernández AL**, Juaréz-Mercado KE, Saldívar-González FI and Medina-Franco JL. Towards the de novo Design of HIV-1 Protease Inhibitors based on Natural Products. *Biomolecules*, **2021**, *11*, 1805. https://doi.org/10.3390/biom11121805.

Se realizó una revisión bibliográfica sobre el diseño *de novo,* utilizando algoritmos de inteligencia artificial, y el análisis de subconjuntos de bases de datos de productos naturales. Los resultados se publicaron en:

4. **Chávez-Hernández AL**, Medina-Franco JL. Natural Products Subsets: Generation and Characterization. *Artif Intell Life Sci*. **2023**, *3*, 100066, https://doi.org/10.1016/j.ailsci.2023.100066.
5. **Chávez-Hernández AL**, López-López E, Medina-Franco JL. Yin-yang in Drug Discovery: Rethinking de Novo Design and Development of Predictive Models. *Front Drug Disc* **2023**, 3, 1222655. https://doi.org/10.3389/fddsv.2023.1222655.

1.2. Artículos de difusión y divulgación

1. Medina-Franco JL* **Chávez-Hernández AL**, López-López E, Saldívar-González FI. Chemical Multiverse: An Expanded View of Chemical Space. *Mol. Inf.* **2022**, *41*, 2200116. https://doi.org/10.1002/minf.202200116.
2. Saldívar-González FI, **Chávez-Hernández AL**, Prado-Romero DL, González-Medina M. ¿Por Qué Hay Que Hablar De Mujeres En Química Computacional Y No solo De Química Computacional? CIENCIAUANL 2023, 26, 8-19. https://doi.org/10.29105/cienciauanl26.121-1
3. Bajorath J, **Chávez-Hernández AL**, Duran-Frigola M, Fernández-de Gortari E, Gasteiger J, López-López E, et al. Chemoinformatics and Artificial Intelligence Colloquium: Progress and Challenges in Developing Bioactive Compounds. *J. Cheminformatics*. **2022**, *14*, 1-12. https://doi.org/10.1186/s13321-022-00661-0.

1.3. Capítulo de libro

1. Medina-Franco JL* Flores-Padilla EA, **Chávez-Hernández AL**. Discovery and Development of Lead Compounds from Natural Sources using Computational Approaches. En Evidence-Based Validation of Herbal Medicine, Pulok Mukherjee, (Ed.) 2nd Ed. Elsevier 2022, pp. 539-560. https://doi.org/10.1016/B978-0-323-85542-6.00009-3.

1.4. Artículo de alumna de licenciatura

1. **Gaytán-Hernández D**, **Chávez-Hernández AL**, López-López E, Miranda-Salas J, Saldívar-González FI & Medina-Franco JL. Art driven by visual representations of chemical space. *J Cheminform*, **2023**, 15, 100. https://doi.org/10.1186/s13321-023-00770-4.

1.5. Mentorías y alumnos

1. Asesoramiento como profesora adjunta del proyecto: Representaciones visuales de espacios y multiversos químicos para la divulgación de conocimiento científico y

expresión artística. Alumna: **Daniela Gaytán Hernández**. Tesis de licenciatura para obtener el grado de Ingeniera Química, Facultad de Química de la UNAM, 2024.

2. Asesoramiento como profesora adjunta del proyecto: Desarrollo de flujos de trabajo aplicados al diseño de fármacos. Alumna: **Jessica Alejandra Román Palafox**. Programa de Estancias Cortas de Investigación, durante el intersemestral 2023-1, Facultad de Química, UNAM.

1.6. Colaboración en el desarrollo del manual de Quimioinformática en español

1. Saldívar-González FI, Prado-Romero DL, Cedillo-González BR, **Chávez-Hernández AL**, Avellaneda-Tamayo JF, Gómez-García A, Medina-Franco JL. A Spanish Chemoinformatics GitBook for Chemical Data retrieval and Analysis using Python Programming. *J. Chem. Educ.* **2024**. https://doi.org/10.1021/acs.jchemed.4c00041.

1.7. Cursos y talleres

1. Ayudante de profesor B en la materia de licenciatura “Introducción a la Quimioinformática” en la Facultad de Química de la UNAM [2023-2024].
2. UAMedia, curso: "Búsqueda, análisis, representación y visualización de información química contenida en bases de datos moleculares". Fecha: 9 al 20 de octubre del 2023.
3. Curso-Taller I (Quimioinformática), con una duración de 8 horas, impartido en el marco del IX Simposio Tendencias actuales en la búsqueda y desarrollo de fármacos, realizado los días 12 y 13 de junio de 2023.
4. Taller: Script de Python para calcular descriptores moleculares y una visualización del espacio químico con PCA. Ponencia del III CURSO TALLER DE BIOQUIMIOINFORMATICA - EL MUNDO DEL DISEÑO Y DESARROLLO DE FÁRMACOS". Evento académico realizado por la Asociación Peruana de Estudiantes de Farmacia y Bioquímica APEFYB-Perú. Fecha: 10 de septiembre 2022.

1.8. Apoyo para obtener financiamiento

1. Becario dentro del espacio de innovación **UNAM-HUAWE**I. Proyecto: Desarrollo y aplicación de algoritmos de inteligencia artificial para el diseño de fármacos aplicables al tratamiento de diabetes mellitus y cáncer. [**2022-2023**].

1.9. Presentación en congresos nacionales e internacionales

1. *Rethinking de novo drug design aided with natural product subsets. American Chemical Society.* Presentación oral. Fall Meeting 2023. San Francisco, California, EUA. Fecha: 16 de agosto de 2023.
2. Subconjuntos de bases de datos de productos naturales: Generación y caracterización. Póster 10 presentado en la 18a Reunión Internacional de Investigación en Productos Naturales organizado por el AMINOPRONAT. Morelia Michoacán, México. Fecha: 25 de mayo de 2023.
3. HANNA: Una huella digital molecular basada en productos naturales y quiralidad utilizando un algoritmo de inteligencia artificial. Póster 34 presentado en la XVII Reunión de la Academia Mexicana de Química Orgánica. Puebla, México. Fecha: 25 de agosto de 2022.
4. Webinar: "HANNA: una huella digital molecular basada en productos naturales utilizando una arquitectura de redes neuronales". Ponencia presentada en el III CURSO TALLER DE BIOQUIMIOINFORMATICA - EL MUNDO DEL DISEÑO Y DESARROLLO DE FÁRMACOS". Evento académico realizado por la Asociación Peruana de Estudiantes de Farmacia y Bioquímica APEFYB-Perú. Fecha: 10 de septiembre de 2022.
5. *Towards the de novo design of HIV-1 protease inhibitors based on natural products*. Presentación oral. XXVII Symposium on Bioinformatics and Computer-Aided Drug Discovery. Rusia. Fecha: 26 de mayo de 2022.
6. Fragment Library of Natural Products for Drug Discovery. Presentación oral. ACS Fall meeting. EUA. Fecha: 22 de agosto de 2021.

7. WORKSHOP ON SECONDARY METABOLITE DISCOVERY. Computational Applications in Secondary Metabolite Discovery. A Fragment Library of Natural Products and Compounds. Alemania. Fecha: 09 de marzo de 2021.
8. Webinar: “La importancia de generar bibliotecas de fragmentos de productos naturales en el desarrollo de fármacos”. Ponencia del Webinar Farmacéutico Internacional de Bioquimioinformática: “La ciencia del descubrimiento, diseño y desarrollo de fármacos”. Evento académico realizado por la Asociación Peruana de Estudiantes de Farmacia y Bioquímica APEFYB-Perú. Fecha: 18 de diciembre de 2020.
9. *A Fragment Library of Natural Products and its Comparative Chemoinformatics Characterization*. Póster 18 presentado en el primer congreso internacional de “*Women in Bioinformatics and Data Science, LA*”. Fecha: septiembre 2020.

1.10. Distinciones científicas y reconocimientos

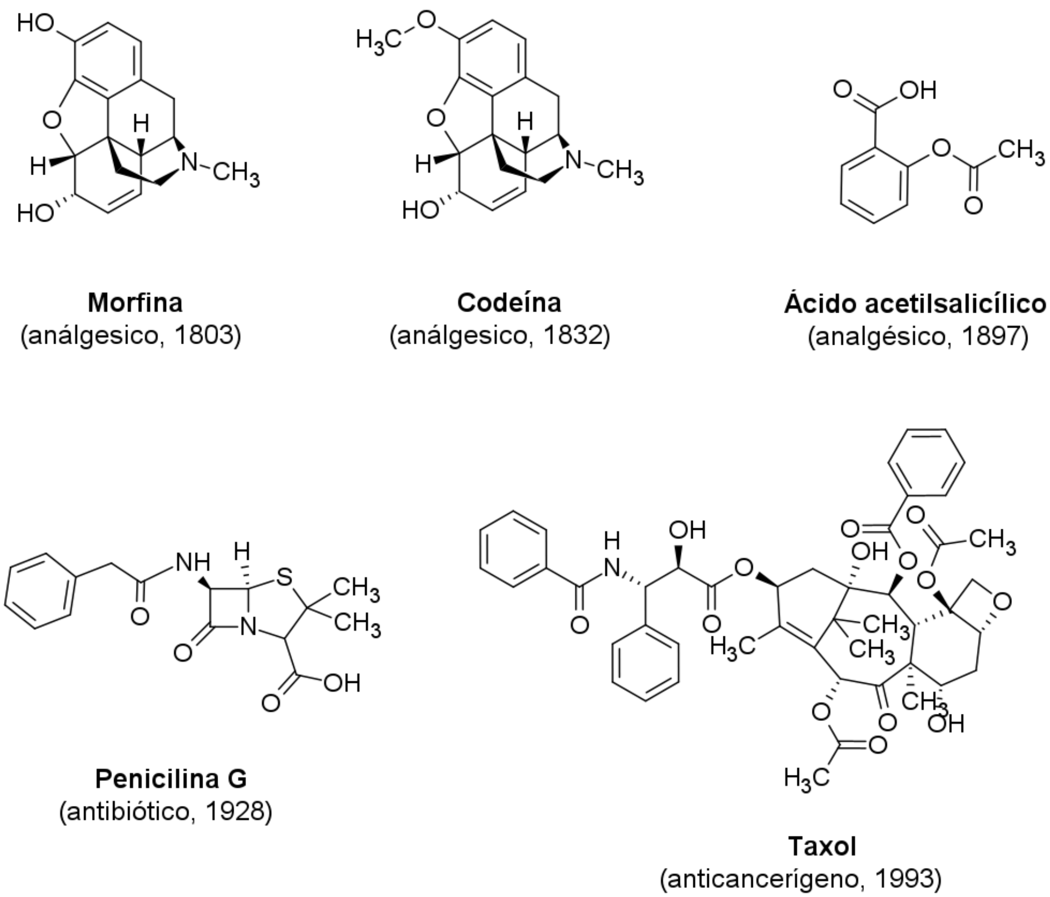
1. **Artículo más citado** 2020-2021. Wiley-Molecular Informatics. **Chávez-Hernández AL**, Sánchez-Cruz N and Medina-Franco JL. A Fragment Library of Natural Products and its Comparative Chemoinformatic Characterization. *Mol. Inf.* **2020**, *39*, 2000050. https://doi.org/10.1002/minf.202000050.
2. **Artículo más leído** 2021-2022**.** Medina-Franco JL* **Chávez-Hernández AL**, López-López E, Saldívar-González FI. Chemical Multiverse: An Expanded View of Chemical Space. *Mol. Inf.* **2022**, *41*, 2200116. https://doi.org/10.1002/minf.202200116.
3. **Artículo publicado en dos volúmenes especiales**: Women in Artificial Intelligence in the Life Sciences y AI in the Life Sciences by Latin Americans. **Chávez-Hernández AL**, Medina-Franco JL. Natural products subsets: Generation and characterization. *Artif Intell Life Sci.* **2023**, 100066.
4. Asesoramiento como profesora adjunta del proyecto: Desarrollo de flujos de trabajo aplicados al diseño de fármacos, **ganador del tercer lugar** en la categoría de Química Farmacéutico Biológica, del programa Estancias Cortas de Investigación, correspondiente al intersemestral **2023-1** de la Facultad de Química, UNAM.

5. *Rethinking de novo drug design aided with natural product subsets. American Chemical Society.* **Ponencia seleccionada** para el informe *Chemical Information and Computation 2023, ACS National Meetings. San Francisco, August 13-17,* ***2023***, organizado por la Dra. Wendy Anne Warr, pionera de la quimioinfomática. https://www.warr.com/morepubs.html#sanfran.

II. Antecedentes

Un producto natural es un compuesto químico sintetizado por un organismo vivo para adaptarse al ecosistema, protegerse contra los depredadores y atraer o advertir a otras especies.[1]

Los productos naturales han atravesado un proceso de adaptación en el organismo que los alberga, por lo tanto, son los ligandos adecuados para interaccionar con diversas dianas biológicas (llamado "*espacio biológico relevante"*).[2,3] En comparación con las moléculas obtenidas mediante síntesis, especialmente síntesis combinatoria, los productos naturales tienen mayor diversidad estructural y mayor número de centros quirales.[4] Por esta razón, los productos naturales han tenido relevancia desde el inicio de la era farmacéutica. La **Figura 1** muestra ejemplos de los primeros fármacos obtenidos a partir de productos naturales o derivados de estos, y entre paréntesis el efecto terapéutico y el año de descubrimiento, como son la morfina (analgésico, 1803),[5] la codeína (analgésico, 1832),[5] el ácido acetilsalicílico (analgésico, 1897),[6,7] la penicilina G o bencilpenicilina (una de varias presentaciones de la penicilina producida de modo natural y usada clínicamente como antibiótico, 1928),[8,9] y el taxol (anticancerígeno, 1993).[10]



Morfina
(análgesico, 1803)

Codeína
(análgesico, 1832)

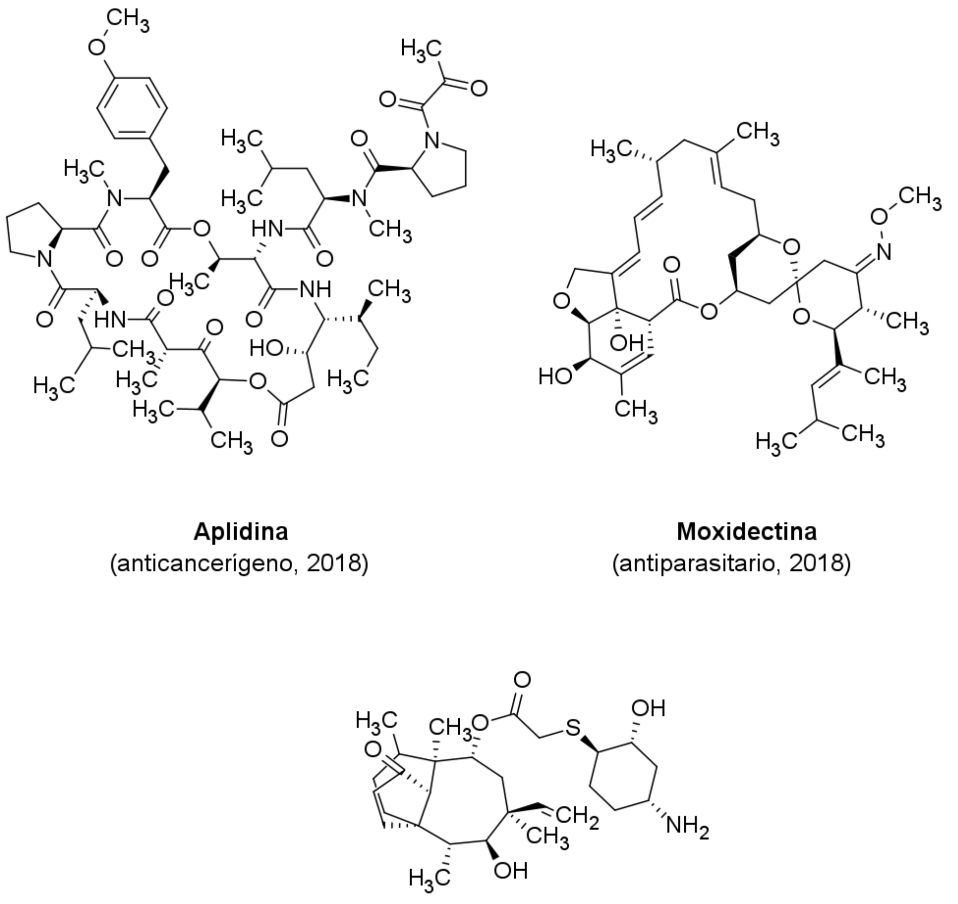
Ácido acetilsalicílico
(analgésico, 1897)

Penicilina G
(antibiótico, 1928)

Taxol
(anticancerígeno, 1993)

Figura 1. Ejemplos de los primeros fármacos obtenidos a partir de productos naturales o derivados de productos naturales. El efecto terapéutico y el año de descubrimiento se indican entre paréntesis.

Así mismo, de las 1,394 moléculas pequeñas aprobadas como fármacos entre 1981 y 2019, 5.1% son productos naturales, 31.2% derivados de productos naturales y 30.5% inspirados en productos naturales (es decir, compuestos que sustituyen al sustrato natural de una diana biológica terapéutica).[11,12] La **Figura 2** muestra la estructura química de la aplidina, moxidectina y lefamulina, fármacos derivados de productos naturales aprobados entre los años 2018 y 2019.[11] La aplidina[13] es un producto natural marino usado para el tratamiento del mieloma múltiple y fue aprobado en Australia en 2018.[11] La moxidectina[14] es un antiparasitario derivado de la nemadectina y aprobado por la FDA (por sus siglas en inglés, *Food and Drug Administration* de los Estados Unidos) en 2018. La lefamulina[15] es un antibiótico derivado de la pleuromutilina y aprobado por la FDA en 2019.



Aplidina
(anticancerígeno, 2018)

Moxidectina
(antiparasitario, 2018)

Lefamulina
(antibacteriano, 2019)

Figura 2. Aplidina, moxidectina y lefamulina, fármacos derivados de productos naturales aprobados entre los años 2017 y 2019. El efecto terapéutico y el año de aprobación se indican entre paréntesis.

Por estas razones, los productos naturales continúan siendo fundamentales en el diseño y desarrollo de nuevos fármacos. El diseño y desarrollo de fármacos tiene dos retos fundamentales que son encontrar moléculas que tengan actividad biológica (*hits*) y, de estas, seleccionar aquellas que tengan actividad biológica robusta y baja toxicidad (compuestos líderes o *leads*).[16] A pesar de que varias moléculas son candidatos a fármacos, una cantidad significativa de estas fallan por su toxicidad y sus propiedades farmacocinéticas inadecuadas. Un método para identificar moléculas con actividad biológica son las pruebas biológicas de alto rendimiento (HTS, por sus siglas del inglés *high-throughput screening*). Las pruebas biológicas de alto rendimiento consisten en el uso de equipos automatizados para evaluar la actividad biológica de miles o millones de compuestos rápidamente.

El diseño de fármacos asistido por computadora es otra estrategia general para identificar moléculas líderes.[17] Dentro de este, el diseño de fármacos basado en fragmentos (FBDD, por sus siglas del inglés *fragment-based drug discovery*)[18] se ha convertido en un método convencional. Por ejemplo, en 2018 se reportaron 26 moléculas bioactivas a partir del FBDD.[18] Los fragmentos moleculares son, generalmente, de peso molecular pequeño (< 300 Da) y tienen menor complejidad estructural que los compuestos estructuralmente similares a los fármacos. Por estas características, se sugiere que los fragmentos moleculares tienen mayor probabilidad de encajar en un sitio de unión y formar interacciones intermoleculares.[17,19] Gracias a la importancia creciente de esta estrategia, se han desarrollado bibliotecas moleculares de fragmentos. A la fecha, estas bibliotecas, primordialmente, se derivan de compuestos sintéticos que tienen una diversidad molecular acotada. Por esta razón es conveniente desarrollar bibliotecas de fragmentos con estructuras diversas que proporcionen una gama amplia de bloques de construcción que sirva para el diseño de nuevos compuestos bioactivos (diseño *de novo*).[20,21] Una alternativa son los productos naturales, ya que poseen mayor diversidad estructural que los compuestos de origen sintético.

2.1. Bases de datos de productos naturales

Una base de datos es una estructura organizada que almacena información, normalmente asociada a un programa computacional. Una finalidad de las bases de datos moleculares es actualizar, responder y recuperar datos almacenados en un sistema de manera eficiente.[22]

La información química almacenada depende del tipo de base de datos. La **Figura 3** muestra una clasificación propuesta[23] de las bases de datos de compuestos químicos que abarca seis categorías que son (1) compuestos bajo demanda, (2) compuestos con actividad biológica, (3) compuestos disponibles comercialmente, (4) bases de datos de productos naturales, (5) compuestos de referencia (en inglés, *benchmark*), y (6) compuestos de tipo señuelo (en inglés *decoy*) y compuestos inactivos.[23,24] Las bases de datos de compuestos bajo demanda tienen protocolos de síntesis química bien establecidos, lo cual facilita su disponibilidad y adquisición.[25,26] Las bases de datos de referencia son compuestos químicos que tienen protocolos de curado de bases de datos bien establecidos, y por ende se usan de referencia para construir modelos predictivos.[27–29] Los compuestos tipo señuelo son presuntamente inactivos contra una diana biológica, pero tienen propiedades fisicoquímicas muy similares a los compuestos biológicamente activos.[30]

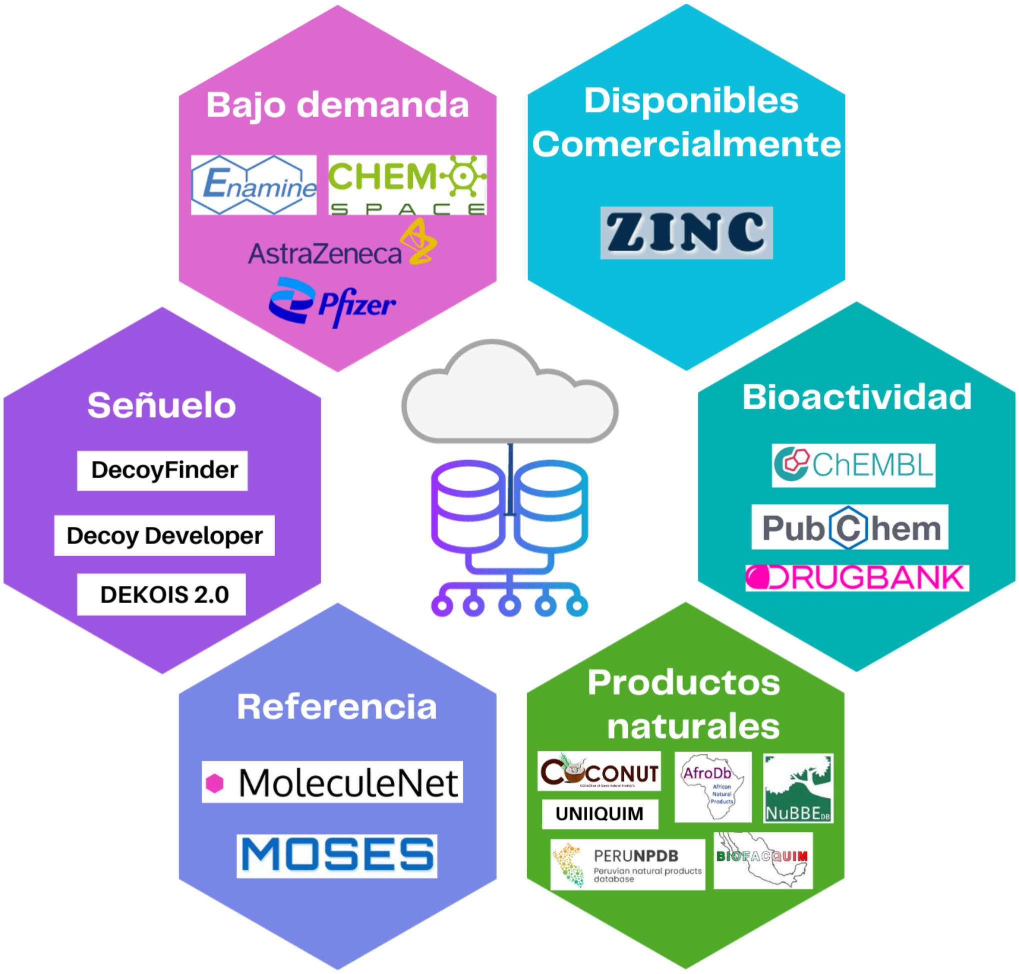


Figura 3. Clasificación de bases de datos de compuestos químicos. Fuente: **Chávez-Hernández AL**, López-López E, Medina-Franco JL. Yin-yang in drug discovery: rethinking de novo design and development of predictive models. *Front. Drug Discov.* **2023**, *3*, 1222655.

Una base de datos de compuestos químicos naturales usualmente contiene información química como la estructura química (ver **sección 2.2**), descriptores moleculares (ver **sección 2.3**), el organismo del que se aisló el compuesto químico y, cuando está disponible, la actividad biológica reportada contra una o más dianas biológicas o un ensayo biológico en general. Las bases de datos de productos naturales son importantes en el desarrollo y descubrimiento de nuevos fármacos[31,32] debido a la diversidad de estructuras químicas y núcleos base que varían del organismo del que provienen.[33] Algunos ejemplos de bases de datos de productos naturales con el mayor número de productos naturales reportados son SuperNatural 3.0[34] con 449,058 productos naturales y derivados, COCONUT[35] (por sus siglas en inglés, *Collection of Open NatUral ProdUcTs*) con 406,076 estructuras químicas únicas y UNPD[36] (por sus siglas en inglés, *Universal Natural Product Database*) con 197,201 estructuras químicas que contienen información sobre la quiralidad de los compuestos. También, las bases de datos de productos naturales pueden recopilar compuestos aislados y caracterizados de diferentes regiones geográficas como China, India, África y Latinoamérica. Por ejemplo, TCM (por sus siglas en inglés, *Chinese Traditional Medicine Database@Taiwan*)[37] es una base de datos de medicina tradicional china de acceso libre con más de 20,00 compuestos. IMPPAT (por sus siglas en inglés, *Indian Medicinal Plants, Phytochemistry and Therapeutics*)[38] contiene 9,596 compuestos fitoquímicos aislados de 1,742 plantas medicinales de la India. AfroDB[39] contienen más de 1,000 compuestos derivados de plantas medicinales de África. LANaPDB (por sus siglas en inglés, *Unified Latin American Natural Product Database*)[31,40] contiene 12,959 productos naturales de seis bases de datos de países de Latinoamérica como Bolivia, Brasil, Colombia, Costa Rica, Ecuador, México, Perú y Venezuela. Algunos ejemplos de bases de datos de productos naturales representativas de Latinoamérica son NuBBE$_{DB}$ (2,223 compuestos, Brasil),[41] SistemaX (9,514 compuestos, Brasil),[42] CIFPMA (354 compuestos, Panamá),[43,44] PeruNPDB (280 compuestos, Perú),[45] UNIIQUIM (1,112 compuestos, México)[48] y BIOFACQUIM (531 compuestos, México).[46,47] UNIIQUIM[48] (por su acrónimo en español, Unidad de Informática de Instituto de Química) es una base de datos de productos naturales creada por el Instituto de Química de la UNAM, compuesta de productos naturales de México y principalmente productos naturales aislados y caracterizados por el departamento de productos naturales del Instituto de Química. Mientras que BIOFACQUIM[46,47] es una base de datos de productos naturales desarrollada por el grupo de investigación Diseño de Fármacos Asistido por Computadora de la Facultad de Química de la UNAM que contienen productos naturales aislados y caracterizados por otros institutos de investigación de México, y cuyos compuestos son derivados de plantas, hongos y propóleo.

Las estructuras químicas almacenadas en las bases de datos de compuestos químicos son representadas utilizando representaciones moleculares.

2.2. Representaciones moleculares

Una representación molecular es cualquier forma de representar a un compuesto químico y, se caracterizan por codificar la conectividad entre pares de átomo de cada molécula.[49,50] Algunos ejemplos de representaciones moleculares son SMILES[51] (por sus siglas en inglés, *Simplified Molecular Input Line Entry System*), SMARTS[52] (por sus siglas en inglés, *arbitrary target specification*), InChI (por sus siglas en inglés, *International Chemical Identifier*) o InChIKey,[53] una representación condensada de un InChI[53] con 27 caracteres fijos que facilita la búsqueda de estructuras químicas.[50] Algunos modelos generativos de diseño *de novo* utilizan representaciones moleculares como SMIRKS,[54,55] una notación genérica que permite a un lenguaje de programación leer una reacción química, y los grafos moleculares.[56] La **Figura 4** muestra la molécula de piperazina representada como SMILES, SMARTS, InChI e InChIKey utilizando el protocolo descrito por Saldívar-González *et al.*[50] Los SMARTS, además de describir la conectividad entre pares de átomo de la molécula de piperazina, mapean los patrones de conectividad que el carbono (**Figura 4-A**) y el nitrógeno (**Figura 4-B y Figura 4-C**) pueden tener como sustituyentes un hidrógeno (**Figura 4-A**), dos hidrógenos (**Figura 4-B**) u otros átomos (**Figura 4-C**). El InChI incluye la fórmula empírica y la posición de los hidrógenos, y el InChIKey indica el tipo de carga, estereoquímica e isótopos.

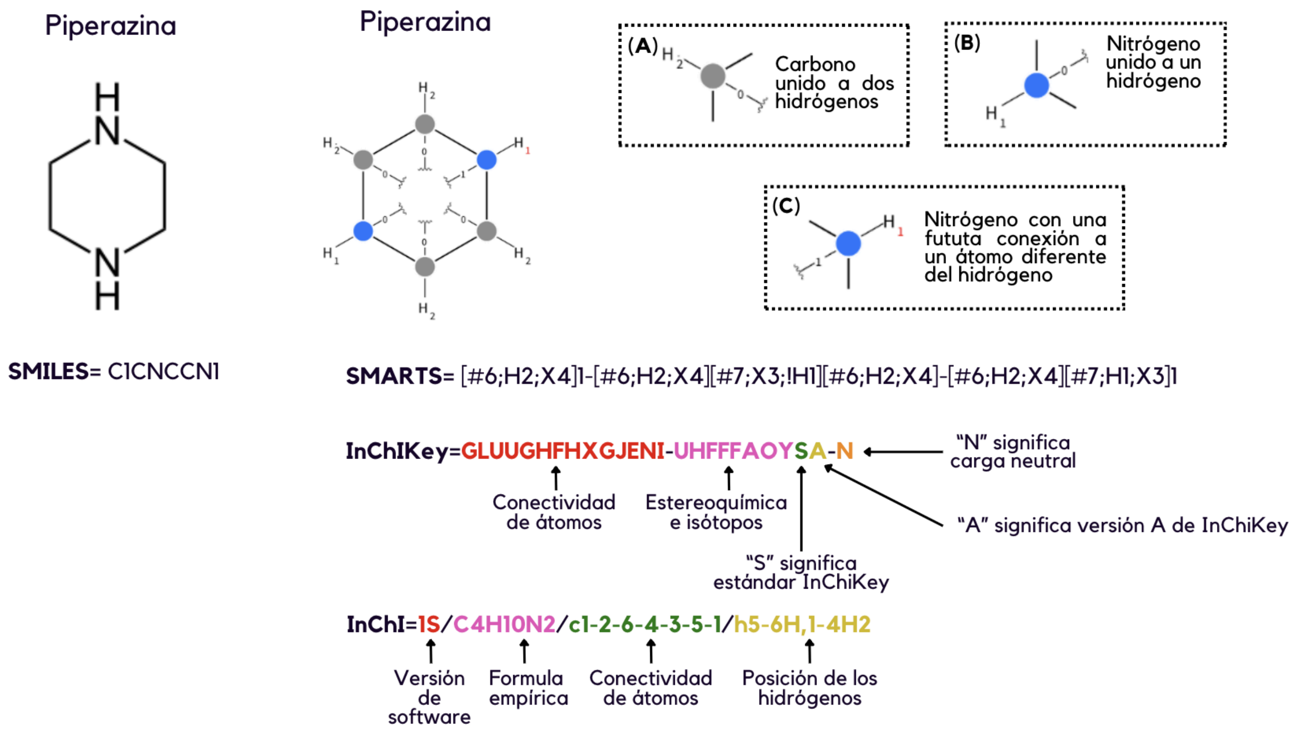

Piperazina
Piperazina
(A) Carbono unido a dos hidrógenos
(B) Nitrógeno unido a un hidrógeno
(C) Nitrógeno con una fututa conexión a un átomo diferente del hidrógeno
SMILES= C1CNCCN1
SMARTS= [#6;H2;X4]1-[#6;H2;X4][#7;X3;!H1][#6;H2;X4]-[#6;H2;X4][#7;H1;X3]1
InChIKey=GLUUGHFHXGJENI-UHFFFAOYSA-N
Conectividad de átomos
Estereoquímica e isótopos
"S" significa estándar InChiKey
"N" significa carga neutral
"A" significa versión A de InChiKey
InChI=1S/C4H10N2/c1-2-6-4-3-5-1/h5-6H,1-4H2
Versión de software
Formula empírica
Conectividad de átomos
Posición de los hidrógenos

Figura 4. Molécula de piperazina representada como SMILES, SMARTS, InChI e InChIKey utilizando el protocolo de Saldívar-González FI, Huerta-García CS, Medina-Franco JL. Chemoinformatics-Based Enumeration of Chemical Libraries: A Tutorial. *J. Cheminform.* **2020**, *12*, 64.

2.2.1. Grafos moleculares

Un grafo molecular permite codificar una molécula en un conjunto de nodos (átomos) y aristas (enlaces químicos).[49] En general, los conjuntos de átomos y enlaces se codifican dentro de dos matrices, una de características y otra de conectividad. La biblioteca Pytorch Geometric, implementada en el lenguaje de programación de Python, añade otra matriz de características para tener tres en total.[57,58] La **Figura 5** muestra un ejemplo de grafo molecular de la molécula R-2-butanol generado mediante Pytorch Geometric, y sus respectivas matrices de conectividad y características. La matriz de características atómicas codifica nueve descriptores moleculares de los átomos de cada molécula como son el número atómico, el tipo de quiralidad (R o S), el número de átomos vecinos, la carga formal, el número de hidrógenos implícitos y no implícitos, el número de electrones desapareados, el tipo de hibridación (SP3 o SP2) y si el átomo está dentro de un anillo aromático o alifático (0=No o 1=Si). La matriz de características de enlace químico describe el tipo de enlace (por ejemplo, 1=simple, 2=doble, 3=triple o 12=aromático), la estereoquímica, y si un enlace es conjugado o no. La matriz de conectividad codifica qué átomos están unidos mediante un enlace químico. Las **Tablas A1** y **A2** (ver Apéndice) resumen las características químicas de átomos y enlaces, y los valores que pueden tomar desglosados en la columna de descriptores moleculares.

Matriz de características de enlace químico (edge_attr)

(1)	(2)	(3)
1	0	0
1	0	0
1	0	0
1	0	0
1	0	0
1	0	0
1	0	0

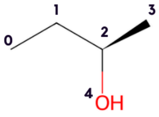
Matriz de características de enlace químico

1) Tipo de enlace (**1: enlace simple, 2: enlace doble**)

2) Estereoquímica del enlace (**0: no tiene estereoquímica**)

3) Enlace conjugado (**0: no tiene enlaces conjugados**)

R-2-butanol



C:0	C:1
C:1	C:0
C:1	C:2
C:2	C:1
C:2	C:3
C:2	O:4
C:3	C:2
O:4	C:2

Matriz de conectividad (edge_index)

Matriz de características atómicas (x)

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
6	0	4	5	3	0	4	0	0
6	0	4	5	2	0	4	0	0
6	1	4	5	1	0	4	0	0
6	0	4	5	3	0	4	0	0
8	0	2	5	1	0	4	0	0

Matriz de características atómicas

(1) Número atómico **6:C, 8:O**

(2) Quiralidad **2:S, 1:R, 0: No**

(3) El grado de un átomo o número de átomos vecinos **O=2, C=4**

(4) Carga formal **5=CERO**

(5) Número de hidrógenos **O=1, C=3, 2, 1**

(6) Número de electrones desapareados **CERO**

(7) Tipo de hibridación **4=SP3**

(8) El átomo está dentro de un anillo aromático **Sí=1, No=0**

(9) El átomo está dentro de un anillo **Sí=1, No=0**

Figura 5. Grafo molecular de la molécula R-2-butanol. Se muestran dos matrices de características, una de características atómicas y otra de características de enlaces químicos. La matriz de características atómicas codifica nueve descriptores moleculares que son el número atómico, quiralidad (R o S), átomos vecinos, carga formal, número de hidrógenos, número de electrones desapareados, tipo de hibridación, si el átomo está dentro de un anillo aromático o alifático. En el centro se muestra la matriz de conectividad con los átomos de la molécula que están unidos mediante un enlace químico. Fuente: https://github.com/DIFACQUIM/HANNA.

2.3. Descriptores moleculares

Las representaciones moleculares descritas anteriormente permiten reconstruir la molécula, pero existen otras notaciones que no lo hacen, y son los descriptores moleculares.[49] Un descriptor molecular es un número que codifica propiedades fisicoquímicas, estructurales, topológicas y electrónicas de una molécula.[49,59]

Los descriptores moleculares constitucionales son los más sencillos y utilizados, ya que reflejan la composición molecular de un compuesto sin ninguna información sobre su geometría molecular.[60]

Algunos ejemplos son el número de átomos de carbono, nitrógeno y oxígeno, el número de anillos aromáticos y alifáticos.

Los descriptores moleculares de complejidad estructural describen la tridimensionalidad de las moléculas, algunos ejemplos son la fracción de átomos con hibridación sp^3 (Fsp^3) y la fracción de centros quirales (FCC por sus siglas en inglés, f*raction of chiral centers*).[61] Fsp^3 es un descriptor molecular de complejidad estructural porque la saturación permite la preparación de moléculas más complejas con mayor tridimensionalidad. Por ejemplo, si átomo de carbono tiene cuatro enlaces simples, es átomo de carbono tendrá una hibridación sp^3 y, por lo tanto, la molécula tendrá una geometría tetraédrica, es decir, una molécula con geometría tridimensional. Mientras que FCC indica la posibilidad de encontrar un número de moléculas únicas con la misma fórmula y peso molecular.[61] Otro ejemplo común es el peso molecular, que es asociado intuitivamente a estructuras químicas complejas.[61,62]

2.4. Huellas digitales moleculares

Una huella digital molecular es un tipo de descriptor molecular que convierte una estructura molecular en una secuencia de bits, cada bit toma en cuenta la presencia o ausencia de una característica molecular.[63] Una huella digital molecular permite describir la similitud molecular entre moléculas a partir del número de bits en común.[63,64] La **Figura 6** muestra tres tipos de huellas digitales moleculares principales que son basadas en claves estructurales, topológicas o basadas en rutas y circulares.[65]

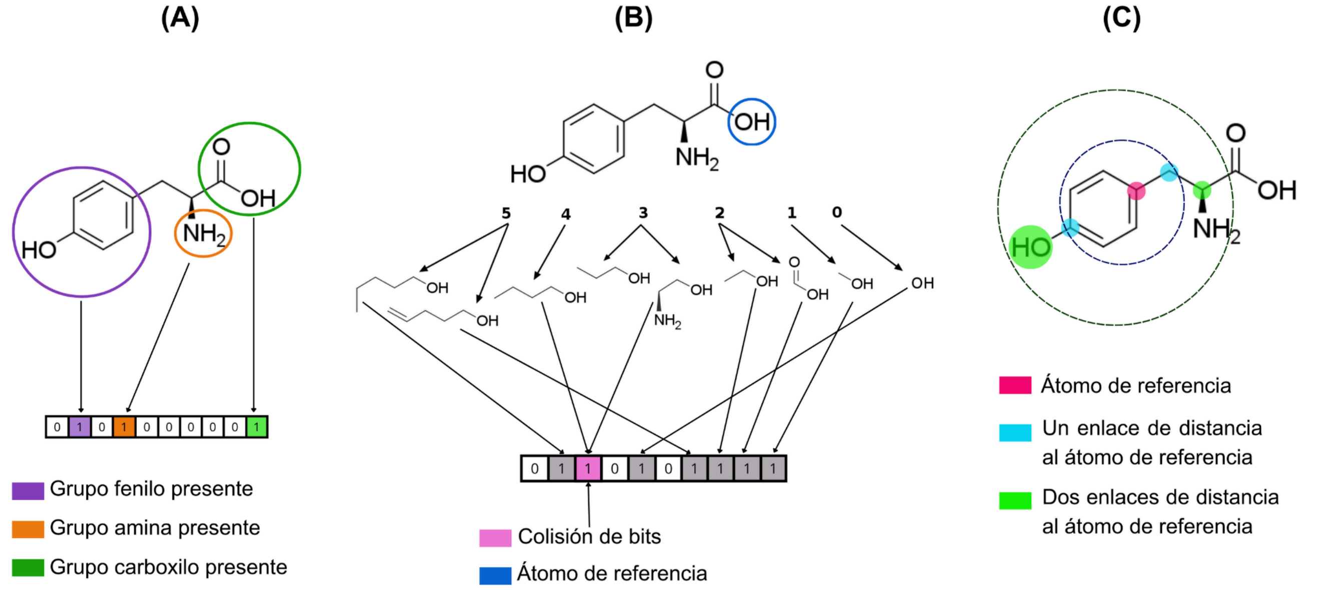


Figura 6. Principales tipos de huellas digitales moleculares. A) Huella digital molecular basada en claves estructurales, B) Huella digital molecular topológica o *hashed* y C) Huella digital molecular circular.

Las huellas digitales moleculares basadas en claves estructurales son cadenas de bits que establecen la presencia de subestructuras o características en una molécula de una lista de claves estructurales dada (**Figura 6A**).[65] El número de bits está determinado por el número de características estructurales, y cada bit está relacionado con la presencia o ausencia de una característica determinada de la molécula.[65] Un ejemplo es MACCS (por sus siglas en inglés, *Molecular ACCess System*) *keys* con 166 y 960 claves estructurales (bits),[66] y PubChem *fingerprint* con 188-bits.[67]

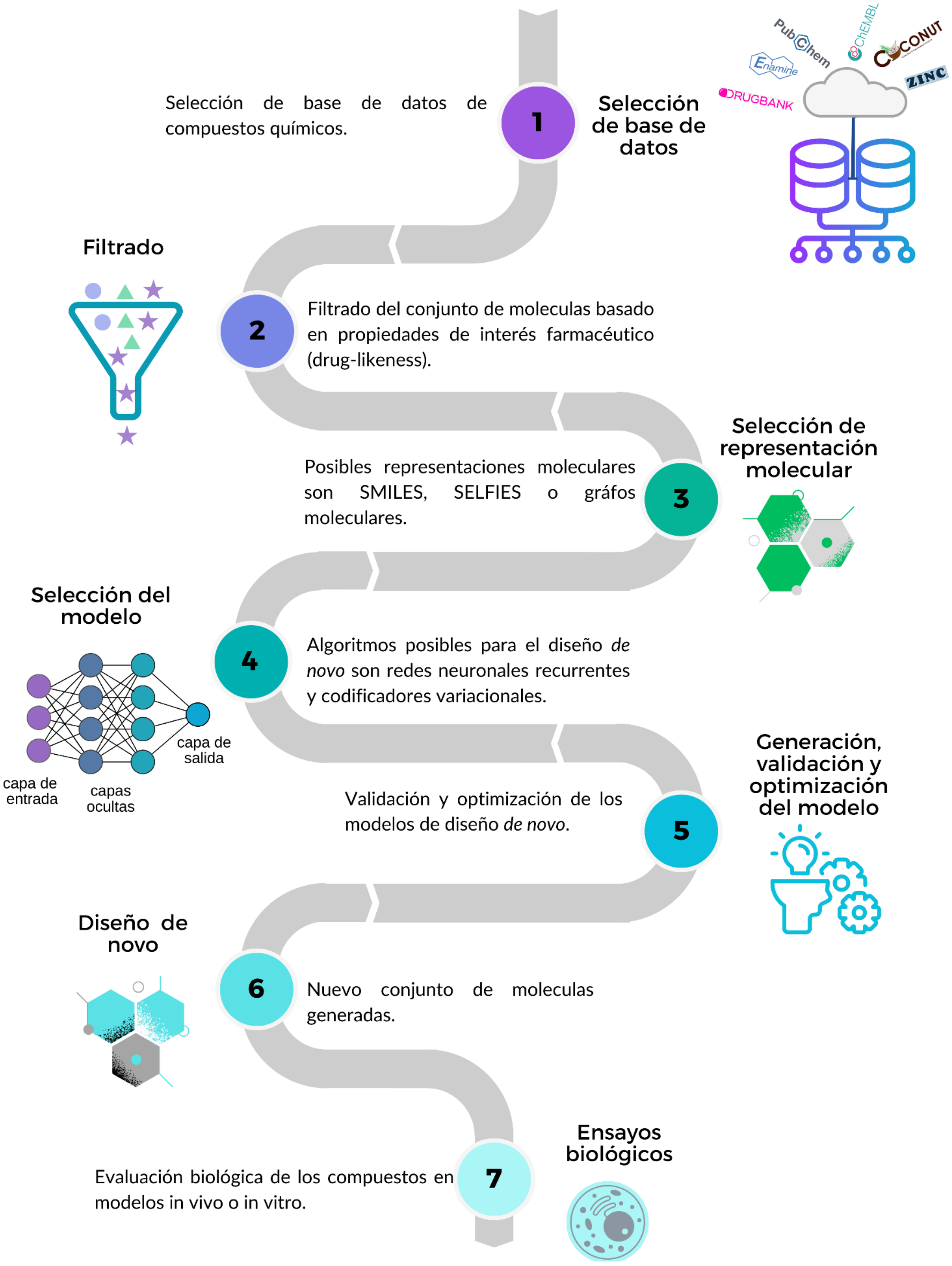
Las huellas digitales moleculares topológicas o basadas en rutas analizan todos los fragmentos de una molécula siguiendo una ruta (usualmente lineal) a cierto número de enlaces y luego realizan un “hashing” (término en inglés que significa convertir uno o varios elementos de entrada en otro elemento) de cada una de estas rutas para crear una huella digital molecular o huella digital molecular *hashed* (**Figura 6B**).[65] Cada molécula puede producir una huella digital molecular significativa, y su longitud puede ser ajustada, aunque a veces puede producirse una “colisión de bits”, es decir, un bit puede ser representado por uno o más características estructurales.[65] Usualmente, este tipo de huella digital molecular es usada para un rápido filtrado o búsqueda por subestructura. Algunos ejemplos son *Daylight fingerprint*[68,69] que codifica todas las rutas de conectividad posibles a través de una molécula con una longitud dada, y *Tree* del programa computacional *OpenEye*[70] cuyas rutas de conectividad no son lineales.

Las huellas digitales moleculares circulares son una variante de las huellas digitales moleculares tipo *hashed*, pero toman en cuenta el ambiente químico a un determinado número de enlaces radiales, en lugar de basarse en rutas (**Figura 6C**).[65] Son ampliamente usadas para la búsqueda de similitud estructural.[65] Algunos ejemplos son *Molprint2D* que codifica los entornos atómicos de cada átomo en una tabla de conectividad molecular, que están representados por cadenas de tamaño variable.[71] ECFP (por sus siglas en inglés, *Extended Connectivity fingerprint*)[72] basado en el algoritmo de Morgan.[73] ECFP representan átomos vecinos circulares y producen huellas digitales moleculares de una longitud variada y la más usada comúnmente es de diámetro 4, es decir, ECFP4.[65] Ejemplos de huellas digitales moleculares similares a ECFP son la huella digital molecular *MinHash* a seis enlaces de distancia (MHFP6)[74] y la huella digital molecular de pares de átomos *minHAshed* a cuatro enlaces de distancia (MAP4).[75]

Otros tipos de huellas digitales moleculares que no entran en las categorías anteriores son las huellas digitales moleculares farmacóforicas, las huellas digitales moleculares basadas en interacciones moleculares y las huellas digitales moleculares basadas en inteligencia artificial. Un farmacóforo representa las características relevantes e interacciones necesarias para que una molécula sea activa contra una diana biológica.[65] Las huellas digitales moleculares farmacóforicas generalmente codifican la información de las características en una lista, la cual representa a la molécula, similar a una huella digital molecular basada en claves estructurales, pero tomando en cuenta la distancia entre estas características.[65] Las huellas digitales moleculares basadas en interacciones moleculares codifican información sobre interacciones proteína-ligando como enlaces de hidrógeno, interacciones únicas y contacto con su residuo de origen, un ejemplo es SIFt (por sus siglas en inglés, *Structural Interaction Fingerprint*).[76] Las huellas digitales moleculares basadas en inteligencia artificial como E3FP (por sus siglas en inglés, *Extended 3-Dimensional FingerPrint*)[77] captura los posibles confórmeros que puede adoptar una molécula en un disolvente.

Algunas aplicaciones de las huellas digitales moleculares en quimioinformática (disciplina que aplica métodos informáticos en la resolución de problemas de la química),[78] incluyendo el diseño de fármacos asistido por computadora (CADD, por sus siglas en inglés, *Computer Aided Drug Discovery*) son búsqueda por similitud estructural,[65] generación de modelos predictivos de relaciones cuantitativas estructura-actividad (QSAR por sus siglas del inglés, *Quantitative Structure Relationships*)[79,80] y filtrado de los compuestos de bases de datos que son utilizados en los primeros pasos del diseño de nuevos compuestos químicos desde cero (diseño *de novo*).[23,81] Por esta razón, se continúan desarrollando nuevas huellas digitales moleculares en diferentes tipos y complejidad.[63]

La **Figura 7** ilustra los pasos de una revisión bibliográfica sobre el diseño *de novo* basado en ligando y utilizando algoritmos de inteligencia artificial.[23] Los pasos son (1) Selección de una base de datos. (2) Filtrado de la base de datos con propiedades deseadas como tipo fármaco. En este ejemplo, los compuestos representados como estrellas cumplen propiedades tipo fármaco como peso molecular (MW) $\leq$500, donadores de puente de hidrógeno (HBD) $\leq$5, aceptores de puente de hidrógeno (HBA) $\leq$10, coeficiente de partición octanol/agua (log P) $\leq$5, enlaces rotables (RB) $\leq$10 y área topológica superficial (TPSA) $\leq$150. (3) Elección de la representación molecular que puede ser SMILES,[51] SELFIES[82] o grafos moleculares[49] (ver **Figura 5**). (4) Selección del algoritmo para el diseño *de novo*. (5) Desarrollo, validación y optimización del modelo. (6) Generación de las moléculas a partir del modelo de diseño *de novo*. (7) Evaluación de la actividad biológica de los compuestos. Cabe mencionar que existen otros enfoques para seleccionar compuestos, mediante el uso de huellas digitales moleculares.



1 Selección de base de datos

Selección de base de datos de compuestos químicos.

2 Filtrado

Filtrado del conjunto de moleculas basado en propiedades de interés farmacéutico (drug-likeness).

3 Selección de representación molecular

Posibles representaciones moleculares son SMILES, SELFIES o gráfos moleculares.

4 Selección del modelo

Algoritmos posibles para el diseño *de novo* son redes neuronales recurrentes y codificadores variacionales.

5 Generación, validación y optimización del modelo

Validación y optimización de los modelos de diseño *de novo*.

6 Diseño de novo

Nuevo conjunto de moleculas generadas.

7 Ensayos biológicos

Evaluación biológica de los compuestos en modelos in vivo o in vitro.

Figura 7. Resumen del diseño *de novo* basado en ligando utilizando algoritmos de inteligencia artificial: 1) Selección de una base de datos de compuestos químicos. 2) Filtrado de compuestos con propiedades fisicoquímicas tipo fármaco (MW ≤500, HBD ≤5, HBA ≤10, log P ≤5, RB ≤10 y TPSA ≤150). 3) Elección de la representación molecular. 4) Selección del algoritmo para el diseño *de novo*. 5) Desarrollo, validación y optimización del modelo. 6) Generación de las moléculas a partir del modelo de diseño *de novo*. 7) Evaluación de la actividad biológica de los compuestos generados. Fuente: **Chávez-Hernández AL**, López-López E, Medina-Franco JL. Yin-yang in drug discovery: rethinking de novo design and development of predictive models. *Front. Drug Discov.* **2023**, *3*,1222655.

Tomando en cuenta el diagrama de la **Figura 7**, las huellas digitales moleculares se pueden utilizar para filtrar compuestos basados en similitud molecular (en las primeras etapas del diseño *de novo*). También, se puede adaptar los pasos 1,3,4 y 5 del esquema al desarrollo de otros algoritmos de aprendizaje profundo (en inglés, *deep learning*) como redes neuronales. Dentro de este último se encuentran los autocodificadores variacionales (en inglés *variational autoencoders,* VAE) que son algoritmos generativos,[83] y su variante, autocodificadores variacionales gráficos (en inglés *graph variational autoencoder*, GVAE).

2.5. Espacio químico

El espacio químico puede definirse como un arreglo multidimensional de compuestos químicos, en el cual cada fila corresponde a una estructura química y, las columnas o dimensiones corresponden a *n* número de descriptores moleculares utilizados para crear ese espacio químico.[84] La **Figura 8** muestra un ejemplo gráfico del espacio químico.

El espacio químico es útil en el análisis de diversidad y de relaciones estructura actividad (en inglés, *Structure Activity Relationships*, SAR) y relaciones estructura propiedad (en inglés, *Structure Property Relationshis*, SPR).[85] En 2022, se amplió el concepto de espacio químico y se propuso el concepto de multiverso químico definido como “un conjunto de todos los espacios químicos posibles para una sola base de datos, en función al conjunto de descriptores moleculares utilizados”.[86] La visualización de los espacios o multiversos químicos requiere de métodos de visualización de datos. Los más utilizados son el análisis de componentes principales (en inglés, *principal component analysis*, PCA)[87] y la incrustación de vecinos estocásticos distribuidos en t (en inglés, *T-distributed stochastic neighbor embedding*, t-SNE).[88]

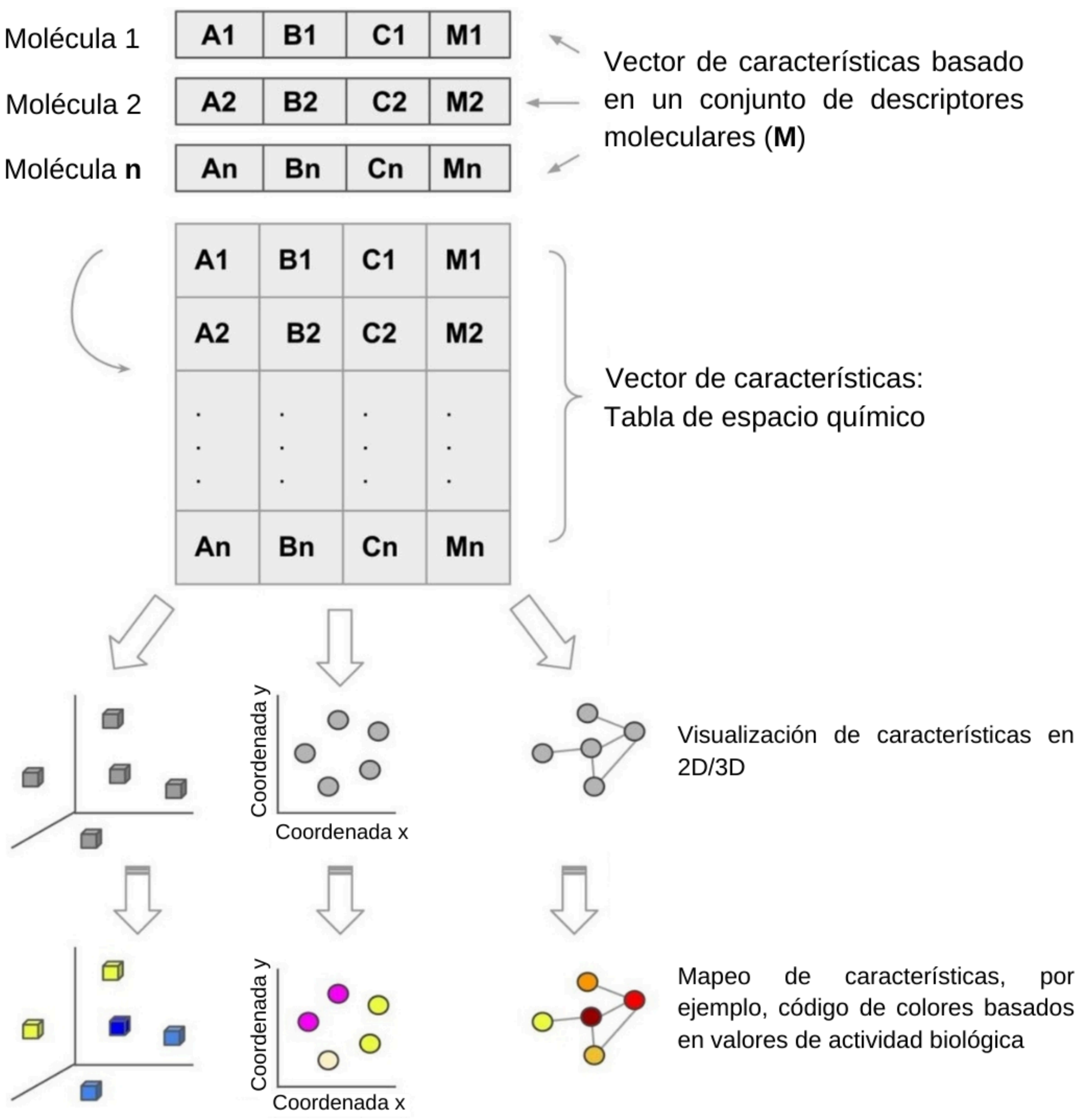

Molécula 1
A1
B1
C1
M1
Molécula 2
A2
B2
C2
M2
Molécula n
An
Bn
Cn
Mn
Vector de características basado en un conjunto de descriptores moleculares (M)
A1
B1
C1
M1
A2
B2
C2
M2
An
Bn
Cn
Mn
Vector de características:
Tabla de espacio químico
Coordenada y
Coordenada x
Visualización de características en 2D/3D
Coordenada y
Coordenada x
Mapeo de características, por ejemplo, código de colores basados en valores de actividad biológica

Figura 8. Espacio químico. Cada molécula (*n*) en una base de datos es representada por M descriptores moleculares. La tabla con el conjunto de moléculas y descriptores moleculares es el espacio químico y puede ser representado con diferentes métodos de visualización. Fuente: Medina-Franco JL, **Chávez-Hernández AL**, López-López E, Saldívar-González FI. Chemical Multiverse: An Expanded View of Chemical Space. *Mol Inform.* **2022**, *41*, e2200116.

III. Problemática y significancia

Este proyecto aborda dos problemas principales. Uno de ellos es la necesidad de generar bibliotecas de fragmentos con mayor diversidad estructural, las cuales puedan servir como bloques de construcción en el diseño *de novo* de moléculas bioactivas y novedosas. El segundo es el diseño de fármacos basados en productos naturales que ha disminuido gradualmente en las industrias farmacéuticas. Esto se ha debido, generalmente, a que se obtienen en pocas cantidades durante los procedimientos de obtención y purificación.[11] Además, estos procesos son más largos y costosos que aquellos obtenidos mediante síntesis combinatoria,[89] y la síntesis total de productos naturales, también, representa un reto por la alta complejidad molecular que tienen algunas estructuras. Por ende, para aprovechar al máximo los productos naturales que ya se han aislado y caracterizado previamente; y partiendo de la importancia del FBDD, el **objetivo** de este proyecto es generar bibliotecas públicas de fragmentos de productos naturales, que puedan ser utilizadas como punto de partida para la síntesis de los llamados *pseudo productos naturales*. La síntesis de pseudo productos naturales es combinar fragmentos de productos naturales de una manera que aún no se ha observado en la naturaleza.[90] Se ha propuesto que la síntesis de pseudo productos naturales coadyuva en la búsqueda de nuevas moléculas bioactivas y aumenta la probabilidad de éxito de estas como fármacos.[91] Un ejemplo reciente fue el descubrimiento de la glupina, a partir de 63 indomorfanos, que disminuye el crecimiento de varias líneas de células cancerosas, entre ellas cáncer de mama.[92] Así mismo, este proyecto busca impulsar el uso de los productos naturales en la industria farmacéutica y otros centros de investigación, por ende, se plantea generar una huella digital molecular basada en productos naturales para caracterizar su diversidad estructural. Además, este proyecto apoya la búsqueda y descubrimiento de nuevos productos naturales, ya que estos enriquecerían significativamente el desarrollo de la huella digital molecular y las bibliotecas de fragmentos, ambas, basadas en productos naturales.

IV. Hipótesis

Los fragmentos derivados de productos naturales conservan las características estructurales de los compuestos originales y cubren regiones del espacio químico diferente a la de los compuestos accesibles sintéticamente y compuestos con actividad biológica.

V. Objetivo general

Desarrollar, analizar y hacer públicas las bibliotecas de fragmentos obtenidos a partir de productos naturales.

VI. Objetivos particulares

1. Generar bibliotecas de fragmentos de productos naturales a partir de las siguientes bases de datos moleculares de acceso libre como son *the Collection of open natural products* (COCONUT) y Food Database (FooDB).
2. Generar bibliotecas de fragmentos de bases de datos de referencia como son la base de datos de REAL (comercializada por Enamine) con compuestos disponibles sintéticamente, compuestos con actividad biológica (ChEMBL), compuestos sin actividad biológica de DCM (por sus siglas en inglés, *Dark Chemical Matter*) pero que recientemente han llevado a la identificación de compuestos bioactivos, una biblioteca de *Chemical Abstract Service* (CAS) enfocada en COVID-19 y una biblioteca de inhibidores de la principal proteasa de SARS-CoV-2 (3CLP).
3. Comparar cuantitativamente las características estructurales, la complejidad estructural y la diversidad estructural de las bibliotecas de compuestos y fragmentos generados derivados de productos naturales y compuestos de referencia.
4. Utilizar los fragmentos moleculares derivados de la base de datos de COCONUT y los fragmentos moleculares comerciales de ChemDiv y Enamine para generar bibliotecas de compuestos análogos a bevirimat (un compuesto inhibidor de la proteasa viral del VIH-1).

5. Comparar el espacio químico y las propiedades fisicoquímicas de interés farmacéutico de las bibliotecas de los compuestos análogos a bevirimat generados mediante fragmentos derivados de productos naturales.
6. Generar tres subconjuntos de productos naturales de la base de datos *Universal Natural Product Database* (UNPD) utilizando el algoritmo MaxMin.

VII. Metodología

La **Figura 9** resume esquemáticamente la estrategia general para alcanzar los objetivos específicos y el objetivo general. La estrategia se desglosa en cuatro pasos (1) curado de las bases de datos, (2) generación de fragmentos moleculares, (3) análisis comparativo de las bases de datos de productos naturales y (4) un caso de estudio de diseño *de novo*. La metodología se detalla en las secciones siguientes.

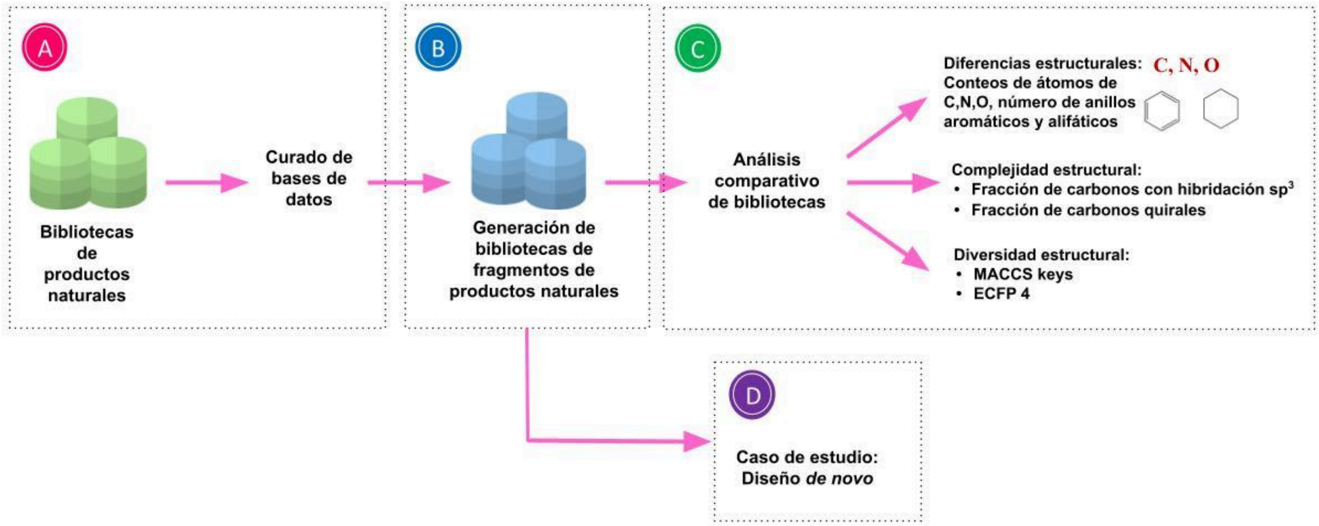

A
Bibliotecas de productos naturales
Curado de bases de datos
B
Generación de bibliotecas de fragmentos de productos naturales
C
Análisis comparativo de bibliotecas
Diferencias estructurales: C, N, O
Conteos de átomos de C,N,O, número de anillos aromáticos y alifáticos
Complejidad estructural:
• Fracción de carbonos con hibridación sp3
• Fracción de carbonos quirales
Diversidad estructural:
• MACCS keys
• ECFP 4
D
Caso de estudio: Diseño de novo

Figura 9. Metodología para desarrollar y analizar bibliotecas de fragmentos basados en productos naturales.

7.1. Curado de bases de datos

El curado de base de datos de compuestos químicos es importante porque permite verificar la exactitud, consistencia, y reproducibilidad de los datos experimentales reportados, lo cual es crucial para generar datos confiables y reproducibles.[93] La **Figura 10** muestra el proceso de curado de bases de datos moleculares que se realizó en este proyecto**.** Los compuestos de las bases de datos son representados mediante cadenas de SMILES[94] (del inglés, *Simplified Molecular Input Line Entry System*). No se incluyó información sobre la estereoquímica porque esta información no está incluida en la mayoría de las bases de datos moleculares públicas. El proceso de preparación se realizó utilizando el protocolo de Sanchez-Cruz *et al.,*[47] RDKit,[95] una biblioteca de software libre para quimioinformática disponible para el lenguaje de programación Python, y las funciones *Standardizer*, *LargestFragmentChoser*, *Uncharger*, *Reionizer*, *TautomerCanonicalizer* implementadas en la biblioteca MolVS.[96] El primer paso es la estandarización de los compuestos que consiste en eliminar los compuestos con errores de valencia y átomos diferentes a los elementos presentes en moléculas orgánicas (H, B, C, N, O, F, Si, P, S, Cl, Se, Br y I), y generar SMILES canónicos a partir de SMILES isoméricos. Los compuestos con múltiples componentes (por ejemplo, compuestos iónicos) fueron separados y el fragmento más grande (el ion más grande) fue retenido. De los compuestos remanentes, se eliminó la información estereoquímica, se neutralizaron las cargas, y se eliminaron los compuestos repetidos.

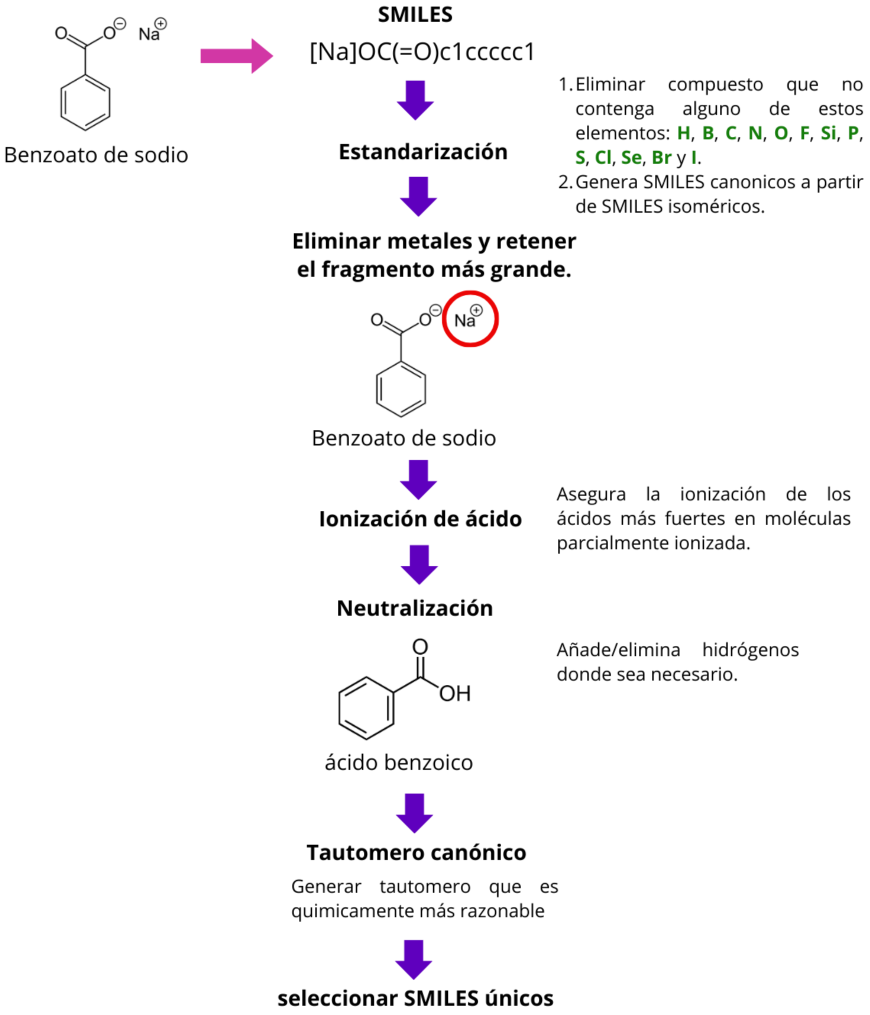

Benzoato de sodio
SMILES
[Na]OC(=O)c1ccccc1
Estandarización
1. Eliminar compuesto que no contenga alguno de estos elementos: H, B, C, N, O, F, Si, P, S, Cl, Se, Br y I.
2. Genera SMILES canonicos a partir de SMILES isoméricos.
Eliminar metales y retener el fragmento más grande.
Benzoato de sodio
Ionización de ácido
Asegura la ionización de los ácidos más fuertes en moléculas parcialmente ionizada.
Neutralización
Añade/elimina hidrógenos donde sea necesario.
ácido benzoico
Tautomero canónico
Generar tautomero que es quimicamente más razonable
seleccionar SMILES únicos

Figura 10. Curado de bases de datos moleculares. Fuente: Saldívar-González FI, Prado-Romero DL, Cedillo-González BR, **Chávez-Hernández AL**, Avellaneda-Tamayo JF, Gómez-García A, Medina-Franco JL. A Spanish Chemoinformatics GitBook for Chemical Data retrieval and Analysis using Python Programming. *J. Chem. Educ.* **2024**. https://doi.org/10.1021/acs.jchemed.4c00041.

7.2. Análisis de diversidad y complejidad estructural

El análisis de diversidad estructural de los compuestos químicos y fragmentos estructurales se divide en tres categorías de descriptores moleculares (1) diferencias estructurales, (2) complejidad estructural y (3) diversidad estructural.

7.2.1. Cálculo de descriptores moleculares

Las diferencias estructurales entre compuestos y fragmentos fueron evaluadas calculando diez descriptores moleculares utilizando como métrica la media de distribución. Los descriptores moleculares calculados fueron el número de átomos de carbono, nitrógeno y oxígeno; el número de átomos pesados, espiro y cabeza de puente; el número de anillos alifáticos y aromáticos, y el número de heterociclos alifáticos y aromáticos,

La complejidad estructural fue evaluada mediante el cálculo de la fracción de carbonos con hibridación sp^3 y la fracción de átomos de carbono quirales.[61,62]

La diversidad estructural de los compuestos y fragmentos fue evaluada mediante el cálculo de dos huellas digitales moleculares: *Extended-connectivity fingerprints* de radio 2 (ECFP-4) y 1024-bits,[72] y *Molecular ACCes System* (MACCS) keys (166-bits).[66] Posteriormente, se calculó la mediana de la distribución de los valores de pares de similitud generados utilizando el coeficiente de Tanimoto,[97] **Ecuación 1.**

$$T = \frac{N_C}{N_A + N_B - N_C}$$

Ecuación 1. Coeficiente de Tanimoto (T). Elementos en el conjunto A (N_A). Elementos en el conjunto B (N_B). Elementos en común entre el conjunto A y B (N_C).

También se calcularon seis propiedades fisicoquímicas de interés farmacéutico para generar visualizaciones de espacio químico utilizando PCA y t-SNE, y para caracterizar los subconjuntos de productos naturales generados en la **Sección 8.3**. Las propiedades fisicoquímicas calculadas fueron peso molecular (MW), donadores de puente de hidrógeno (HBD), aceptores de puente de hidrógeno

(HBA), coeficiente de partición octanol/agua (log P), enlaces rotables (RB) y área topológica superficial (TPSA).

7.2.2. Visualizaciones de espacio químico (PCA, t-SNE y TMAP)

Se realizaron visualizaciones del espacio químico y multiverso químico utilizando tres métodos de reducción de dimensiones y de visualización: PCA, t-SNE y Tree MAP (TMAP). Las visualizaciones de espacio químico utilizando PCA y t-SNE fueron generados usando seis propiedades de interés farmacéutico (MW, HBD, HBA, log P, RB y TPSA) como descriptores moleculares. Las visualizaciones de espacio químico utilizando TMAP fueron generadas usando como descriptor molecular la huella digital molecular ECFP-4 de 1024-bits.

PCA[87] es un método de reducción de dimensiones lineal que reduce los datos con muchas dimensiones en dos o tres nuevas variables llamadas componentes principales.[87] PCA conserva la mayor relación posible entre los datos. t-SNE es un método de reducción de dimensiones no lineal que genera gráficos que organizan los compuestos en subconjuntos de datos.[88] Compuestos similares forman grupos y compuestos disimilares que son distantes uno entre otros. TMAP es un método que permite visualizar muchos compuestos químicos a través de la distancia entre subconjuntos de compuestos.[98] La función HASH sensible a la localidad (en inglés *Local Sensitive Hashing, LSH*) es una función matemática que permite agrupar jerárquicamente las estructuras químicas en común. La cercanía se mide a partir de la distancia euclidiana entre cada compuesto y los compuestos más cercanos van formando conjuntos que se van uniendo a otros mediante ramas como si fuera un árbol.[47,98] El número de vecinos más cercanos (*k*=50) y el factor usado para argumentar la calidad de los vecinos, (*kc*=10) fueron usados para desarrollar los gráficos de TMAP, y cuyos parámetros se ha aplicado previamente.[47]

7.2.3. Visualizaciones de espacio químico utilizando una huella digital de bases de datos basada en estadística

Se construyeron dos huellas digitales moleculares basadas en el concepto de Huella Digital de Bases de datos Basada en Estadística (en inglés, *Statistical-Based Database Fingerprint*, SB-DFP),[99] una para productos naturales a partir de COCONUT (COCONUT SB-DFP) y otra para compuestos accesibles sintéticamente a partir de REAL (REAL SB-DFP). Se utilizó como representación molecular a ECFP-4 (1024 bits). SB-DFP parte de una representación molecular, en este caso ECFP-4, y la frecuencia de ocurrencia de cada bit en un conjunto de datos (A) que es comparado con un conjunto de datos de referencia (B). Si la frecuencia del bit en A es estadísticamente mayor que en B, se le asigna un valor de 1 al bit, y en el caso contrario, un valor de 0. El 100% de los compuestos de ChEMBL, el 100 % de los fragmentos moleculares generados de COCONUT, REAL y ChEMBL, y el 40 % restante de los compuestos de COCONUT y REAL fueron utilizados para calcular los valores de similitud de los compuestos utilizando SB-DFP y el coeficiente de Tanimoto (valores entre 0 y 1). Los valores de similitud fueron utilizados para generar visualizaciones de espacio químico de productos naturales, compuestos accesibles sintéticamente y compuestos de relevancia biológica (ver **Sección 8.1.1.3**).

7.3. Generación de fragmentos moleculares

Los fragmentos moleculares se generaron utilizando el algoritmo de análisis combinatorio retrosintético por sus siglas en inglés RECAP,[100] implementado en la biblioteca RDKit. El algoritmo es basado en romper once enlaces claves que pueden regenerarse a partir de sustancias químicas conocidas, y son (1) amida, (2) éster, (3) amina, (4) urea, (5) éter, (6) olefina, (7) nitrógeno cuaternario, (8) nitrógeno aromático-carbono alifático, (9) carbono alifático-nitrógeno lactámico, (10) carbono aromático-carbono aromático y (11) sulfonamida. El algoritmo rompe cualquiera de esos enlaces si están presentes en una molécula, **Figura 11**.

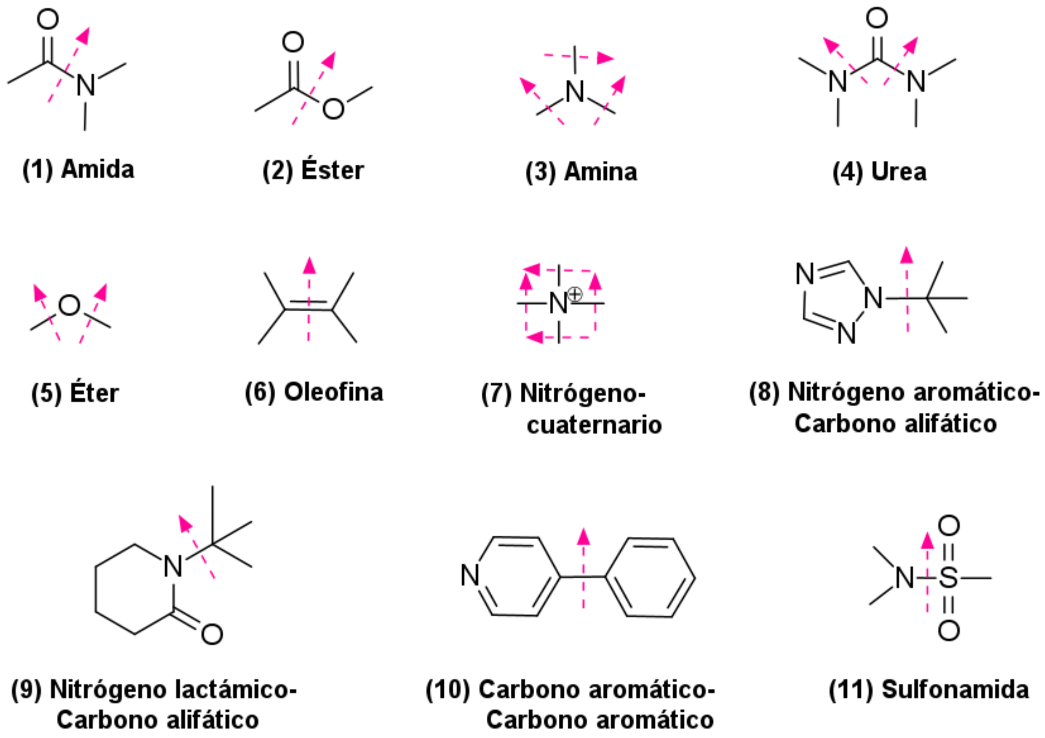

(1) Amida
(2) Éster
(3) Amina
(4) Urea
(5) Éter
(6) Oleofina
(7) Nitrógeno-cuaternario
(8) Nitrógeno aromático-Carbono alifático
(9) Nitrógeno lactámico-Carbono alifático
(10) Carbono aromático-Carbono aromático
(11) Sulfonamida

Figura 11. Enlaces clave que rompen el algoritmo RECAP para generar fragmentos moleculares. Fuente: Lewell X. Q, Judd DB, Watson SP, Hann MM. RECAP--Retrosynthetic Combinatorial Analysis Procedure: A Powerful New Technique for Identifying Privileged Molecular Fragments with Useful Applications in Combinatorial Chemistry. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 511–522.

7.4. Diseño de novo utilizando fragmentos de productos naturales

Se desarrolló una biblioteca enfocada a compuestos inhibidores de la proteasa viral del VIH-1 a partir de fragmentos de productos naturales utilizando el protocolo de Saldívar-González *et al.*[50] para enumerar bibliotecas de compuestos químicos,[50] el lenguaje de programación Python, la biblioteca de Python RDKit,[95] y el esquema de reacción propuesto por Zhao *et al.*[101] El esquema de reacción de Zhao *et al.*[101] derivado de un análisis de relaciones estructura actividad (SAR por sus siglas en inglés, *structure activity relationships*) sugiere la optimización de bevirimat, un compuesto derivado del ácido betulínico (**Figura 12**) y en pruebas clínicas.[102,103] Bevirimat se une a la poliproteína Gag inhibiendo la acción de la proteasa viral en su último evento de la última escisión de la proteína de la cápside y del péptido espaciador 1 (CA-SP1).[104,105] El esquema propuesto para construir nuevos compuestos químicos derivados de bevirimat se muestra en la **Figura 13**.

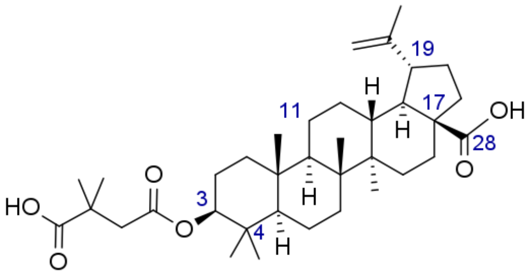
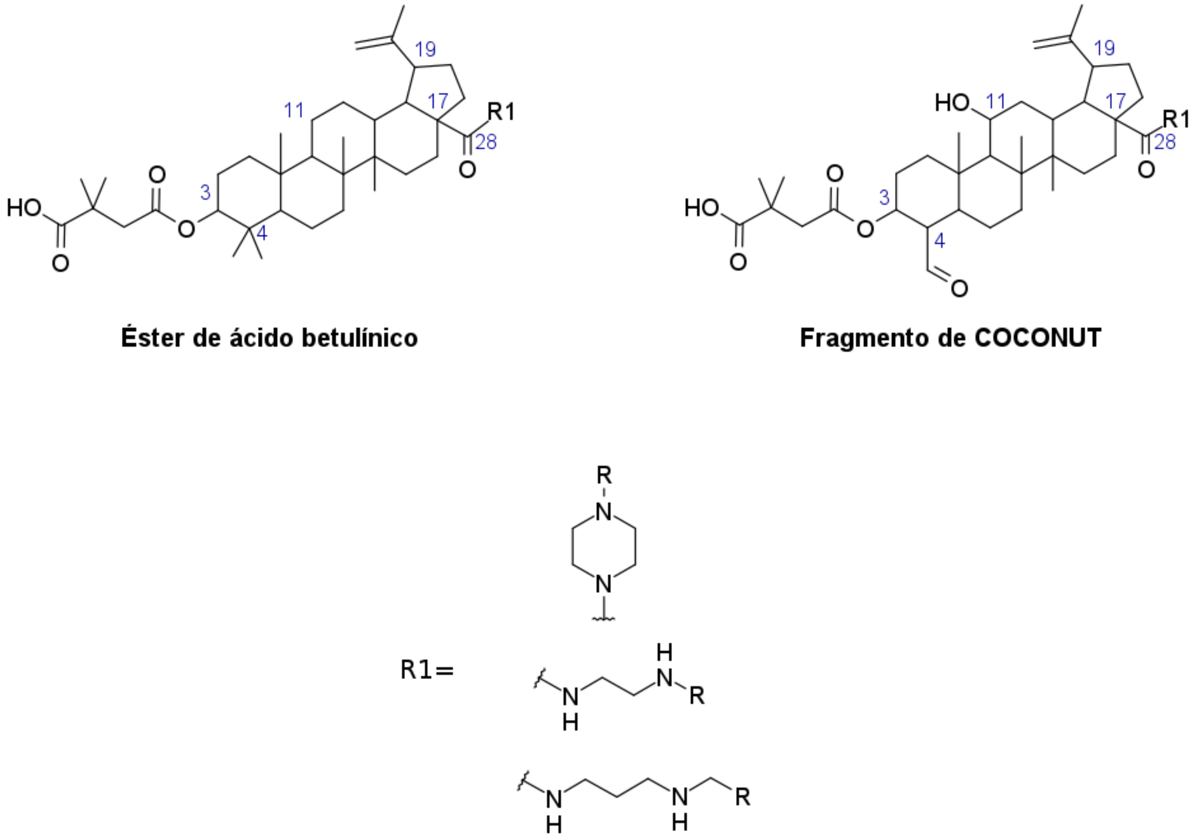


Figura 12. Estructura química de bevirimat.



Éster de ácido betulínico

Fragmento de COCONUT

Figura 13. Esquema para construir nuevos compuestos químicos similares a bevirimat usando el éster del ácido betulínico y éster de un fragmento de COCONUT derivado del ácido 24-nor-3α,11α-dihydroxy-lup-20(29)-en-23,28-dioico.

A partir de este esquema, se generaron SMARTS y SMIRKS usando SMARTviewer.[106] Los SMARTS fueron usados para filtrar ciclohexanol, piperazina, 1,2-diaminoetano, 1,3-diaminopropano y el sistema cíclico derivado del ácido betulínico. Los SMIRKS se utilizaron para representar las reacciones de esterificación y aminación entre los fragmentos de las bibliotecas COCONUT, ChemDiv y Enamine.

Después, se calculó la diversidad estructural de los compuestos generados como se describe en la **sección 7.2.1** y se generaron visualizaciones de espacio químico mediante PCA y TMAP siguiendo los protocolos de la **sección 7.2.2**. Para reducir el espacio de búsqueda en el espacio químico se calcularon seis propiedades fisicoquímicas de interés farmacéutico de las bibliotecas virtuales enfocadas de compuestos inhibidores de la proteasa viral del VIH-1 generadas a partir de fragmentos moleculares y los compuestos inhibidores de la proteasa viral del VIH-1 aprobados por la FDA. Se tomaron los valores máximos de las propiedades fisicoquímicas obtenidas de los inhibidores de la proteasa viral del VIH-1 aprobados por la FDA. Las moléculas retenidas debía cumplir con al menos cuatro condiciones, entre ellas log P. Estos conjuntos de propiedades y valores fueron usados como regla heurística que es ligeramente menos estricta que las normas Lipinski[107] y Veber.[108]

7.4.1. Accesibilidad sintética

Se calculó la viabilidad sintética de los compuestos generados usando la función de puntuación SAscore (por sus siglas en inglés, *sintetic accesibility score*). Se sugiere las moléculas que tiene un valor de SAscore < 6 son fáciles de sintetizar químicamente.[109]

El SAscore se calcula como la diferencia entre la puntuación del fragmento y la penalización por complejidad (ver **Ecuación 2**).

$$SAscore \ = \ Puntuación\, de\, fragmentos\, moleculares \ - \ Penalización\, por\, complejidad$$

Ecuación 2. Algoritmo SAscore para calcular accesibilidad sintética.

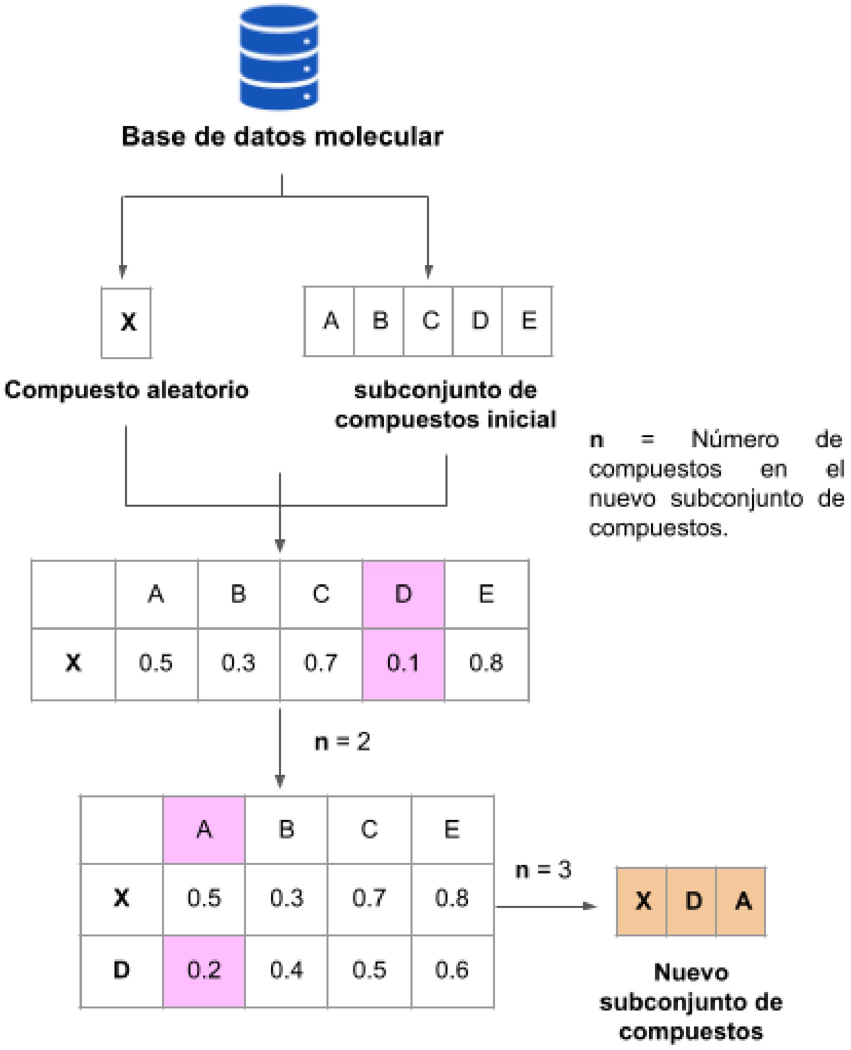
La puntuación de fragmentos moleculares (ver Apéndice, **Ecuación A1**) captura las características estructurales en un gran número de moléculas ya sintetizadas (934,046 moléculas representativas de PubChem). Las moléculas son fragmentadas utilizando la función *ECFP_4# fragments*, y la puntuación del fragmento es calculada como una suma de contribuciones de todos los fragmentos en la molécula y dividido por el número de fragmentos en la molécula. La frecuencia de los fragmentos está relacionada con su accesibilidad sintética y, por lo tanto, las subestructuras fáciles de preparar están presentes en las moléculas frecuentemente.

La penalización por complejidad (ver Apéndice, **Ecuación A2**) es calculada como la suma de complejidad de anillos (átomos cabeza de puente y átomos espiro), el número de estereocentros, el número de macrociclos (anillos con más de ocho átomos pesados aumentan la complejidad molecular) y el peso molecular.

La función de puntuación SAscore se calculó para las bibliotecas de compuestos virtuales enfocadas a inhibidores de la proteasa viral de VIH-1 generadas a partir de los fragmentos de COCONUT, ChemDiv y Enamine, y dos bases de datos de referencia como son fármacos aprobados por la FDA y compuestos inhibidores de la proteasa viral del VIH-1 aprobados por la FDA.[110] SAscore fue calculado usando un script de Python publicado por Ertl y Schuffenhauer.[109]

7.4.1. Subconjuntos de productos naturales

Se generaron tres subconjuntos con 14,994, 7,497 y 4,998 productos naturales únicos y derivados del UNPD utilizando el algoritmo MaxMin,[111] el cual se describe en la **Figura 14**. Primero 153,375 compuestos del UNPD codificados como SMILES fueron estandarizados utilizando el lenguaje de programación Python, ver **Sección 7.1**, pero conservando la estereoquímica de los compuestos. El UNPD se dividió en tres formas diferentes, es decir, primero en 30 subconjuntos iniciales con 5,000 compuestos cada uno (experimento-A); después en 15 subconjuntos iniciales con 10,000 compuestos cada uno (experimento-B), y por último, 10 subconjuntos iniciales con 15,000 compuestos cada uno (experimento-C). Para cada subconjunto inicial, un compuesto aleatorio (X) es seleccionado. En la **Figura 14**, se muestra un subconjunto de compuestos representados por A, B, C, D y E. La similitud molecular es calculada entre el compuesto X y cada uno de los compuestos del subconjunto inicial remanente utilizando como métrica el coeficiente de Tanimoto[97] (**Ecuación 1**). El compuesto que tiene el valor de similitud más pequeño, es decir, el compuesto más diverso (compuesto D) es seleccionado, se quita del subconjunto remanente, y se agrupa con el compuesto X. Después, se calcula la similitud molecular de los compuestos remanentes con los compuestos D y X. Se selecciona el compuesto con el valor más pequeño de similitud molecular (compuesto A); se quita el compuesto A del conjunto remanente, y se une al nuevo subconjunto formado por los compuestos X y D. El proceso se repite hasta generar un subconjunto con el número de compuestos deseados. De cada experimento, se seleccionaron 500 compuestos y se unieron en un solo subconjunto final, y se obtuvieron tres subconjuntos de productos naturales finales derivados del UNPD con cerca de 15,000 (UNPD-A), 10,000 (UNPD-B) y 5,000 compuestos (UNPD-C). Los tres subconjuntos finales derivados del UNPD se encuentran disponibles públicamente en GitHub: https://github.com/DIFACQUIM/Natural-products-subsets-generation.


Base de datos molecular
X
A B C D E
Compuesto aleatorio
subconjunto de compuestos inicial
n = Número de compuestos en el nuevo subconjunto de compuestos.
n = 2
n = 3
X D A
Nuevo subconjunto de compuestos

	A	B	C	D	E
X	0.5	0.3	0.7	0.1	0.8

	A	B	C	E
X	0.5	0.3	0.7	0.8
D	0.2	0.4	0.5	0.6

Figura 14. Algoritmo MaxMin. De una base de datos molecular (UNPD) se seleccionó un compuesto aleatorio (X). Después, la base de datos sin X fue dividida en un número dado de subconjuntos de compuestos. En esta figura se muestra un subconjunto de compuestos representados por A, B, C, D y E. La similitud molecular fue calculada entre el compuesto X y cada compuesto del subconjunto (A, B, C, D y E). El compuesto con el valor de similitud molecular más pequeño se retira del subconjunto y se genera un nuevo subconjunto de compuestos (X y D). El proceso continúa *n* número de veces hasta obtener el número de compuestos deseados, en este ejemplo *n*=3.

VIII. Resultados y discusión

8.1. Generación de fragmentos moleculares

Se desarrollaron dos publicaciones científicas sobre generación y caracterización de fragmentos de productos naturales descritas en las **secciones 8.1.1** y **8.1.2**.

8.1.1. Caracterización de fragmentos de productos naturales, compuestos sintéticos y compuestos de relevancia biológica

Se curaron 15,547,017 compuestos disponibles sintéticamente de la base de datos de REAL (comercializada por Enamine),[112] 412,903 productos naturales de COCONUT[35] y 1,844,434 compuestos con actividad biológica de la base de datos ChEMBL.[113,114] Después, seis propiedades fisicoquímicas de interés farmacéutico fueron calculadas y se retuvieron los compuestos que cumplieran con la regla de Lipinski[107] y Veber[108] (MW ≤ 500, log P ≤ 5, HBA ≤ 10, HBD ≤ 5, RB ≤10 y TPSA ≤ 140), y que no tuviera PAINS (por sus siglas en inglés, *pan-assay interference compounds*). Las reglas de Lipisnki y Veber son reglas empíricas que determinan si un compuesto químico con una actividad biológica tiene probabilidad de ser un fármaco administrado por vía oral. Mientras que los PAINS son compuestos químicos que podrían perturbar la tecnología de ensayo utilizada en el cribado para notificar la actividad biológica, y que no son activos frente a la diana biológica prevista.[115–117]

Tabla 1. Compuestos y fragmentos moleculares generados de COCONUT, REAL y ChEMBL.

Bases de datos	Compuestos iniciales	Compuestos procesados	Fragmentos generados	Referencias
COCONUT	412,903	190,139	205,904	[35]
REAL (Enamine)	15,547,017	15,297,437	11,243,073	[112]
ChEMBL	1,844,434	1,074,335	1,177,361	[113,114]

La **Tabla 1** resume el número de compuestos originales, compuestos procesados que cumplieron con las reglas de Lipinski y Verber, y sin PAINS, y fragmentos moleculares generados de COCONUT, REAL y ChEMBL.[118] Se obtuvieron 190,139 compuestos de COCONUT, 15,297,437 compuestos de

REAL y 1,074, 336 de compuestos ChEMBL. Se encontraron 16,529,500 compuestos únicos entre las bases de datos de COCONUT, REAL y ChEMBL. El 83.1 % COCONUT, 97.0 % ChEMBL y 99.9 % REAL correspondieron a compuestos únicos (ver Apéndice, **Figura A1a**,). COCONUT y REAL tuvieron 22 compuestos en común y COCONUT y ChEMBL tuvieron 32,053 compuestos en común. A partir de las bases de datos de compuestos procesados se generaron las bibliotecas de fragmentos moleculares utilizando el algoritmo RECAP[100] (ver **sección 7.3**). Se obtuvieron 205,904 fragmentos moleculares de COCONUT, 11,243,073 de REAL y 1,177,361 de ChEMBL. Uniendo las tres bases de datos de fragmentos moleculares, se obtuvieron 2,497,641 fragmentos únicos (99 %) (ver Apéndice, **Figura A1c**). El porcentaje de fragmentos únicos y generados fueron 72.2% COCONUT, 82.6 % ChEMBL y 99.3 % REAL. También, se generaron los núcleos estructurales base de cada base de datos utilizando la definición de Bemis y Murcko.[119] Se obtuvieron 6,852,628 núcleos estructurales base únicos, 99.1 % (ver Apéndice, **Figura A1b**), y para cada base de datos fueron 68.7 % COCONUT, 82.6 % ChEMBL y 99.3 % REAL, una tendencia muy similar a la de los fragmentos moleculares. Esto sugiere que al menos el 68.7 % de núcleos estructurales base y el 72.2 % de fragmentos moleculares de productos naturales (COCONUT) no están totalmente cubiertos por los compuestos accesibles sintéticamente (REAL) o los compuestos probados biológicamente (ChEMBL), y sustenta la premisa de que los fragmentos moleculares derivados de productos naturales tiene estructuras novedosas que pueden servir como bloques de construcción para el diseño de nuevos fármacos a partir del diseño *de novo*.[120]

8.1.1.1. Análisis de fragmentos moleculares

Se calculó la media de distribución de los descriptores moleculares asociados a las diferencias estructurales de los fragmentos moleculares únicos y en común de COCONUT, REAL y ChEMBL. La **Tabla 2** muestra que COCUNUT, REAL y ChEMBL tuvieron fragmentos moleculares con alrededor de 20 átomos pesados. COCONUT tuvo el mayor número de átomos de oxígeno (grupos hidroxilos y epóxidos), anillos alifáticos y biciclos de acuerdo al número de átomos de cabeza de puente y átomos espiro, y calculados mediante la media de distribución de átomos de oxígeno, anillos alifáticos, átomos de cabeza de puente y átomos espiro: 3.793, 1.522, 0.282 y 0.11, respectivamente; en comparación a REAL (2.080, 1.036, 0.108 y 0.053) y ChEMBL (2.130, 0.647, 0.052 y 0.022). Sin embargo, los fragmentos moleculares de COCONUT tuvieron menor número de átomos de nitrógeno y anillos aromáticos (media de distribución de átomos de nitrógeno y anillos aromáticos: 0.847 y 0.957, respectivamente) en comparación a los fragmentos de REAL (3.006, 1.341) y ChEMBL (2.562, 1.857).

Tabla 2. Diversidad estructural de fragmentos moleculares únicos y en común entre COCONUT, REAL (Enamine) y ChEMBL.

Características estructurales	COCONUT	REAL (Enamine)	ChEMBL	Fragmentos en común
Átomos de oxígeno	3.793	2.080	2.130	1.300
Átomos de nitrógeno	0.847	3.006	2.562	1.119
Átomos pesados	20.922	19.583	19.784	10.788
Número de anillos	2.479	2.377	2.504	1.172
Número de anillos Alifáticos	1.522	1.036	0.647	0.252
Número de anillos aromáticos	0.957	1.341	1.857	0.920
Número de heterociclos	1.077	1.556	1.371	0.538
Número de heterociclos alifáticos	0.707	0.694	0.487	0.184
Número de heterociclos aromáticos	0.369	0.862	0.884	0.354
Átomos spiro	0.110	0.053	0.022	0.001
Átomos cabeza de puente	0.282	0.108	0.052	0.020

[a]Media de distribución.

La **Figura A2a-c** (ver Apéndice) muestra las estructuras químicas de los diez fragmentos moleculares únicos y más frecuentes de COCONUT, REAL y ChEMBL. Los fragmentos de COCONUT tuvieron un mayor número biciclos, átomos de oxígeno, grupos hidroxilo, mientras que REAL y ChEMBL tuvieron un mayor número de átomos de nitrógeno y anillos aromáticos. Usualmente, los productos naturales contienen grupos funcionales con átomos de oxígeno como hidroxilos, anillos de epóxido, ésteres y peróxidos, mientras que las moléculas obtenidas sintéticamente tienen un mayor contenido de átomos de nitrógeno y grupos funcionales más accesibles como amida, urea, sulfonanida, sulfona y sustituyeres como flúor.[121]

La **Figura A2d** y la **Tabla A2** muestran que los fragmentos comunes entre COCONUT, REAL y ChEMBL fueron los menos diversos y de menor tamaño (media de distribución del número de átomos pesados, átomos de oxígeno, átomos de nitrógeno, átomos espiro, átomos de cabeza de puente, anillos alifáticos y anillos aromáticos: 10.788, 1.3, 1.119, 0.001, 0.020, 0.252 y 0.920, respectivamente).

8.1.1.2. Diversidad y complejidad estructural

La diversidad estructural de los fragmentos moleculares y las bases de datos fue calculada a partir de la mediana de similitud de pares de moléculas usando dos huellas digitales moleculares MACCs Keys (166-bits) y ECFP-4 (1024-bits). La **Tabla 3** y la **Tabla 4** muestran la diversidad y complejidad estructural de las bases de datos y los fragmentos moleculares generados. La **Tabla 3** muestra que COCONUT tuvo los compuestos más diversos, en ambas huellas digitales moleculares (mediana de similitud con MACCS keys y ECFP-4: 0.344, 0.111, respectivamente), seguida de ChEMBL (0.377, 0.119) y REAL (0.420, 0.123). La misma tendencia se observó en las bases de datos de fragmentos moleculares La **Tabla 4** muestra que COCONUT tuvo los fragmentos moleculares más diversos, en ambas huellas digitales moleculares (mediana de similitud con MACCS keys y ECFP-4: 0.314, 0.117, respectivamente), seguido por ChEMBL (0.334, 0.122) y REAL (0.408, 0.134).

Tabla 3. Diversidad estructural basada en huellas digitales moleculares y complejidad molecular basada en la fracción de carbonos sp^3 y la fracción de carbonos quirales de COCONUT, REAL y ChEMBL.

Bases de datos	MACCS Keys[a] (166-bits)	ECFP-4[a] (1024-bits)	Media de la fracción de carbonos sp^3	Media de la fracción de carbonos quirales
COCONUT	0.344	0.111	0.453	0.112
REAL (Enamine)	0.420	0.123	0.526	0.068
ChEMBL	0.377	0.119	0.318	0.033

[a]Mediana de similitud.

Tabla 4. Diversidad estructural basada en huellas digitales moleculares y complejidad molecular basada en la fracción de carbonos sp^3 y la fracción de carbonos quirales de los fragmentos moleculares generados de COCONUT, REAL y ChEMBL.

Bases de datos	MACCS Keys[a] (166-bits)	ECFP-4[a] (1024-bits)	Media de la fracción de carbonos sp^3	Media de la fracción de carbonos quirales
COCONUT	0.314	0.117	0.518	0.175
REAL (Enamine)	0.408	0.134	0.516	0.074
ChEMBL	0.334	0.122	0.335	0.046

[a]Mediana de similitud.

La complejidad estructural fue calculada mediante el cálculo de la media de carbonos con hibridación sp^3 y la fracción de carbonos quirales. La **Tabla 3** y **Tabla 4** muestran la complejidad estructural de los compuestos y fragmentos moleculares de COCONUT, REAL y ChEMBL. Los compuestos de COCONUT fueron los más diversos en términos de la media de la fracción de carbonos con hibridación sp^3 y la fracción de carbonos quirales: 0.518, 0.175, respectivamente; seguidos por REAL (0.516, 0.074) y ChEMBL (0.335, 0.046). Los fragmentos moleculares de COCONUT fueron los más complejos en términos de la media de fracción de carbonos con hibridación sp^3 y la fracción de carbonos quirales: 0.519, 0.175, respectivamente; seguidos por REAL (0.516, 0.074) y ChEMBL (0.335, 0.046). Los fragmentos moleculares preservan la diversidad y complejidad estructural de los compuestos originales (producto natural, compuesto accesible sintéticamente o compuesto de relevancia biológica), es decir, conservan las diferencias asociadas al origen de los compuestos.[4]

8.1.1.3. Visualización de espacio químico

Se realizó una visualización del espacio químico de los compuestos y fragmentos moleculares de COCONUT, REAL y ChEMBL utilizando las huellas digitales moleculares COCONUT SB-DFP (huella digital molecular representativa de los productos naturales), REAL SB-DFP (huella digital molecular representativa de los compuestos sintéticos) utilizando como representación molecular ECFP-4 de 1024 bits (ver **Sección 7.2.3**). La **Figura 15** muestra una visualización de espacio químico basado en similitudes de SB-DFP. En el gráfico, cada estructura química es graficada de acuerdo a su valor de similitud utilizando COCONUT SB-DFP y REAL SB-DFP. La Figura **15a-c** muestra los compuestos de COCONUT (**Figura 15a**) y REAL (**Figura 15b**) que no se utilizaron para construir las huellas digitales moleculares representativas de productos naturales y de compuestos sintéticos. La **Figura 15c** muestra todos los compuestos de ChEMBL. La **Figura 15d-f** muestra los fragmentos moleculares generados de las bases de datos de COCONUT (**Figura 15d**), REAL (**Figura 15e**) y ChEMBL (**Figura 15f**). En cada sección de la **Figura 15** (**a-f**), el número de compuestos es representado con una escala de color continua entre morado (región menos densa de compuestos químicos) y amarillo (región más densa de compuestos químicos). La visualización de espacio químico muestra que los productos naturales tienden a ocupar un espacio químico cercano a la huella digital molecular representativa de los productos naturales, COCONUT SB-DFP, (**Figura 3a**), mientras que los compuestos disponibles sintéticamente son muy cercanos a la huella digital molecular representativa de los compuestos sintéticos, REAL SB-DFP, (**Figura 3b**). Los compuestos de ChEMBL compartieron espacio químico con los compuestos de COCONUT y REAL, siendo más cercanos a los compuestos de REAL (**Figura 3c**). Los fragmentos moleculares derivados de productos naturales, compuestos accesibles sintéticamente y de relevancia biológica siguen la misma tendencia que los compuestos originales de los que parten, y sustenta la premisa que los fragmentos moleculares preservan las propiedades estructurales de los compuestos originales de los cuales fueron generados.

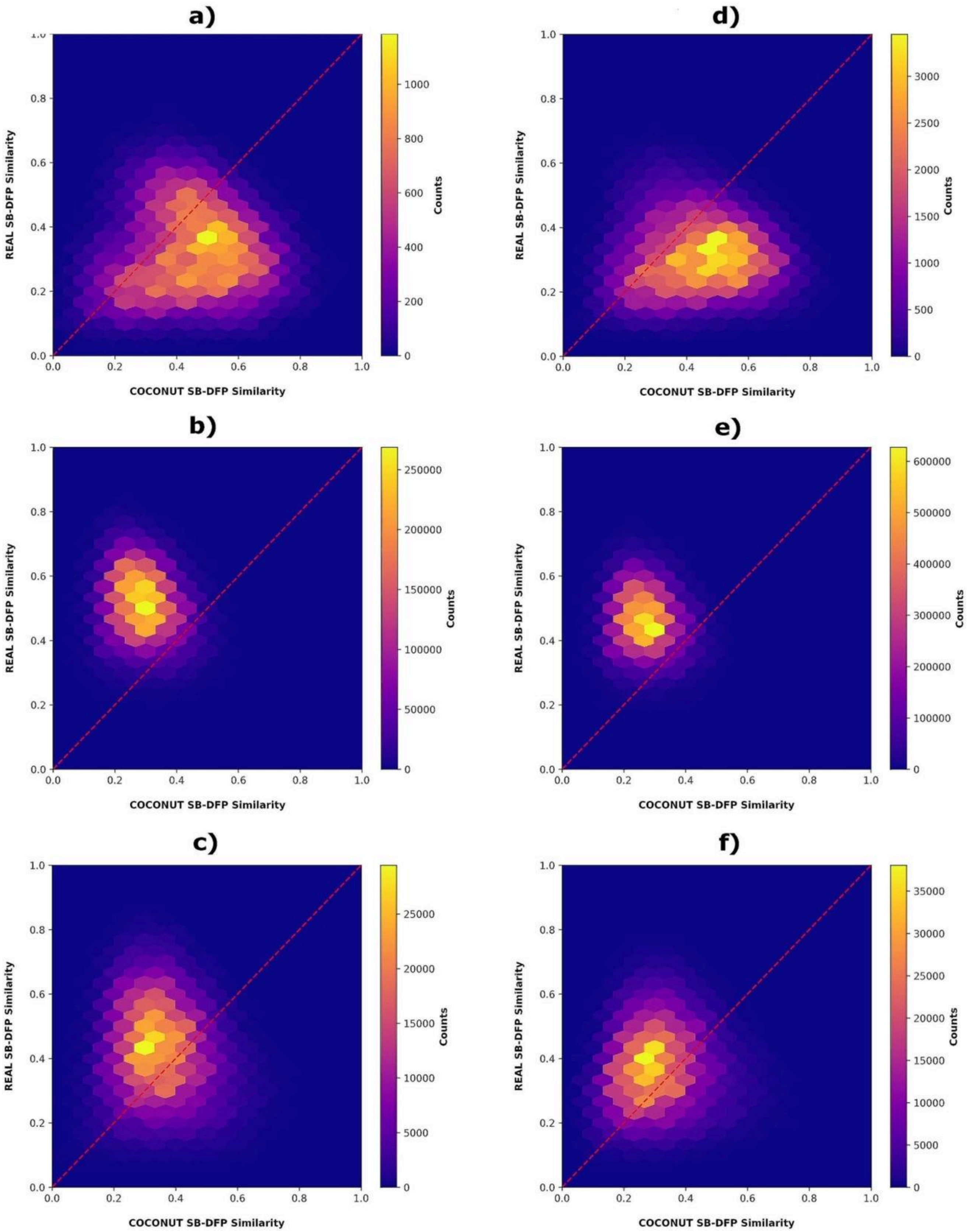


Figura 15. Representación visual del espacio químico de compuestos y fragmentos de productos naturales, compuestos sintéticos y compuestos de relevancia biológica. El número de compuestos es representado con una escala continúa de color. Bases de datos de compuestos químicos: a) COCONUT, b) REAL y c) ChEMBL. Bases de datos de fragmentos moleculares: d) COCONUT, e) REAL y f) ChEMBL.

8.1.2. Caracterización de fragmentos de productos naturales, compuestos químicos alimentarios y compuestos enfocados a COVID-19

Se curó COCONUT[35] *version 4* con 432,706 productos naturales y se comparó con cuatro bases de datos de referencia: *Food Database* (FooDB)[122] con 23,883 compuestos químicos de alimentos que están fuertemente asociados a los productos naturales. *Dark Chemical Matter* (DCM)[123] con 139,352 compuestos sin actividad biológica, pero que recientemente han llevado a la identificación de compuestos bioactivos. Ante la pandemia de la enfermedad causada por coronavirus del 2019 (COVID-19), se seleccionaron dos bibliotecas con relevancia en el descubrimiento de fármacos y relacionados con esta enfermedad: *Chemical Abstract Service* (CAS),[124] enfocada en COVID-19 con 48,876 compuestos, y un conjunto de 280 compuestos inhibidores de la principal proteasa de SARS-CoV-2 (3CLP).[125]

8.1.2.1. Análisis de fragmentos moleculares

La **Tabla 5** muestra los compuestos originales, los compuestos procesados (compuestos curados y con peso molecular menor o igual a 1300 Da) y los fragmentos moleculares generados de COCONUT, FooDB, DCM, CAS y 3CLP.[126]

Tabla 5. Compuestos y fragmentos generados de COCONUT, FooDB, DCM, CAS y 3CLP.

Bases de datos	Compuestos originales	Compuestos procesados	Fragmentos generados	Referencia
COCONUT	432,706	382,248 (88 %)	52,630	[35]
FooDB	23,883	21,319 (89 %)	3,186	[122]
DCM	139,352	139,326 (99 %)	14,001	[123]
CAS	48,876	44,692 (91 %)	8,432	[124]
3CLP	280	256 (91 %)	108	[125]

De los compuestos procesados, se retuvieron 382,248 compuestos de COCONUT (88 %), 21,319 compuestos de FooDB (89 %), 139,326 compuestos de DCM (99 %), 44,692 compuestos de CAS (91 %) y 256 compuestos de 3CLP (91 %). Los compuestos con peso molecular menor o igual a 1300 Da fueron seleccionados utilizando este criterio porque se permite fragmentar más del 88 % de los compuestos, en poco tiempo. Posteriormente, se generaron los fragmentos moleculares de los compuestos procesados utilizando el algoritmo RECAP,[100] y fueron 52,630 fragmentos moleculares de COCONUT, 3,186 fragmentos moleculares de FooDB, 14,001 fragmentos moleculares de DCM, 8,432 fragmentos moleculares de CAS y 108 fragmentos moleculares de 3CLP; de los cuales se encontraron 28 fragmentos moleculares en común entre COCONUT, FooDB, DCM, CAS y 3CLP (ver Apéndice, **Figura A3**).

8.1.2.2. Diversidad y complejidad estructural

La diversidad estructural de los compuestos y fragmentos moleculares se calculó mediante la mediana de similitud utilizando dos huellas digitales moleculares ECFP-4 (1024-bits) y MACCS Keys (166-bits). Los compuestos y fragmentos moleculares con mediana de similitud cercana a cero, son los más diversos porque tienen menos subestructuras en común (ver **Ecuación 1**). La **Tabla 6** y la **Tabla 7** resumen la diversidad estructural de los compuestos y fragmentos moleculares generados de COCONUT, FooDB, DCM, CAS y 3CLP. Los compuestos de FooDB fueron los más diversos (mediana de similitud con MACCS Keys y ECFP-4: 0.322, 0.092, respectivamente), seguido de los compuestos de COCONUT (0.380, 0.107), ver **Tabla 6**. Del mismo modo, los fragmentos de FooDB fueron los más diversos (mediana de similitud con MACCS Keys y ECFP-4: 0.241,0.106, respectivamente), seguidos de los fragmentos de COCONUT (0.300, 0.111), ver **Tabla 7**. Los compuestos de CAS fueron menos diversos estructuralmente (0.473, 0.117), lo cual es consistente porque CAS es una base de datos enfocada a COVID-19. Además, los fragmentos moleculares de CAS fueron más diversos para ECFP-4 (0.095) que los fragmentos de DCM (0.125) y 3CLP (0.147).

Tabla 6. Diversidad estructural basada en huellas digitales moleculares y complejidad molecular basada en la fracción de carbonos sp^3 y la fracción de carbonos quirales de los compuestos de COCONUT, FooDB, DCM, CAS y 3CLP.

Bases de datos	ECFP-4[a] (1024-bits)	MACCS Keys[a] (166-bits)	Media de la fracción de carbonos sp^3	Media de la fracción de carbonos quirales
COCONUT	0.107	0.380	0.506	0.154
FooDB	0.092	0.322	0.620	0.152
DCM	0.136	0.407	0.342	0.028
CAS	0.117	0.473	0.489	0.145
3CLP	0.127	0.403	0.291	0.069

[a]Mediana de similitud.

Tabla 7. Diversidad estructural basada en huellas digitales moleculares y complejidad molecular basada en la fracción de carbonos sp^3 y la fracción de carbonos quirales de los fragmentos moleculares generados de COCONUT, FooDB, DCM, CAS y 3CLP.

Bases de datos	ECFP-4[a] (1024-bits)	MACCS Keys[a] (166-bits)	Media de la fracción de carbonos sp^3	Media de la fracción de carbonos quirales
COCONUT	0.111	0.300	0.557	0.189
FooDB	0.106	0.241	0.615	0.199
DCM	0.125	0.243	0.330	0.054
CAS	0.095	0.222	0.656	0.240
3CLP	0.147	0.214	0.298	0.071

[a]Mediana de similitud.

La complejidad estructural fue calculada mediante la media de carbonos con hibridación sp^3 y la media de la fracción de carbonos quirales. La **Tabla 6** y la **Tabla 7** muestran la complejidad estructural de los compuestos y los fragmentos moleculares generados de COCONUT, FooDB, DCM, CAS y 3CLP. Los compuestos de FooDB fueron los más diversos en términos de la media de la fracción de carbonos con hibridación sp^3 y la media de la fracción de carbonos quirales: 0.620, 0.152, respectivamente; seguidos por COCONUT (0.506, 0.154), CAS (0.489, 0.145), DCM (0.342, 0.028) y 3CLP (0.291, 0.069). Mientras que los fragmentos moleculares de CAS fueron los más diversos en términos de la media de la fracción de carbonos con hibridación sp^3 y la media de la fracción de carbonos quirales: 0.656, 0.240, respectivamente; seguidos por FooDB (0.615, 0.199), COCONUT (0.557, 0.189), DCM (0.330, 0.054) y 3CLP (0.298, 0.071). Estos valores indican que los fragmentos y compuestos de FooDB y COCONUT que representan a los productos naturales fueron los más diversos y complejos estructuralmente. La diversidad y complejidad estructural es una característica buscada en compuestos candidatos a fármacos, como los fragmentos moleculares de CAS, que fueron los más diversos; por lo tanto, los compuestos químicos de alimentos y los productos naturales poseen estructuras químicas que pueden ser utilizadas en el desarrollo de nuevos fármacos, en este caso de estudio para el tratamiento del COVID-19.

8.1.2.3. Visualización de espacio químico

Se realizó una visualización de espacio químico de los compuestos y los fragmentos moleculares de COCONUT, FooDB, DCM, CAS y 3CL, utilizando TMAP y la huella digital molecular ECFP-4 (1024-bits) como descriptor molecular. Cabe mencionar que TMAP permite visualizar un mayor número de compuestos, por ejemplo, más de 380,000 moléculas de COCONUT. La **Figura 16** y la **Figura 17** muestran las visualizaciones de espacio químico de los compuestos y fragmentos moleculares de las bases de datos representados en diferente color como COCONUT (cyan), FooDB (naranja), DCM (gris), CAS (rosa) y 3CLP (oliva). Para mejorar la claridad de cada visualización de espacio químico, se muestran cada conjunto de compuestos y fragmentos únicos en diferentes paneles, y su comparación directa con COCONUT, es decir, COCONUT-FooDB (morado), COCONUT-DCM (verde), COCONUT-CAS (negro) y COCONNUT-3CLP (rosa).

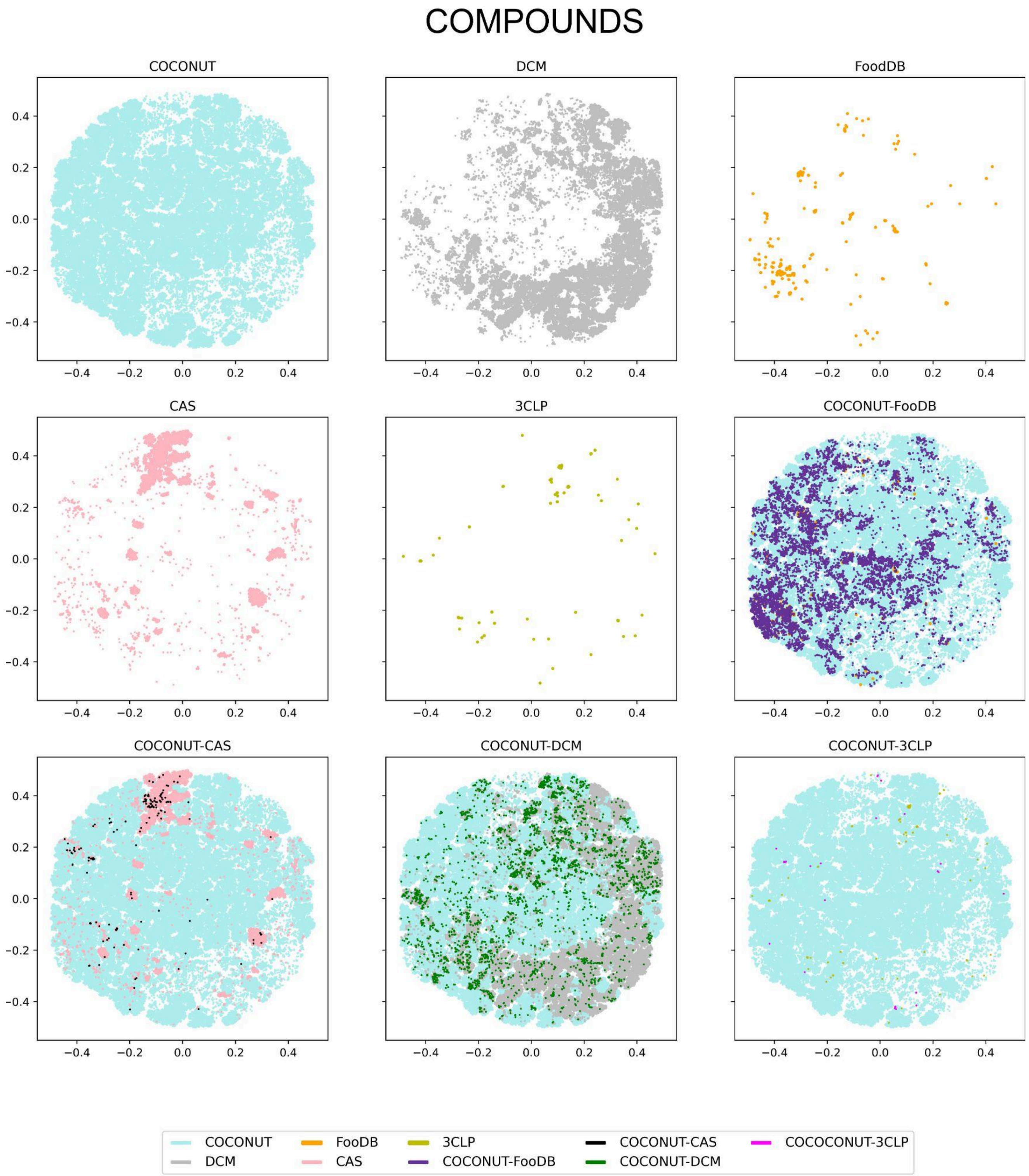


Figura 16. Visualización del espacio químico utilizando TMAP de las bases de datos de COCONUT, FooDB, DCM, CAS y 3CLP. Las bases de datos están representadas en color cyan (COCONUT), gris (DCM), naranja (FooDB), rosa (CAS) y oliva (3CLP). Los compuestos en común son representados en color púrpura (COCONUT-FooDB), negro (COCONUT-CAS), verde (COCONUT-DCM) y magenta (COCONUT-3CLP).

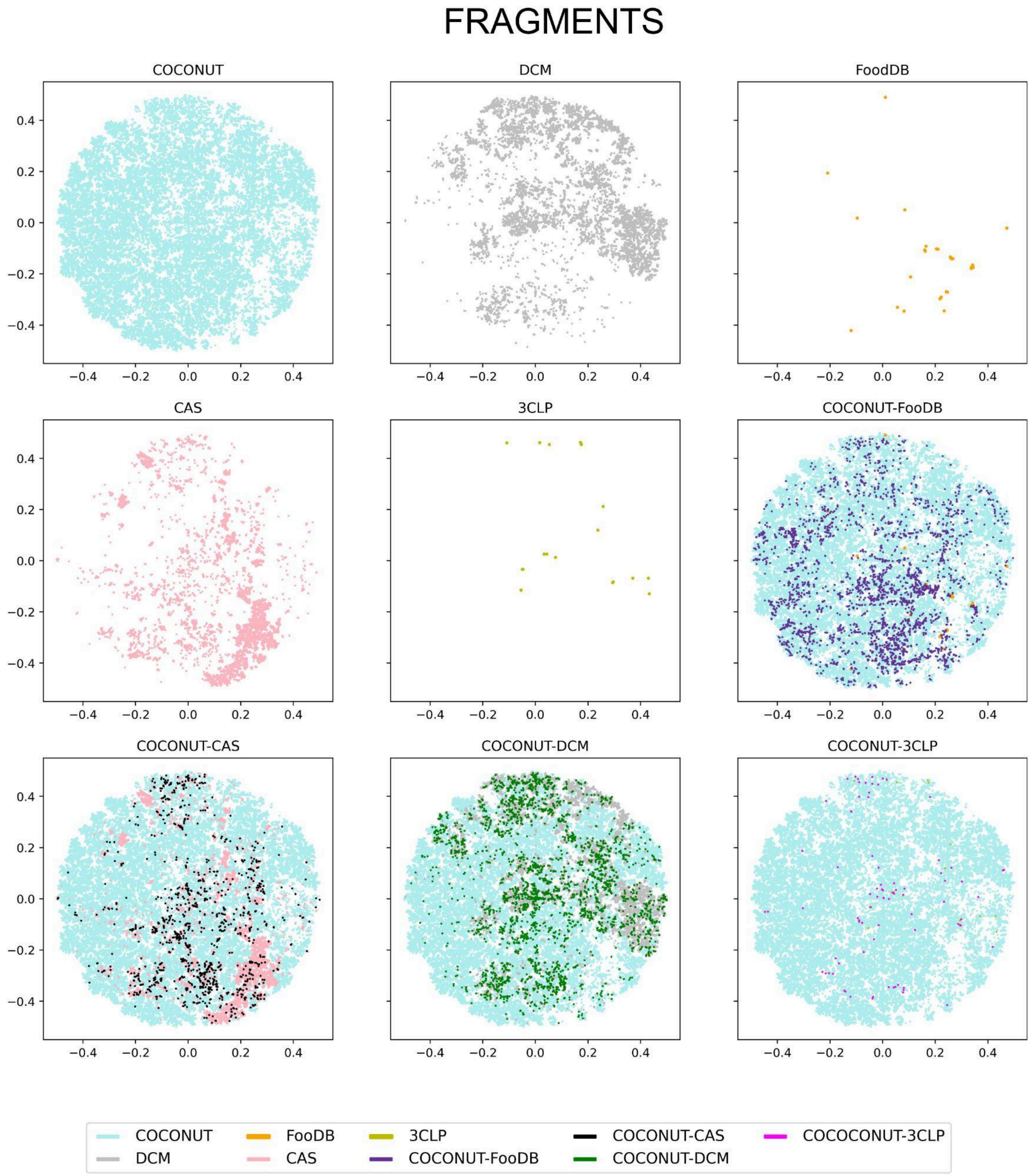


Figura 17. Visualización del espacio químico utilizando TMAP de las bases de datos de los fragmentos generados de COCONUT, FooDB, DCM, CAS y 3CLP. Las bases de datos de fragmentos moleculares están representadas en color cyan (COCONUT), gris (DCM), naranja (FooDB), rosa (CAS) y oliva (3CLP). Los compuestos en común son representados en color púrpura (COCONUT-FooDB), negro (COCONUT-CAS), verde (COCONUT-DCM) y magenta (COCONUT-3CLP).

La **Figura 16** muestra que todos los compuestos convergen en el espacio químico definido principalmente por COCONUT, seguido por DCM (**Figura 16**). FooDB tuvo 21,591 compuestos (90 %) en común con COCONUT (ver Apéndice, **Figura A3**). La **Figura 17** muestra que el espacio químico de los fragmentos moleculares fue definido en su mayoría por los fragmentos de COCONUT. En este caso, se encontró que FooDB tuvo 3,150 de fragmentos (99 %) en común con COCONUT (ver Apéndice, **Figura A3**). DCM tuvo 3,693 compuestos (2.6 %) en común con COCONUT y 2,993 fragmentos moleculares (21 %) de DCM en común con los fragmentos de COCONUT (ver Apéndice, **Figura A3**).

Las **Figuras 16** y **17** muestran que las moléculas pequeñas con escasa actividad biológica, como DCM, cubren una gran región del espacio químico cubierto por COCONUT (productos naturales) y algunas regiones del espacio químico de CAS (compuestos enfocados en COVID-19). Para abordar este punto, se realizó una comparación directa entre los compuestos y los fragmentos moleculares de DCM y CAS.

Las **Figuras 18** y **19** muestran una comparación directa entre los compuestos y fragmentos moleculares de CAS y DCM. Los compuestos de CAS y DCM convergen muy poco en el espacio químico, mientras que los fragmentos de CAS y DCM convergen en una mayor región del espacio químico, y concuerda con la **Figura A3,** donde DCM tiene 671 compuestos (0.03 %) en común con CAS. Mientras que los fragmentos moleculares de DCM tiene 46 fragmentos moleculares (4.8 %) en común con CAS. Esto sugiere que los fragmentos moleculares de DCM pueden utilizarse como bloques de construcción en el diseño *de novo* de moléculas enfocadas a COVID-19, a pesar de que los compuestos de DCM no tienen una actividad biológica reportada.

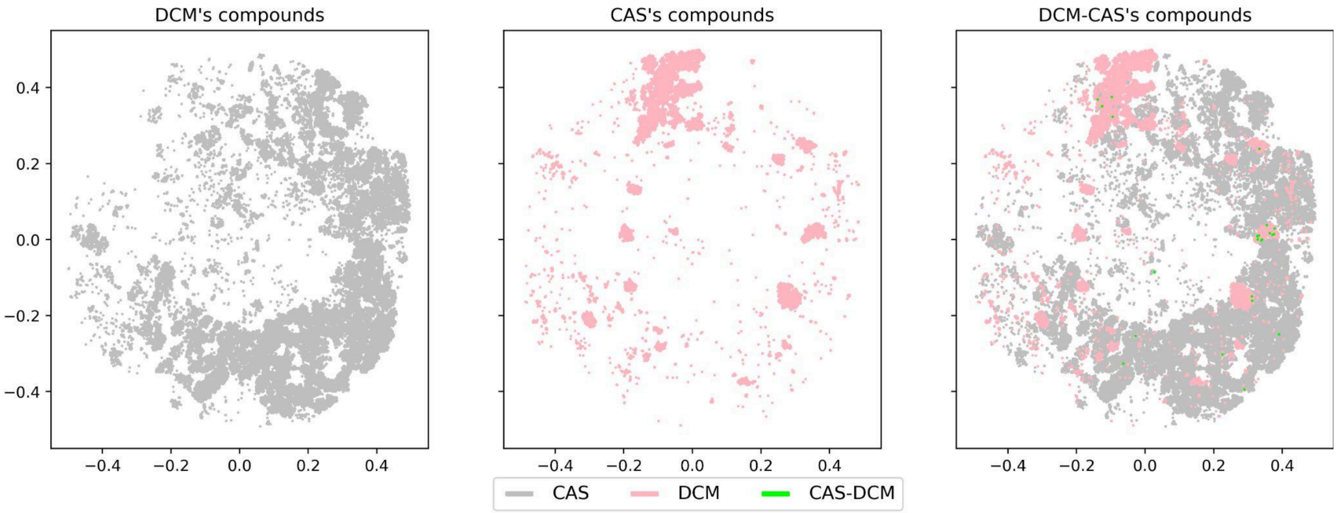


Figura 18. Visualización del espacio químico utilizando TMAP de los compuestos de CAS y DCM. Las bases de datos están representadas en color rosa (CAS), gris (DCM), y verde (los compuestos en común entre CAS y DCM).

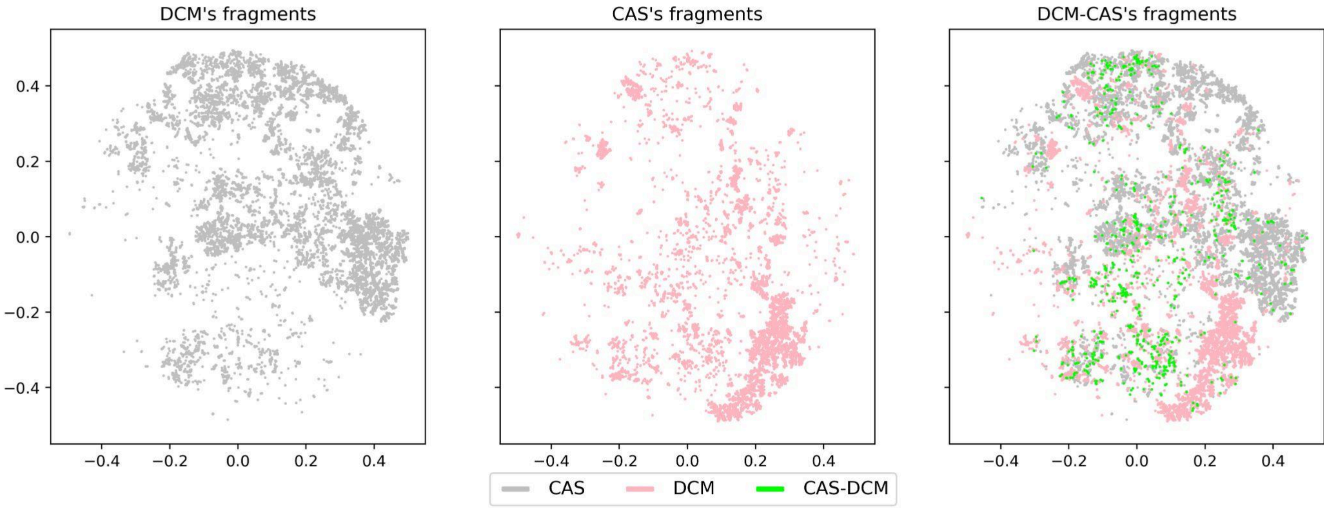


Figura 19. Visualización del espacio químico utilizando TMAP de los fragmentos de CAS y DCM. Las bases de datos están representadas en color rosa (CAS), gris (DCM), y verde (fragmentos en común entre CAS y DCM).

8.2. Diseño de novo de compuestos inhibidores de la proteasa viral de VIH-1

El estudio descrito en la sección **8.1.2.** mostró que los fragmentos de productos naturales convergen en el espacio químico de los fragmentos de compuestos antivirales enfocados a COVID-19, por lo cual los fragmentos de productos naturales pueden utilizarse para desarrollar nuevos compuestos antivirales. Para seguir explorando el espacio químico de compuestos antivirales generados a partir de fragmentos de productos naturales, se generaron bibliotecas enfocadas de compuestos inhibidores de la proteasa viral del VIH-1 a partir de fragmentos de COCONUT, ChemDiv y Enamine utilizando el protocolo descrito en la **sección 7.4**. Las bibliotecas de fragmentos moleculares fueron 184,769 fragmentos de productos naturales generados a partir de COCONUT y dos bibliotecas disponibles comercialmente con 4,063 fragmentos de ChemDiv enriquecidos con carbonos sp^3 y 4,160 fragmentos de productos naturales de Enamine.

Se generaron seis SMARTS para filtrar fragmentos con ciclohexanol, ácido 2,2-dimetil succínico, piperazina, 1,2-diaminoetano, 1,3-diaminopropano y el sistema cíclico derivado del ácido betulínico mostrados en la **Tabla A5** (ver Apéndice). Se usaron fragmentos de COCONUT con un sistema cíclico similar al ácido betulínico, un grupo hidroxilo unido al átomo de carbono 3, y un ácido carboxílico unido al átomo de carbono 17, como se muestra en la **Figura 20**. El fragmento de COCONUT derivado del ácido 24-nor-3α,11α-dihydroxi-lup-20(29)-en-23,28-dioico (COCONUT ID: CNP0243494 o Reaxys ID: 6547020) fue utilizado para construir la biblioteca enfocada de compuestos inhibidores de la proteasa viral del VIH-1 a partir de fragmentos de COCONUT. El ácido betulínico fue utilizado para construir la biblioteca enfocada de compuestos inhibidores de la proteasa viral del VIH-1 a partir de fragmentos de ChemDiv y Enamine (bibliotecas comerciales de fragmentos moleculares) porque no se encontraron en estas bibliotecas fragmentos similares al sistema cíclico derivado del ácido betulínico o triterpenos análogos.

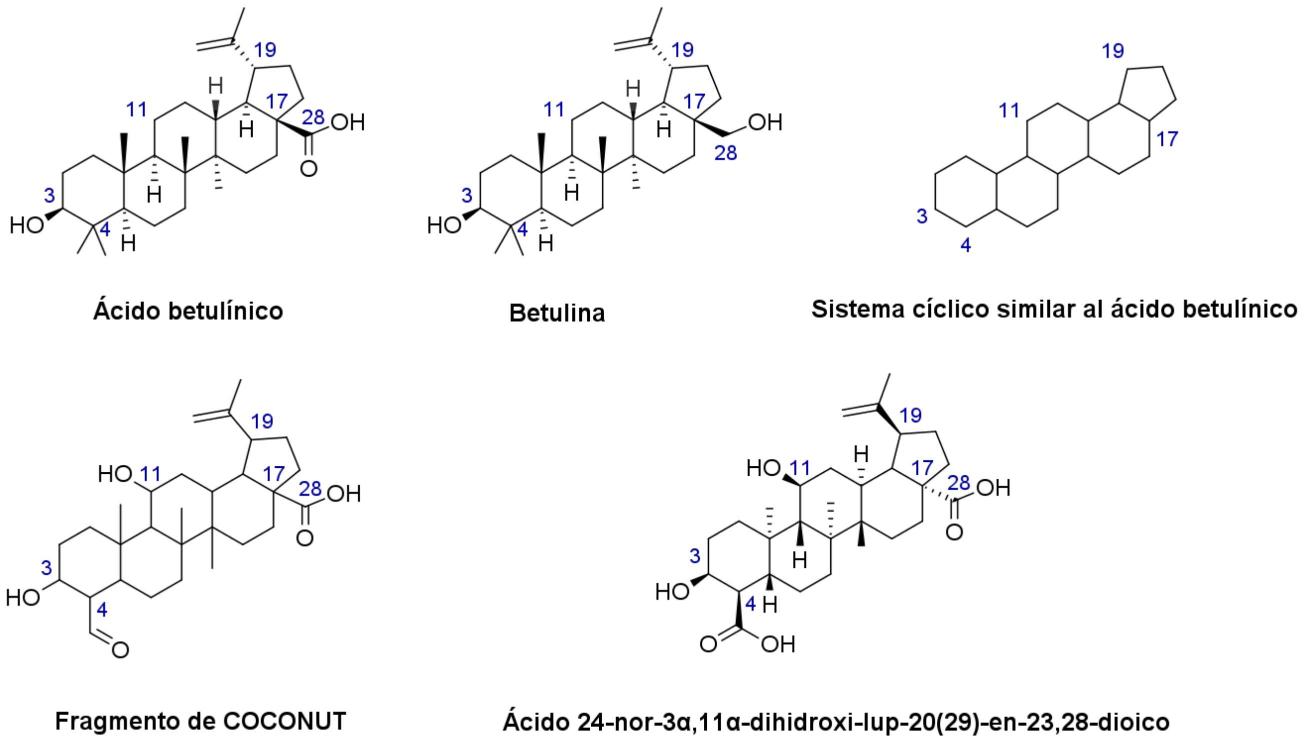

19
11
17
28
OH
HO
3
4
H
O
Ácido betulínico
Betulina
Sistema cíclico similar al ácido betulínico
Fragmento de COCONUT
Ácido 24-nor-3α,11α-dihidroxi-lup-20(29)-en-23,28-dioico

Figura 20. Estructuras químicas del ácido betulínico, betulina, y sistema cíclico derivado del ácido betulínico y el fragmento de COCONUT derivado con el sistema cíclico similar al ácido betulínico, pero derivado de la fragmentación del ácido 24-nor-3α,11α-dihidroxi-lup-20(29)-en-23,28-dioico.

Se construyeron cuatro SMIRKS como se muestra en la **Tabla A6** (ver Apéndice) para una reacción de esterificación y tres reacciones de amidación. La reacción 1, esterificación, fue realizada entre el alcohol del triterpeno y el ácido 2,2-dimetil succinico usando el SMIRKS 1 (ver Apéndice, **Tabla A6**, **SMIRKS 1**). La reacción 2, amidación, fue construida entre el grupo carboxilo unido al átomo de carbono 17 como se muestra en la **Figura 20** usando los fragmentos unidos a la piperazina, 1,3-diaminoetano, y 1,3-diaminopropano encontrados en los fragmentos de COCONUT, ChemDiv y Enamine. Los SMIRKS 2.1-2.3 fueron usados en la reacción 2 (ver Apéndice, **Tabla A6**, **SMIRKS 2.1-2.3**). Después, las estructuras generadas con errores de valencia fueron removidas. Los SMILES canónicos fueron generados y las moléculas duplicadas fueron removidas.

Se generaron tres bibliotecas enfocadas de compuestos inhibidores de la proteasa viral del VIH-1 con 1,534 moléculas utilizando fragmentos de COCONUT, 62 moléculas de fragmentos de ChemDiv, y once moléculas de fragmentos de Enamine. Los fragmentos unidos a 1,3-diaminopropano y 1,2-diaminoetano no se encontraron en las bibliotecas de fragmentos de ChemDiv y Enamine.

8.2.1. Similitud molecular

La mediana de similitud fue generada utilizando las huellas digitales moleculares ECFP-4 y MACCS keys son mostrados entre paréntesis y están descritos en la **Tabla A7** del Apéndice. Fármacos aprobados por la FDA (mediana de similitud con ECFP-4 y MACCS keys: 0.096, 0293, respectivamente) e inhibidores de la proteasa viral del VIH-1 aprobados por la FDA (0.253, 0.558) fueron las bases de datos más diversas, seguidas por los compuestos derivados de los fragmentos de COCONUT (0.605, 0.817), fragmentos de ChemDiv (0.676, 0.821) y fragmentos de Enamine (0.682, 0.823). Los compuestos generados computacionalmente de bases de datos de fragmentos fueron los menos diversos, lo cual era de esperarse, ya que son bibliotecas de compuestos similares a bevirimat.

8.2.2. Espacio químico

Se generaron dos visualizaciones de espacio químico utilizando PCA y TMAP. La **Figura 21** muestra una representación del espacio químico basado en seis propiedades fisicoquímicas de interés farmacéutico (MW, HBD, HBA, log P, TPSA y RB) utilizando PCA (**sección 7.2.2**). El componente principal 1 recuperó el 73.6% de la varianza, el componente principal 2 el 21.2% de la varianza (varianza acumulada de los dos componentes principales fue 94.8%). En esta visualización de espacio químico, los compuestos generados de las tres bibliotecas enfocadas de compuestos inhibidores de la proteasa viral del VIH-1 a partir de fragmentos de COCONUT, ChemDiv y Enamine se encuentran dentro del espacio químico de propiedades fisicoquímicas de interés farmacéutico de los fármacos aprobados por la FDA. Asimismo, algunos compuestos generados de los fragmentos de COCONUT tienen propiedades fisicoquímicas similares a los inhibidores de la proteasa viral del VIH-1 aprobados por la FDA.

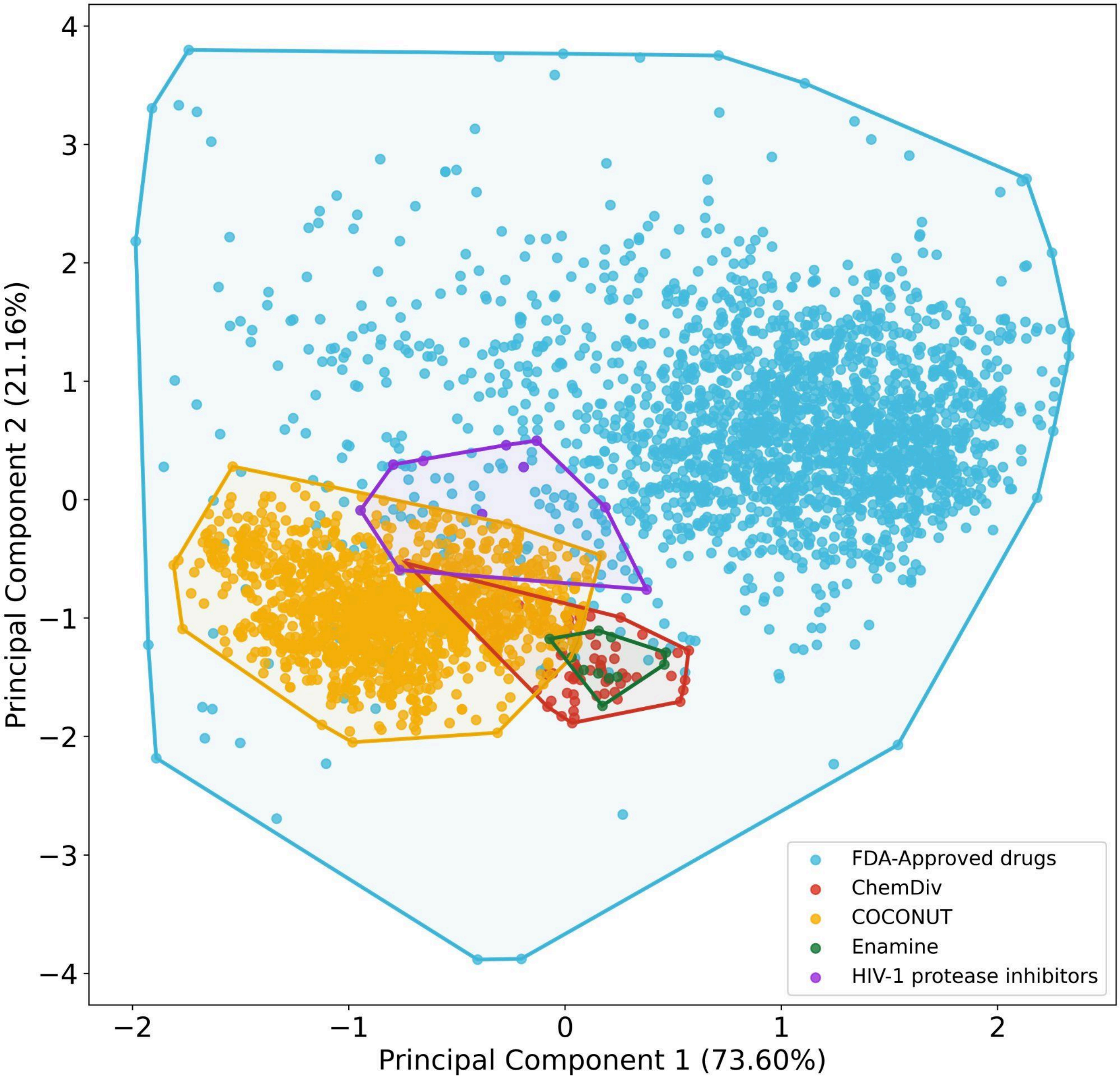


Figura 21. Visualización de espacio químico de la biblioteca de compuestos enfocados inhibidores de la proteasa viral del VIH-1 derivados de fragmentos de productos naturales y dos bibliotecas de compuestos de referencia utilizando PCA basado en propiedades fisicoquímicas. Las bibliotecas de compuestos de referencia se muestran en color azul (fármacos aprobados por la FDA), morado (inhibidores de la proteasa viral del VIH-1 aprobados por la FDA). Las bibliotecas de nuevos compuestos generados a partir de fragmentos moleculares se muestran en color naranja (COCONUT), rojo (ChemDiv) y verde (Enamine).

Se realizó un análisis de *convex hull* derivado del PCA para cada biblioteca de compuestos para definir cuantitativamente qué base de datos es la más diversa (ver Apéndice, **Figura A4**). El *convex hull* es definido como el polígono mínimo convexo, de modo que el conjunto de puntos se encuentre dentro de este polígono o en su frontera.[127,128] El área de *convex hull* calculada para las bases de datos fue fármacos aprobados (737.59), inhibidores de la proteasa viral del VIH-1 aprobados por la FDA (1.11), compuestos derivados de fragmentos de COCONUT (3.18), compuestos derivados de fragmentos de ChemDiv(0.79) y fragmentos derivados de fragmentos de Enamine (0.18). El resultado de este análisis fue similar a los resultados del análisis de diversidad basados en huellas digitales moleculares. Las bases de datos de referencia fueron las más diversas que las tres bibliotecas enfocadas de compuestos inhibidores de la proteasa viral del VIH-1 generadas a partir de fragmentos de COCONUT, ChemDiv y Enamine. Los compuestos químicos derivados de los fragmentos de COCONUT fueron los más diversos, seguidos por los compuestos químicos derivados de los fragmentos de ChemDiv y Enamine.

La **Figura 22** muestra una visualización de espacio químico basada en huellas digitales moleculares usando TMAP. La versión interactiva de esta visualización se encuentra disponible en el siguiente enlace https://figshare.com/s/ceb58d58e8f5585ce67e. Las estructuras químicas de las tres bibliotecas enfocadas de compuestos inhibidores de la proteasa viral del VIH-1 generadas a partir de fragmentos de COCONUT, ChemDiv y Enamine fueron muy diferentes en comparación a los fármacos aprobados por la FDA y los inhibidores de la proteasa viral del VIH-1 aprobados por la FDA. En algunos casos, las estructuras químicas de las bibliotecas enfocadas de compuestos inhibidores de la proteasa viral del VIH-1 generadas a partir de fragmentos de COCONUT tuvieron estructuras químicas muy similares a fármacos aprobados por la FDA como palbociclib y pipecuronium, lo cual sugiere un posible reposicionamiento de estos compuestos como inhibidores de la proteasa viral del VIH-1.

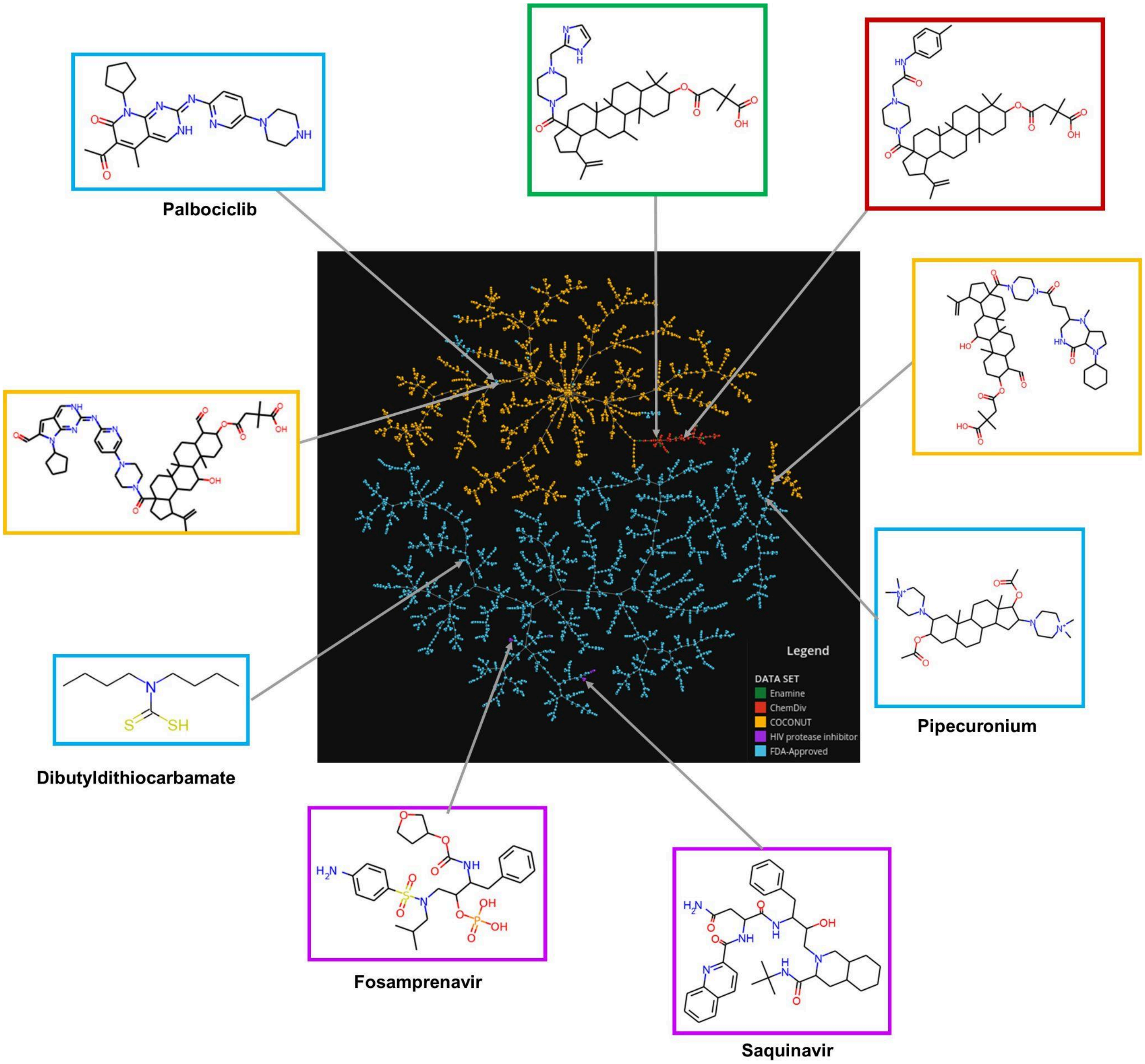


Figura 22. Visualización de espacio químico de la biblioteca de compuestos enfocados inhibidores de la proteasa viral del VIH-1 derivados de fragmentos de productos naturales y dos bibliotecas de compuestos de referencia utilizando TMAP basado en huellas digitales moleculares. Las bibliotecas de compuestos de referencia se muestran en color azul (fármacos aprobados por la FDA), morado (inhibidores de la proteasa viral del VIH-1 aprobados por la FDA). Las bibliotecas de nuevos compuestos generados a partir de fragmentos moleculares están en naranja (COCONUT), rojo (ChemDiv) y verde (Enamine). Disponible en https://rawcdn.githack.com/DIFACQUIM/De-novo-desing-of-HIV-1-inhibitors/45dedd8152b643a6f2884d88003f0c56bc48bef8/TMAP/TMAP_chemical_space_visualization.html (fecha de acceso: 15 de mayo de 2024).

8.2.3. Propiedades fisicoquímicas

Se calcularon seis propiedades fisicoquímicas de interés farmacéutico (MW, HBD, HBA, log P, TPSA y RB) para los inhibidores de la proteasa viral del VIH-1 aprobados por la FDA y las bibliotecas enfocadas de compuestos inhibidores de la proteasa viral del VIH-1 a partir de fragmentos de COCONUT, ChemDiv y Enamine. El valor máximo de las propiedades fisicoquímicas obtenidas para los inhibidores de la proteasa viral del VIH-1 fueron HBD ≤ 6, HBA ≤ 13, log P ≤ 6.7, MW ≤ 720.30, TPSA ≤ 174.60 y RB ≤ 17 (ver Apéndice, **Tabla A8**), y se utilizaron de referencia para construir una regla empírica (ver **sección 7.4**). 352 compuestos derivados de fragmentos de COCONUT (20 %) y un compuesto derivado de los fragmentos de ChemDiv (2 %) cumplieron con al menos cuatro condiciones, entre ellas log P, el filtro de regla empírica. La **Figura 23** muestra un gráfico de caja de las seis propiedades después de aplicar el filtro de reglas empíricas (ver **sección 7.4**).

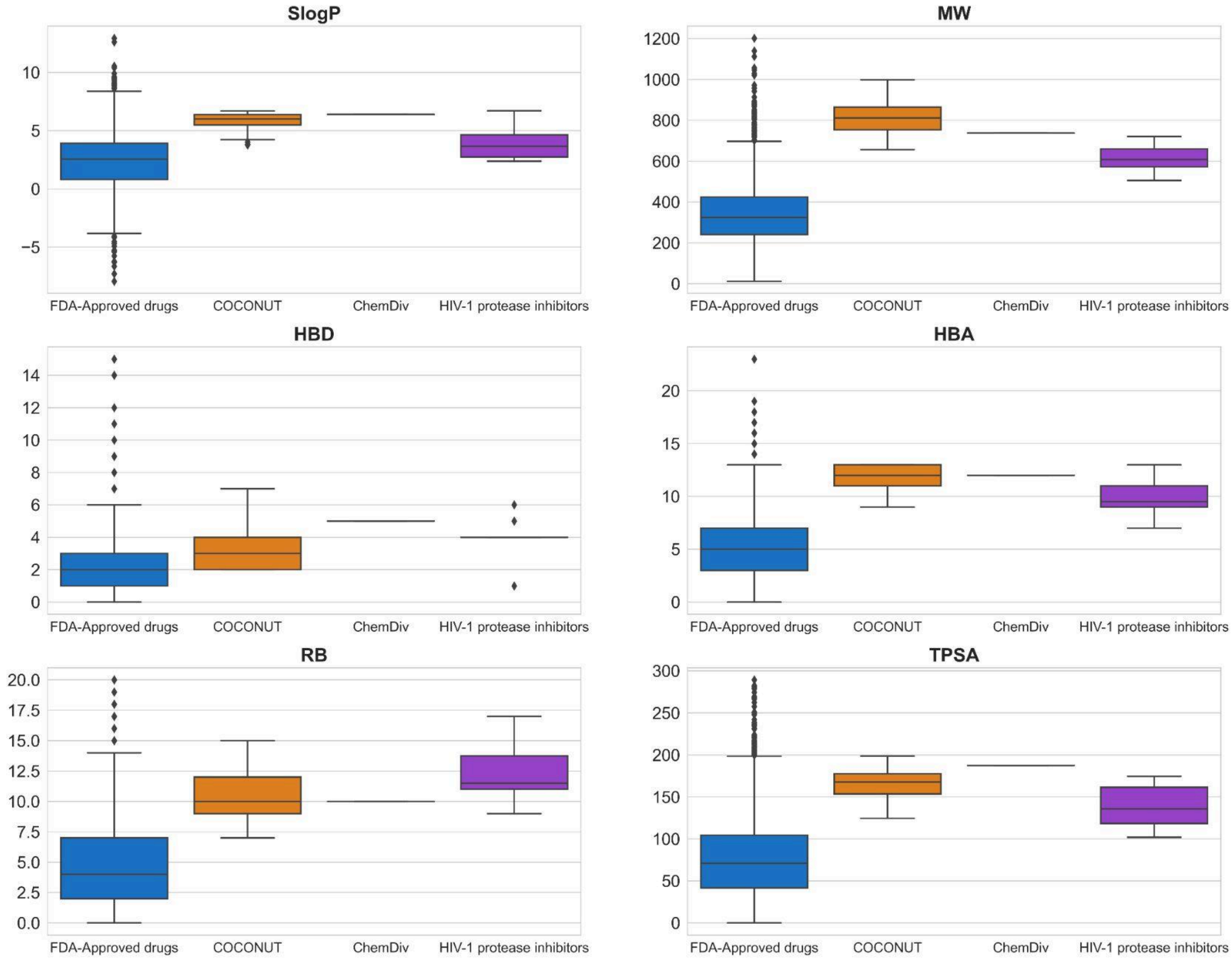


Figura 23. Gráfico de caja de seis propiedades fisicoquímicas de interés farmacéutico de fármacos aprobados por la FDA (azul), compuestos inhibidores de la proteasa viral del VIH-1 aprobados por la FDA (morado), y nuevos compuestos químicos generados a partir de fragmentos moleculares derivados de COCONUT (naranja) y un compuesto químico generado a partir de los fragmentos moleculares de ChemDiv (barra horizontal negra) después de aplicar el filtro de propiedades fisicoquímicas obtenidas para los inhibidores de la proteasa viral del VIH-1. Los diamantes negros muestras los puntos atípicos.

Las propiedades fisicoquímicas calculadas para las bases de datos fueron: log P ≤ 12.94, MW ≤ 1201.84, RB ≤ 20, TPSA ≤ 286.50, HBA ≤ 23 y HBD ≤ 15 para fármacos aprobados por la FDA, y log P ≤ 6.70, MW ≤ 720.31, RB ≤ 17, TPSA ≤ 174.56, HBA ≤ 13 y HBD ≤ 6 para los inhibidores de la proteasa viral del VIH-1 aprobados por la FDA. Mientras que las propiedades fisicoquímicas de los compuestos generados a partir de fragmentos de COCONUT que cumplieron con la regla empirica fueron log P ≤ 6.69, MW ≤ 998.63, RB ≤ 15, TPSA ≤ 198.54, HBA ≤ 13 y HBD ≤ 7, y las propiedades fisicoquímicas del compuesto derivado de los fragmentos de ChemDiv que cumplió con la regla empirica fueron log P = 6.4, MW = 737.47, RB = 10, TPSA= 187.47, HBA = 12 y HBD = 5.

Los valores de log P, RB y HBA de los compuestos generados a partir de los fragmentos de COCONUT y ChemDiv fue menor que el de los inhibidores de la proteasa viral del VIH-1 aprobados por la FDA. Los valores de MW, TPSA y HBD de compuestos generados a partir de fragmentos de COCONUT fueron menores que los inhibidores de la proteasa viral del VIH-1 aprobados por la FDA y menor que los valores de MW, TPSA y HBD calculados para los fármacos aprobados por la FDA. Ganesan *et al.* menciona que los productos naturales que violan la regla de Lipinski al menos cumplen en términos de log P y HBD.[129] Él considera que "la naturaleza ha aprendido a mantener un perfil bajo de hidrofobicidad y el potencial de donación de enlaces H intermoleculares cuando se necesita fabricar compuestos biológicamente activos con un peso molecular elevado y un gran número de enlaces rotables". La mayoría de los fármacos aprobados que superan el HBD = 5 o el HBA = 10 son derivados de productos naturales.[130]

8.2.4. Accesibilidad sintética

La accesibilidad sintética fue calculada para fármacos aprobados por la FDA, inhibidores de la proteasa viral del VIH-1 y la biblioteca enfocada de compuestos inhibidores de la proteasa viral del VIH-1 generada a partir de fragmentos de COCONUT, y ChemDiv que tienen propiedades fisicoquímicas similares a los inhibidores de la proteasa viral del VIH-1 aprobados por la FDA. La **Figura 24** muestra los resultados de accesibilidad sintética utilizando la función SAscore. El 97% de los fármacos aprobados por la FDA tuvieron un SAscore < 6, y los inhibidores de la proteasa viral del VIH-1 aprobados por la FDA tuvieron un SAscore ≤ 4.24. 65% de los compuestos generados a partir de fragmentos de COCONUT tuvieron un SAscore ≤ 6.03 y el compuesto generado a partir de ChemDIv tuvo un SAscore =5.54. Aunque los compuestos generados a partir de los fragmentos de COCONUT tuvieron 5.50 ≤ SAscore ≤ 6.03 se encuentran en el rango recomendado para ser accesibles sintéticamente. Además, el valor de SAscore fue mayor en los compuestos generados en relación con los inhibidores de la proteasa viral del VIH-1 aprobados por la FDA debido a los diez estereocentros del ácido betulínico y el ácido 24-nor-3α,11α-dihidroxi-lup-20(29)-en-23,28-dioico.

Teniendo en cuenta que estos estereocentros no tienen que generarse desde cero, el valor de SAscore sería menor, facilitando aún más la síntesis química de los compuestos propuestos.

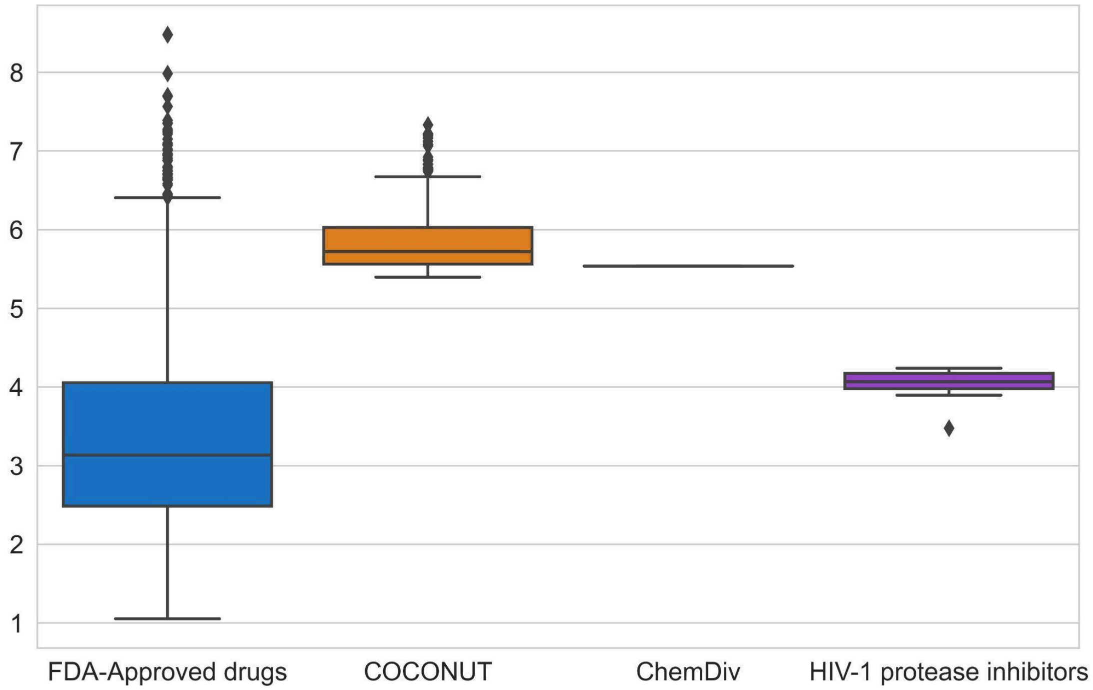


Figura 24. Gráfico de caja de accesibilidad sintética calculada para los fármacos probados por la FDA (azul), compuestos inhibidores de la proteasa viral del VIH-1 aprobados por la FDA (morado), y nuevos compuestos químicos generados a partir de fragmentos moleculares derivados de COCONUT (naranja) y ChemDiv (rojo) con propiedades fisicoquímicas similares a los inhibidores de la proteasa viral del VIH-1 aprobados por la FDA. Los diamantes negros muestras los puntos atípicos.

8.3. Subconjuntos de productos naturales

Debido al coste computacional que conlleva realizar una red neuronal para cerca de 153,000 moléculas, se seleccionaron subconjuntos con los compuestos más diversos del UNPD. La metodología para generar estos subconjuntos de productos naturales derivados del UNPD se publicó en la referencia.[131] Se generaron tres subconjuntos con 14,994, 7,497 y 4,998 estructuras de productos naturales únicos y derivados del UNPD utilizando el algoritmo MaxMin.[111] Los tres subconjuntos de productos naturales fueron nombrados como UNPD-A (14,994 compuestos), UNPD-B (7,497 compuestos) y UNPD-C (4,998 compuestos).

El UNPD, los subconjuntos de productos naturales (UNPD-A, UNPD-B, UNPD-C), y dos bases de datos de referencia, compuestos inhibidores de ADN metil transferasa 1 (DNMT1) y BIOFACQUIM fueron caracterizados por medio del cálculo computacional de la diversidad estructural y de seis propiedades fisicoquímicas de interés farmacéutico.

La diversidad estructural se evaluó utilizando una función de distribución comulativa (en inglés *commulative distribution function*, CDF) mostrada en la **Figura 25**. Esta figura muestra la similitud entre pares de moléculas utilizando como métrica el coeficiente de Tanimoto y tres huellas digitales moleculares como representación molecular: *Molecular ACCes System* (MACCS) *keys*[66] de 166-bits y *extended connectivity fingerprint* (ECFP) con 1024-bits[72] de diámetro 4 (ECFP4) y diámetro 6 (ECFP-6). Las bases de datos utilizadas son representadas en una línea continúa y usando diferentes colores: rosa (UNPD-A), verde (UNPD-B), rojo (UNPD-C), azul (UNPD), y dos bases de datos de referencia BIOFACQUIM, una base de datos de productos naturales aislados y caracterizados de diferentes centros de investigación en México[46,47] (amarillo) y un subconjunto de compuestos inhibidores de ADN metiltransferasa 1 (en inglés *DNA methyltransferase 1*, DNMT1), una diana biológica relevante en el tratamiento temprano del cancer[132,133], extraída de la base de datos ChEMBL[113,114] (cyan). Los compuestos más diversos usando CDF son aquellos cuya pendiente es más pronunciada, por esta razón, los subconjuntos UNPD-A, UNPD-B y UNPD-C tuvieron las estructuras químicas más diversas, seguido del UNPD, BIOFACQUIM y el subconjunto de DNMT1.

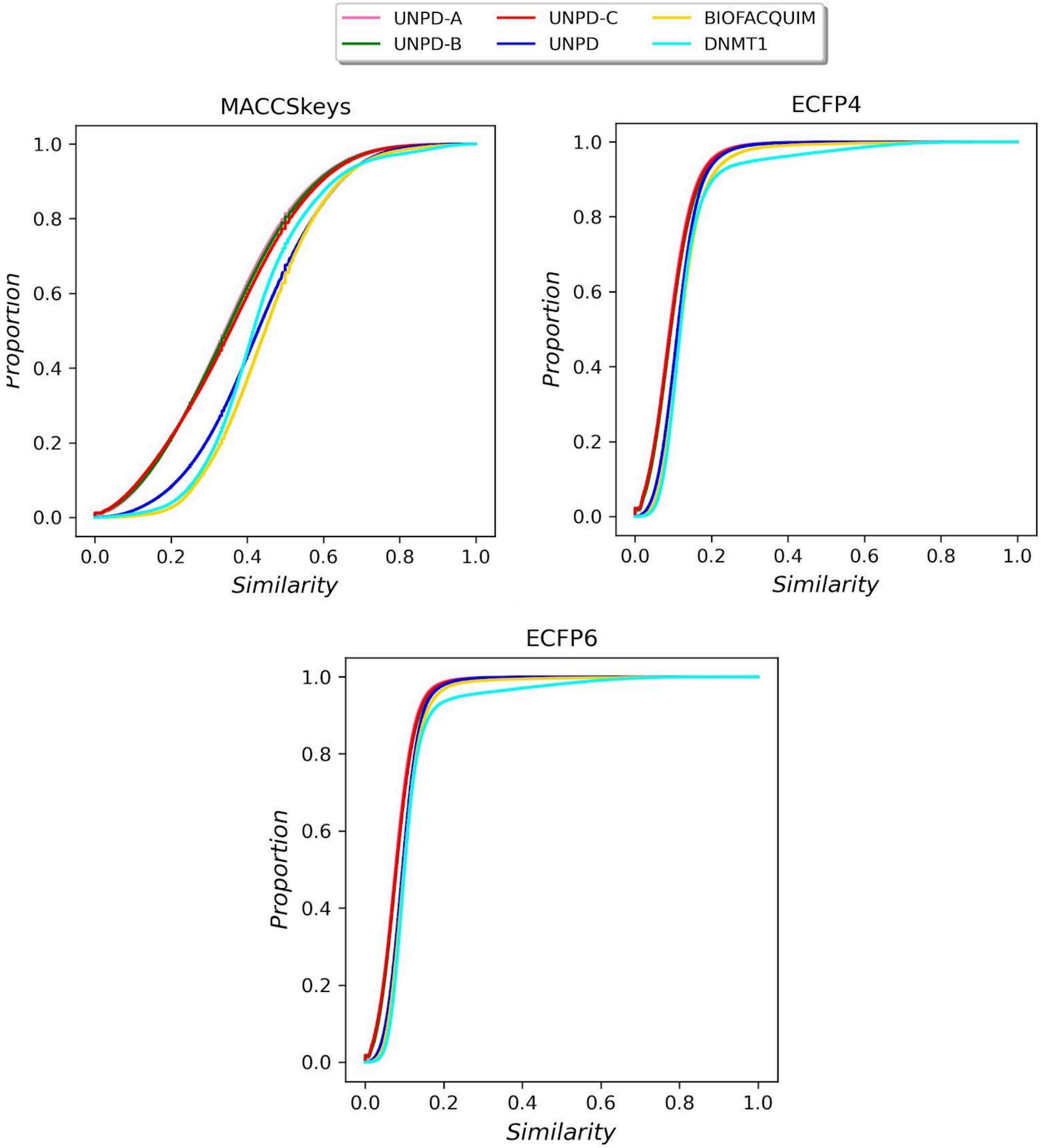


Figura 25. Función de distribución comulativa de similitud molecular de pares de moléculas utilizando el coeficiente de Tanimoto y como representación molecular las huellas digitales moleculares MACCS keys (166-bits), ECFP4 y ECFP6. Las bases de datos son representadas en una línea continúa y usando diferentes colores: UNPD-A (rosa), UNPD-B (verde), UNPD-C (rojo), UNPD (azul), BIOFACQUIM (amarillo) y DNMT1 (cyan).

Los subconjuntos del UNPD y todo el UNPD fue caracterizado evaluando la diversidad de los valores calculados de seis propiedades fisicoquímicas de interés farmacéutico como HBD, HBA, TPSA, RB, MW y log P. La **Figura 26** muestra un gráfico de caja con la distribución de estas propiedades fisicoquímicas donde las bases de datos son representadas en diferentes colores: UNPD-A (rosa), UNPD-B (verde), UNPD-C (rojo) y todo el UNPD (azul), y los valores atípicos son representados en diamantes negros. Tomando de referencia el tercer cuartil (75 % compuestos), la **Figura 26** muestra que los compuestos presentes en los tres subconjuntos del UNPD y en todo el UNPD tienen valores similares de log P (log P= 4.32-4.48). El UNPD-C tuvo valores de HBA y HBD más similares al UNPD (HBA=4 y para el RB, el UNPD-B fue más similar al UNPD (RB=7-8). La **Figura 27** muestra una visualización de espacio químico usando el algoritmo t-SNE). Se observa que los subconjuntos de compuestos derivados del UNPD cubren regiones similares del espacio químico cubierto por la base de datos original (UNPD) y son representativos del UNPD con respecto a sus propiedades de interés farmacéutico.

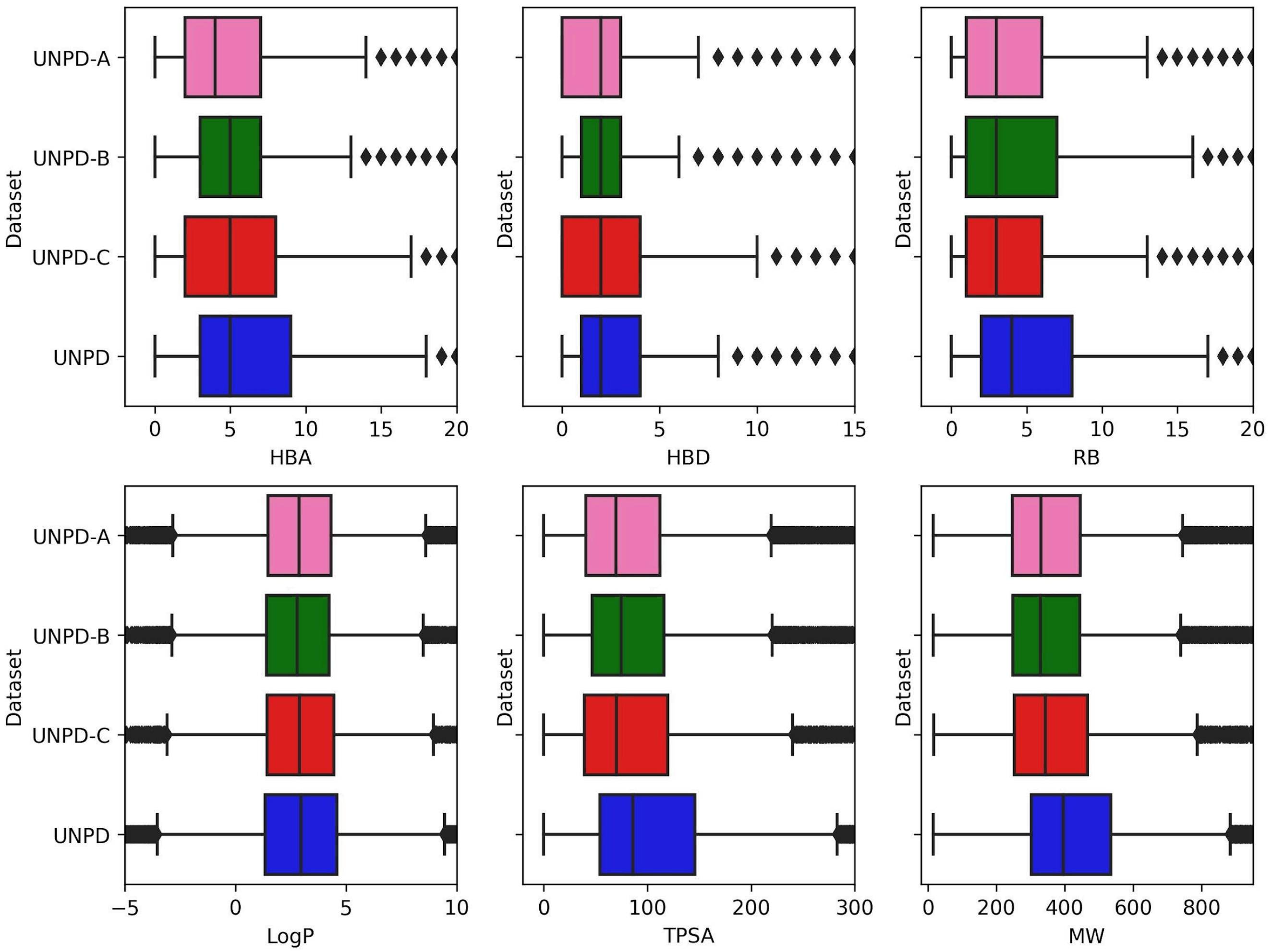


Figura 26. Gráfico de cajas de los subconjuntos del UNPD y el UNPD usando seis propiedades fisicoquímicas de interés farmacéutico: donadores de puente de hidrógeno (HBD), aceptores de puente de hidrógeno (HBA), área topológica superficial (TPSA), número de enlaces rotables (RB), peso molecular (MW) y coeficiente de partición octanol/agua (log P). Las bases de datos son representadas en diferentes colores: UNPD-A (rosa), UNPD-B (verde), UNPD-C (rojo), y todo el UNPD (azul). Los diamantes negros muestran los puntos atípicos.

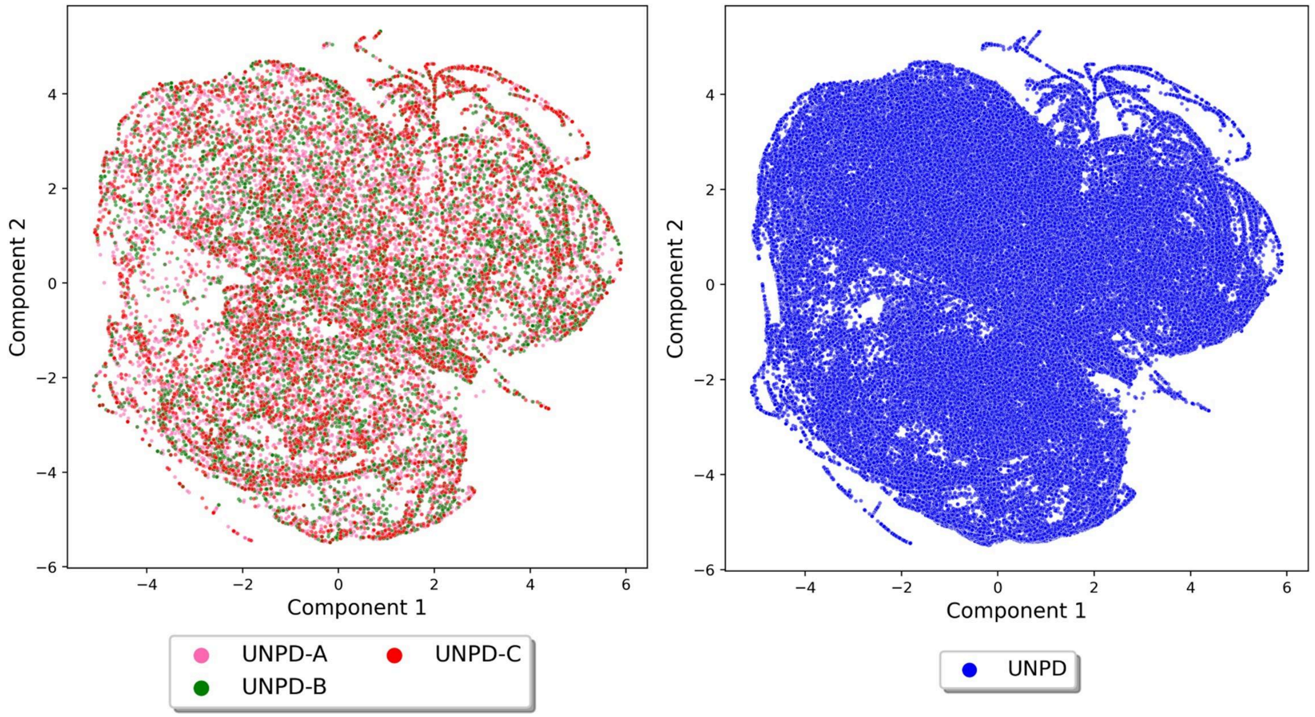


Figura 27. Visualización del espacio químico de los subconjuntos del UNPD y el UNPD usando el algoritmo t-SNE basado en seis propiedades fisicoquímicas de interés farmacéutico. Las bases de datos son representadas en diferentes colores: UNPD-A (rosa), UNPD-B (verde), UNPD-C (rojo) y todo el UNPD (azul).

Los tres subconjuntos generados del UNPD tienen estructuras químicas muy diversas. El subconjunto de UNPD-C tiene valores de propiedades fisicoquímicas de interés farmacéutico muy similares al UNPD y es recomendable usarlo para entrenar redes neuronales. Los tres subconjuntos generados del UNPD se pueden utilizar para entrenar redes neuronales de aprendizaje profundo en grupos de investigación con recursos computacionales limitados.

IX. Conclusiones

Se desarrollaron, analizaron e hicieron públicas las bibliotecas de fragmentos de productos naturales de COCONUT y compuestos químicos de alimentos de FooDB. También se hicieron los mismos análisis para bibliotecas de referencia de los fragmentos de compuestos accesibles sintéticamente (REAL), compuestos de relevancia biológica (ChEMBL), compuestos sin actividad biológica reportada (DCM), compuestos enfocados al COVID-19 (CAS) y compuestos inhibidores de la principal proteasa de SARS-CoV-2 (3CLP). Los resultados de estos análisis están publicados.[118,126] Análisis quimioinformáticos de las bases de datos mostraron que los compuestos y los fragmentos moleculares derivados de productos naturales y compuestos químicos de alimentos fueron los más diversos (utilizando las huellas digitales moleculares ECFP-4 y MACCC keys) y complejos (utilizando la fracción de carbonos con hibridación sp^3 y la fracción de carbonos quirales) en comparación a los compuestos accesibles sintéticamente, compuestos con relevancia biológica, bibliotecas enfocadas a COVID-19, y compuestos sin actividad biológica reportada. Los fragmentos moleculares y compuestos de productos naturales, compuestos sintéticos y compuestos con actividad biológica conservaron la diversidad y complejidad estructural de los compuestos originales.

El 83.1 % de compuestos de COCONUT y el 72.2 % de los fragmentos moleculares de COCONUT fueron únicos, y no están cubiertos totalmente por los compuestos accesibles sintéticamente o los compuestos probados de relevancia biológica. Los fragmentos moleculares de COCONUT cumplieron con un perfil tipo fármaco (es decir, cumplieron con las reglas de Lipinski y Verber, y no contienen PAINS), y podrían ser utilizados en el diseño de nuevos candidatos a fármaco.

Los fragmentos moleculares de DCM convergen en varias regiones del espacio químico cubierto por los fragmentos moleculares de CAS, lo que refuerza la premisa de que la base de datos de DCM es una fuente importante de bloques de construcción para el diseño de nuevas moléculas bioactivas.

A partir de los fragmentos de COCONUT, se generó una biblioteca virtual enfocada en compuestos inhibidores de la proteasa viral del VIH-1. Los compuestos fueron generados utilizando de referencia a bevirimat, un inhibidor de la proteasa viral del VIH-1. 251 compuestos de 1,534 compuestos generados a partir fragmentos de COCONUT tuvieron propiedades fisicoquímicas similares a los inhibidores de la proteasa viral del VIH-1 aprobados por la FDA y se estimaron viables sintéticamente, cuyos resultados se publicaron en.[134]

Se realizó una revisión bibliográfica sobre la metodología de diseño *de novo* utilizando algoritmos de inteligencia artificial y se publicó en.[23] Se encontró que una aproximación inicial para el diseño *de novo* u otro modelo de aprendizaje de máquina es comenzar con un subconjunto de datos con estructuras químicas y propiedades fisicoquímicas diversas; para lo cual se generaron y caracterizaron tres subconjuntos de productos naturales con 14,994, 7,497 y 4,998 compuestos derivados del UNPD utilizando el algoritmo MaxMin, y se publicaron en.[131] Se encontró que los tres subconjuntos tuvieron estructuras químicas muy diversas y propiedades fisicoquímicas de interés farmacéutico muy similares a los compuestos del UNPD (base de datos original). Los subconjuntos del UNPD pueden utilizarse para entrenar redes neuronales de aprendizaje profundo en grupos de investigación con recursos computacionales limitados.

Recursos informáticos generales / aportaciones del trabajo Doctoral

Como parte de la tesis doctoral, se desarrollaron varios códigos y scripts de acceso libre que se resumen en la **Tabla 8**. Estas herramientas informáticas son generales y se pueden emplear en otros proyectos, como se describe en la **Tabla 8**.

Tabla 8. Resumen de códigos y recursos libres generados.

Título de la herramienta	Aplicación o uso	Vínculo (URL)
Bibliotecas de fragmentos basadas en productos naturales.	Diseño *de novo* o híbridos de fragmentos basados en productos naturales.	https://doi.org/10.6084/m9.figshare.11997951 https://doi.org/10.6084/m9.figshare.13064231.v1
El código utilizado para construir compuestos análogos a bevirimat.	Construcción de compuestos análogos a bevirimat u otros compuestos derivados del ácido betulínico.	https://github.com/DIFACQUIM/De-novo-desing-of-HIV-1-inhibitors

Tabla 8 (continuación). Resumen de códigos y recursos libres generados.

Título de la herramienta	Aplicación o uso	Vínculo (URL)
Visualización de espacio químico de la biblioteca de compuestos enfocados inhibidores de la proteasa viral del VIH-1 derivados de fragmentos de productos naturales y fármacos aprobados por la FDA utilizando TMAP basado en huellas digitales moleculares.	Las estructuras químicas pueden ser analizadas y utilizadas por los químicos farmacéuticos para sintetizar análogos de posibles inhibidores de la proteasa viral del VIH-1 u otros antivirales.	https://rawcdn.githack.com/DIFACQUIM/De-novo-desing-of-HIV-1-inhibitors/45dedd8152b643a6f2884d88003f0c56bc48bef8/TMAP/TMAP_chemical_space_visualization.html
Visualizaciones de espacio químico utilizando el PCA, t-SNE y TMAP.	Generación de visualizaciones de espacio químico utilizando otras bases de datos moleculares y descriptores moleculares. Se generan visualizaciones de TMAP interactivas (ver **Figura 22**).	https://github.com/DIFACQUIM/Art-Driven-by-Visual-Representations-of-Chemical-Space-
Visualizaciones interactivas de espacio químico utilizando PCA y t-SNE.	Construcción de visualizaciones de espacio químico utilizando otras bases de datos y descriptores moleculares.	https://github.com/DIFACQUIM/Cursos/blob/main/07_Espacio_Qu%C3%ADmico_PCA.ipynb https://github.com/DIFACQUIM/Cursos/blob/main/07_Espacio_Qu%C3%ADmico_tSNE.ipynb
Subconjuntos del UNPD y script de Python para generar subconjuntos de bases de datos utilizando el algoritmo MaxMin.	Generación de subconjuntos de bases de datos utilizando el algoritmo MaxMin. Los subconjuntos de productos naturales de UNPD pueden utilizarse para entrenar redes neuronales de aprendizaje profundo en grupos de investigación con recursos computacionales limitados.	https://github.com/DIFACQUIM/Natural-products-subsets-generation.

Tabla 8 (continuación). Resumen de códigos y recursos libres generados.

Título de la herramienta	Aplicación o uso	Vínculo (URL)
Código para generar grafos moleculares	Generación de grafos moleculares de productos naturales, fármacos aprobados, compuestos sintéticos y compuestos organometálicos. Los grafos moleculares pueden ser utilizados en la construcción, entrenamiento y prueba de arquitecturas de redes neuronales para proponer compuestos a partir del diseño *de novo* utilizando algoritmos de inteligencia artificial.	https://github.com/DIFACQUIM/HANNA

X. Perspectivas

Utilizar las bibliotecas de fragmentos obtenidos a partir de productos naturales en futuras investigaciones para el diseño *de novo* o híbridos de fragmentos basados en productos naturales.

Realizar un cribado virtual de los 251 compuestos generados a partir fragmentos de COCONUT que tuvieron propiedades fisicoquímicas similares a los inhibidores de la proteasa viral del VIH-1 aprobados por la FDA y se estimaron viables sintéticamente. Realizar la síntesis química y la evaluación biológica de los compuestos seleccionados del cribado virtual.

Utilizar el protocolo, el código, los SMARTS y SMIRKS para construir compuestos análogos a bevirimat u otros compuestos derivados del ácido betulínico.

Los subconjuntos de productos naturales del UNPD pueden utilizarse para entrenar redes neuronales de aprendizaje profundo en grupos de investigación con recursos computacionales limitados.

XI. Referencias

(1) Dewick P. Secondary Metabolism: The Building Blocks and Construction Mechanisms. In *Medicinal Natural Products*; John Wiley & Sons, Ltd: Chichester, UK, 2009; pp 7–38.

(2) Pahl A, Waldmann H, Kumar K. Exploring Natural Product Fragments for Drug and Probe Discovery. *Chimia* **2017**, *71*, 653–660.

(3) Beutler JA. Natural Products as a Foundation for Drug Discovery. *Curr. Protoc. Pharmacol.* **2019**, *86*, e67.

(4) Feher M, Schmidt JM. Property Distributions: Differences between Drugs, Natural Products, and Molecules from Combinatorial Chemistry. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 218–227.

(5) Campos DS. Medicamentos, Plantas Medicinales Y Productos Naturales. *fármacos* **2003**, *16*, 13–20.

(6) Rishton GM. Natural Products as a Robust Source of New Drugs and Drug Leads: Past Successes and Present Day Issues. *Am. J. Cardiol.* **2008**, *101*, 43D – 49D.

(7) Sneader W. The Discovery of Aspirin: A Reappraisal. *BMJ* **2000**, *321*, 1591–1594.

(8) Fleming A. On the Antibacterial Action of Cultures of a Penicillium, with Special Reference to Their Use in the Isolation of B. Influenzæ. *Br. J. Exp. Pathol.* **1929**, *10*, 226.

(9) *Penicillin G*. https://pubchem.ncbi.nlm.nih.gov/compound/Penicillin-G (accessed 2023-12-29).

(10) Suffness M. Chapter 32. Taxol: From Discovery to Therapeutic Use. In *Annual Reports in Medicinal Chemistry*; Bristol, J. A., Ed.; Academic Press, 1993; Vol. 28, pp 305–314.

(11) Newman DJ, Cragg GM. Natural Products as Sources of New Drugs over the Nearly Four Decades from 01/1981 to 09/2019. *J. Nat. Prod.* **2020**, *83*, 770–803.

(12) Atanasov AG, Zotchev SB, Dirsch VM. International Natural Product Sciences Taskforce; Supuran, C. T. Natural Products in Drug Discovery: Advances and Opportunities. *Nat. Rev. Drug Discov.* **2021**, *20*, 200–216.

(13) *Plitidepsin*. https://pubchem.ncbi.nlm.nih.gov/compound/Plitidepsin (accessed 2023-12-28).

(14) *Moxidectin*. https://pubchem.ncbi.nlm.nih.gov/compound/Moxidectine (accessed 2023-12-27).

(15) *Lefamulin*. https://pubchem.ncbi.nlm.nih.gov/compound/Lefamulin (accessed 2023-12-28).

(16) Ertl P. Cheminformatics Analysis of Organic Substituents: Identification of the Most Common Substituents, Calculation of Substituent Properties, and Automatic Identification of Drug-like Bioisosteric Groups. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 374–380.

(17) Erlanson DA, Fesik SW, Hubbard RE, Jahnke W, Jhoti H. Twenty Years on: The Impact of Fragments on Drug Discovery. *Nat. Rev. Drug Discov.* **2016**, *15*, 605–619.

(18) Erlanson DA, de Esch IJP, Jahnke W, Johnson CN, Mortenson PN. Fragment-to-Lead Medicinal Chemistry Publications in 2018. *J. Med. Chem.* **2020**, *63*, 4430–4444.

(19) Congreve M, Carr R, Murray C, Jhoti H. A “Rule of Three” for Fragment-Based Lead Discovery? *Drug Discov. Today* **2003**, *8*, 876–877.

(20) Hartenfeller M, Schneider G. De Novo Drug Design. *Methods Mol. Biol.* **2011**, *672*, 299–323.

(21) Mouchlis VD, Afantitis A, Serra A, Fratello M, Papadiamantis AG, Aidinis V, Lynch I, Greco D, Melagraki G. Advances in de Novo Drug Design: From Conventional to Machine Learning Methods. *Int. J. Mol. Sci.* **2021**, *22*, 1676.

(22) Saldívar González FI, Chávez Ponce de León DE, López López E, Hernández Luis F, Lira Rocha A, Medina Franco JL. Bases de Datos: Bibliotecas de Compuestos Químicos. In *Manual de Quimioinformática*; UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO, 2018; pp 8–26.

(23) Chávez-Hernández AL., López-López E, Medina-Franco JL. Yin-Yang in Drug Discovery: Rethinking de Novo Design and Development of Predictive Models. *Front. Drug Discov. (Lausanne)* **2023**, *3*, 1222655.

(24) Yang J, Wang D, Jia C, Wang M, Hao G, Yang G. Freely Accessible Chemical Database Resources of Compounds for In Silico Drug Discovery. *Curr. Med. Chem.* **2019**, *26*, 7581–7597.

(25) Warr WA, Nicklaus MC, Nicolaou CA, Rarey M. Exploration of Ultralarge Compound Collections for Drug Discovery. *J. Chem. Inf. Model.* **2022**, *62*, 2021–2034.

(26) Korn M, Ehrt C, Ruggiu F, Gastreich M, Rarey M. Navigating Large Chemical Spaces in Early-Phase Drug Discovery. *Curr. Opin. Struct. Biol.* **2023**, *80*, 102578.

(27) Wu Z, Ramsundar B, Feinberg EN, Gomes J, Geniesse C, Pappu AS, Leswing K, Pande V. MoleculeNet: A Benchmark for Molecular Machine Learning. *Chem. Sci.* **2018**, *9*, 513–530.

(28) Brown N, Fiscato M, Segler MHS, Vaucher AC. GuacaMol: Benchmarking Models for de Novo Molecular Design. *J. Chem. Inf. Model.* **2019**, *59*, 1096–1108.

(29) Polykovskiy D, Zhebrak A, Sanchez-Lengeling B, Golovanov S, Tatanov O, Belyaev S, Kurbanov R, Artamonov A, Aladinskiy V, Veselov M, Kadurin A, Johansson S, Chen H, Nikolenko S, Aspuru-Guzik A, Zhavoronkov A. Molecular Sets (MOSES): A Benchmarking Platform for Molecular Generation Models. *Front. Pharmacol.* **2020**, *11*, 565644.

(30) Réau M, Langenfeld F, Zagury JF, Lagarde N, Montes M. Decoys Selection in Benchmarking Datasets: Overview and Perspectives. *Front. Pharmacol.* **2018**, *9*, 11.

(31) Gómez-García A, Medina-Franco JL. Progress and Impact of Latin American Natural Product Databases. *Biomolecules* **2022**, *12*, 1202.

(32) Saldívar-González FI, Aldas-Bulos VD, Medina-Franco JL, Plisson F. Natural Product Drug Discovery in the Artificial Intelligence Era. *Chem. Sci.* **2022**, *13*, 1526–1546.

(33) Medina-Franco JL, Flores-Padilla EA, Chávez-Hernández AL. Chapter 23 - Discovery and Development of Lead Compounds from Natural Sources Using Computational Approaches. In *Evidence-Based Validation of Herbal Medicine (Second Edition)*; Mukherjee, P. K., Ed.; Elsevier, 2022; pp 539–560.

(34) Gallo K. Kemmler E, Goede A, Becker F, Dunkel M, Preissner R, Banerjee P. SuperNatural 3.0-a Database of Natural Products and Natural Product-Based Derivatives. *Nucleic Acids Res.*

2023, *51*, D654–D659.

(35) Sorokina M, Merseburger P, Rajan K, Yirik MA, Steinbeck C. COCONUT Online: Collection of Open Natural Products Database. *J. Cheminform.* **2021**, *13*, 2.

(36) 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.

(37) Chen YC. TCM Database@Taiwan: The World's Largest Traditional Chinese Medicine Database for Drug Screening in Silico. *PLoS One* **2011**, *6*, e15939.

(38) Mohanraj K, Karthikeyan BS, Vivek-Ananth RP, Chand RPB, Aparna SR, Mangalapandi P, Samal A. IMPPAT: A Curated Database of Indian Medicinal Plants, Phytochemistry And Therapeutics. *Sci. Rep.* **2018**, *8*, 4329.

(39) Ntie-Kang F, Zofou D, Babiaka SB, Meudom R, Scharfe M, Lifongo LL, Mbah JA, Mbaze LM, Sippl W, Efange SMN. AfroDb: A Select Highly Potent and Diverse Natural Product Library from African Medicinal Plants. *PLoS One* **2013**, *8*, e78085.

(40) Gómez-García A, Jiménez DAA, Zamora WJ, Barazorda-Ccahuana HL, Chávez-Fumagalli MA, Valli M, Andricopulo AD, Bolzani V, da S, Olmedo DA, Solís PN, Núñez MJ, Rodríguez Pérez JR, Valencia Sánchez HA, Cortés Hernández HF, Medina-Franco JL. Navigating the Chemical Space and Chemical Multiverse of a Unified Latin American Natural Product Database: LANaPDB. *Pharmaceuticals* **2023**, *16*, 1388.

(41) 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.

(42) Scotti MT, Herrera-Acevedo C, Oliveira TB, Costa RPO, Santos SYKO, Rodrigues RP, Scotti L, Da-Costa FB. SistematX, an Online Web-Based Cheminformatics Tool for Data Management of Secondary Metabolites. *Molecules* **2018**, *23*, 103.

(43) Olmedo DA, González-Medina M, Gupta MP, Medina-Franco JL. Cheminformatic Characterization of Natural Products from Panama. *Mol. Divers.* **2017**, *21*, 779–789.

(44) A. Olmedo, D.; L. Medina-Franco, J. Chemoinformatic Approach: The Case of Natural Products of Panama. In *Cheminformatics and its Applications*; IntechOpen, 2020.

(45) Barazorda-Ccahuana HL, Ranilla LG, Candia-Puma MA, Cárcamo-Rodriguez EG, Centeno-Lopez AE, Davila-Del-Carpio G, Medina-Franco JL, Chávez-Fumagalli MA. PeruNPDB: The Peruvian Natural Products Database for in Silico Drug Screening. *Sci. Rep.* **2023**, *13*, 7577.

(46) Pilón-Jiménez BA, Saldívar-González FI, Díaz-Eufracio BI, Medina-Franco JL. BIOFACQUIM: A Mexican Compound Database of Natural Products. *Biomolecules* **2019**, *9*, 31.

(47) Sánchez-Cruz N, Pilón-Jiménez BA, Medina-Franco JL. Functional Group and Diversity Analysis of BIOFACQUIM: A Mexican Natural Product Database. *F1000Res.* **2019**, *8*, Chem Inf

Sci-2071.

(48) *UNIIQUIM*. UNIIQUIM. https://uniiquim.iquimica.unam.mx/ (accessed 2023-05-13).

(49) David L, Thakkar A, Mercado R, Engkvist O. Molecular Representations in AI-Driven Drug Discovery: A Review and Practical Guide. *J. Cheminform.* **2020**, *12*, 56.

(50) Saldívar-González FI, Huerta-García CS, Medina-Franco JL. Chemoinformatics-Based Enumeration of Chemical Libraries: A Tutorial. *J. Cheminform.* **2020**, *12*, 64.

(51) Weininger D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31–36.

(52) *Inc D. Daylight Theory: SMARTS-A Language for describing molecular patterns*. https://www.daylight.com/dayhtml/doc/theory/theory.smarts.html (accessed 2023-12-28).

(53) Heller SR, McNaught A, Pletnev I, Stein S, Tchekhovskoi D. InChI, the IUPAC International Chemical Identifier. *J. Cheminform.* **2015**, *7*, 23.

(54) *5. SMIRKS - A Reaction Transform Language*. Daylight Chemical Information Systems, Inc. https://www.daylight.com/dayhtml/doc/theory/theory.smirks.html (accessed 2024-01-18).

(55) O'Donnell TJ. *Design and Use of Relational Databases in Chemistry*; CRC Press, 2008.

(56) Simonovsky M, Komodakis N. GraphVAE: Towards Generation of Small Graphs Using Variational Autoencoders. In *Artificial Neural Networks and Machine Learning – ICANN 2018*; Springer International Publishing, 2018; pp 412–422.

(57) Fey M, Lenssen JE. Fast Graph Representation Learning with PyTorch Geometric. *arXiv.org* **2019**, doi:arXiv:1903.02428v3.

(58) *PyG Documentation*. PyG (PyTorch Geometric). https://pytorch-geometric.readthedocs.io/en/latest/.

(59) Todeschini R, Consonni V. *Handbook of Molecular Descriptors*; Wiley & Sons, Limited, John, 2008.

(60) Todeschini R, Consonni V, Mannhold R, Kubinyi H, Folkers G. *Molecular Descriptors for Chemoinformatics : Volume I : Alphabetical Listing / Volume II: Appendices, References*; Wiley & Sons, Limited, John, 2010.

(61) Méndez-Lucio O, Medina-Franco JL. The Many Roles of Molecular Complexity in Drug Discovery. *Drug Discov. Today* **2017**, *22*, 120–126.

(62) Oprea TI, Bologa C. Molecular Complexity: You Know It When You See It. *J. Med. Chem.* **2023**, *66*, 12710–12714.

(63) Bajorath J. Selected Concepts and Investigations in Compound Classification, Molecular Descriptor Analysis, and Virtual Screening. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 233–245.

(64) Seo M, Shin HK, Myung Y, Hwang S, No KT. Development of Natural Compound Molecular Fingerprint (NC-MFP) with the Dictionary of Natural Products (DNP) for Natural Product-Based Drug Development. *J. Cheminform.* **2020**, *12*, 6.

(65) Cereto-Massagué A, Ojeda MJ, Valls C, Mulero M, Garcia-Vallvé S, Pujadas G. Molecular Fingerprint Similarity Search in Virtual Screening. *Methods* **2015**, *71*, 58–63.

(66) Durant JL, Leland BA, Henry DR, Nourse JG. Reoptimization of MDL Keys for Use in Drug Discovery. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1273–1280.

(67) Bolton EE, Wang Y, Thiessen PA, Bryant SH. Chapter 12 - PubChem: Integrated Platform of Small Molecules and Biological Activities. In *Annual Reports in Computational Chemistry*; Wheeler, R. A., Spellmeyer, D. C., Eds.; Elsevier, 2008; Vol. 4, pp 217–241.

(68) *6. Fingerprints - Screening and Similarity*. Daylight Chemical Information Systems, Inc. https://www.daylight.com/dayhtml/doc/theory/theory.finger.html (accessed 2024-01-04).

(69) *Daylight Fingerprints*. Daylight Chemical Information Systems Inc. https://www.daylight.com/meetings/summerschool01/course/basics/fp.html (accessed 2024-01-04).

(70) *Fingerprint Generation*. OpenEye scientific, Cadence Design Systems, Inc. https://docs.eyesopen.com/toolkits/python/graphsimtk/fingerprint.html (accessed 2024-01-04).

(71) Bender A, Mussa HY, Glen RC, Reiling S. Molecular Similarity Searching Using Atom Environments, Information-Based Feature Selection, and a Naïve Bayesian Classifier. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 170–178.

(72) Rogers D. Hahn M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742–754.

(73) Morgan HL. The Generation of a Unique Machine Description for Chemical Structures-A Technique Developed at Chemical Abstracts Service. *J. Chem. Doc.* **1965**, *5*, 107–113.

(74) Probst, D.; Reymond, J.-L. A Probabilistic Molecular Fingerprint for Big Data Settings. *J. Cheminform.* **2018**, *10*, 66.

(75) Capecchi A, Probst D, Reymond JL. One Molecular Fingerprint to Rule Them All: Drugs, Biomolecules, and the Metabolome. *J. Cheminform.* **2020**, *12*, 43.

(76) Deng Z, Chuaqui C, Singh J. Structural Interaction Fingerprint (SIFt): A Novel Method for Analyzing Three-Dimensional Protein−Ligand Binding Interactions. *J. Med. Chem.* **2004**, *47*, 337–344.

(77) Axen SD, Huang XP, Cáceres EL, Gendelev L, Roth BL, Keiser MJ. A Simple Representation of Three-Dimensional Molecular Structure. *J. Med. Chem.* **2017**, *60*, 7393–7409.

(78) López-López E, Bajorath J, Medina-Franco JL. Informatics for Chemistry, Biology, and Biomedical Sciences. *J. Chem. Inf. Model.* **2021**, *61*, 26–35.

(79) Muratov EN, Bajorath J, Sheridan RP, Tetko IV, Filimonov D, Poroikov V, Oprea TI, Baskin II, Varnek A, Roitberg A, Isayev O, Curtalolo S, Fourches D, Cohen Y, Aspuru-Guzik A, Winkler DA, Agrafiotis D, Cherkasov A, Tropsha A. QSAR without Borders. *Chem. Soc. Rev.* **2020**, *49*, 3525–3564.

(80) Shi J, Zhao G, Wei Y. Computational QSAR Model Combined Molecular Descriptors and Fingerprints to Predict HDAC1 Inhibitors. *Med Sci* **2018**, *34*, 52–58.

(81) Kadurin A, Aliper A, Kazennov A, Mamoshina P, Vanhaelen Q, Khrabrov K, Zhavoronkov A. The Cornucopia of Meaningful Leads: Applying Deep Adversarial Autoencoders for New Molecule Development in Oncology. *Oncotarget* **2017**, *8*, 10883–10890.

(82) Krenn M, Häse F, Nigam A, Friederich P, Aspuru-Guzik A. Self-Referencing Embedded Strings (SELFIES): A 100% Robust Molecular String Representation. *Mach. Learn.: Sci. Technol.* **2020**, *1*, 045024.

(83) Lim J, Ryu S, Kim JW, Kim WY. Molecular Generative Model Based on Conditional Variational Autoencoder for de Novo Molecular Design. *J. Cheminform.* **2018**, *10*, 31.

(84) Virshup AM, Contreras-García J, Wipf P, Yang W, Beratan DN. Stochastic Voyages into Uncharted Chemical Space Produce a Representative Library of All Possible Drug-Like Compounds. *J. Am. Chem. Soc.* **2013**, *135*, 7296–7303.

(85) Saldívar-González FI, Medina-Franco JL. Approaches for Enhancing the Analysis of Chemical Space for Drug Discovery. *Expert Opin. Drug Discov.* **2022**, *17*, 789–798.

(86) Medina-Franco JL, Chávez-Hernández, AL, López-López E, Saldívar-González FI. Chemical Multiverse: An Expanded View of Chemical Space. *Mol. Inform.* **2022**, *41*, e2200116.

(87) Greener JG, Kandathil SM, Moffat L, Jones DT. A Guide to Machine Learning for Biologists. *Nat. Rev. Mol. Cell Biol.* **2022**, *23*, 40–55.

(88) Gmail L, Hinton G. Visualizing Data Using T-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.

(89) Lam KS. New Aspects of Natural Products in Drug Discovery. *Trends Microbiol.* **2007**, *15*, 279–289.

(90) Cremosnik GS, Liu J, Waldmann H. Guided by Evolution: From Biology Oriented Synthesis to Pseudo Natural Products. *Nat. Prod. Rep.* **2020**, *37*, 1497–1510.

(91) Christoforow A, Wilke J, Binici A, Pahl A, Ostermann C, Sievers S, Waldmann H. Design, Synthesis, and Phenotypic Profiling of Pyrano-Furo-Pyridone Pseudo Natural Products. *Angew. Chem. Int. Ed Engl.* **2019**, *58*, 14715–14723.

(92) Ceballos J, Schwalfenberg M, Karageorgis G, Reckzeh ES, Sievers S, Ostermann C, Pahl A, Sellstedt M, Nowacki J, Carnero Corrales MA, Wilke J, Laraia L, Tschapalda K, Metz M, Sehr DA, Brand S, Winklhofer K, Janning P, Ziegler S, Waldmann H. Synthesis of Indomorphan Pseudo-Natural Product Inhibitors of Glucose Transporters GLUT-1 and -3. *Angew. Chem. Int. Ed Engl.* **2019**, *58*, 17016–17025.

(93) Fourches D, Muratov E, Tropsha A. Trust, but Verify II: A Practical Guide to Chemogenomics Data Curation. *J. Chem. Inf. Model.* **2016**, *56*, 1243–1252.

(94) Weininger D, Weininger A, Weininger JL. SMILES. 2. Algorithm for Generation of Unique SMILES Notation. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 97–101.

(95) *RDKit*. https://www.rdkit.org (accessed 08 January 08 2022).

(96) *MolVS*. https://molvs.readthedocs.io/en/latest/ (accessed 08 accessed January 2022).

(97) Jaccard P. Etude Comparative de La Distribution Florale Dans Une Portion Des Alpes et Des Jura. *Bull. Soc. Vaud. sci. nat.* **1901**, *37*, 547–579.

(98) Probst D, Reymond JL. Visualization of Very Large High-Dimensional Data Sets as Minimum Spanning Trees. *J. Cheminform.* **2020**, *12*, 12.

(99) Sánchez-Cruz N, Medina-Franco JL. Statistical-Based Database Fingerprint: Chemical Space Dependent Representation of Compound Databases. *J. Cheminform.* **2018**, *10*, 55.

(100) Lewell X. Q, Judd DB, Watson SP, Hann MM. RECAP--Retrosynthetic Combinatorial Analysis Procedure: A Powerful New Technique for Identifying Privileged Molecular Fragments with Useful Applications in Combinatorial Chemistry. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 511–522.

(101) Zhao Y, Chen CH, Morris-Natschke SL, Lee KH. Design, Synthesis, and Structure Activity Relationship Analysis of New Betulinic Acid Derivatives as Potent HIV Inhibitors. *Eur. J. Med. Chem.* **2021**, *215*, 113287.

(102) Martin DE, Salzwedel K, Allaway GP. Bevirimat: A Novel Maturation Inhibitor for the Treatment of HIV-1 Infection. *Antivir. Chem. Chemother.* **2008**, *19*, 107–113.

(103) Lazerwith SE, Siegel D, McFadden RM, Mish MR, Tse WC. 5.19—New Antiretrovirals for HIV and Antivirals for HBV; Chackalamannil S, Rotella D, Ward SE, Eds.; Elsevier: Oxford, UK, **2017**; pp. 628–664.

(104) Qian K, Kuo RY, Chen CH, Huang L, Morris-Natschke SL, Lee KH. Anti-AIDS Agents 81. Design, Synthesis, and Structure-Activity Relationship Study of Betulinic Acid and Moronic Acid Derivatives as Potent HIV Maturation Inhibitors. *J. Med. Chem.* **2010**, *53*, 3133–3141.

(105) Huang QX, Chen HF, Luo XR, Zhang YX, Yao X, Zheng X. Structure and Anti-HIV Activity of Betulinic Acid Analogues. *Curr Med Sci* **2018**, *38*, 387–397.

(106) Schomburg K, Ehrlich HC, Stierand K, Rarey M. Chemical Pattern Visualization in 2D – the SMARTSviewer. *J. Cheminform.* **2011**, *3*, O12.

(107) 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.* **1997**, *23*, 3–25.

(108) 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.

(109) Ertl P, Schuffenhauer A. Estimation of Synthetic Accessibility Score of Drug-like Molecules Based on Molecular Complexity and Fragment Contributions. *J. Cheminform.* **2009**, *1*, 8.

(110) Wishart DS, Feunang YD, Guo AC, Lo EJ, Marcu A, Grant JR, Sajed T, Johnson D, Li C, Sayeeda Z, Assempour N, Iynkkaran I, Liu Y, Maciejewski A, Gale N, Wilson A, Chin L, Cummings R, Le D, Pon A, Knox C, Wilson M. DrugBank 5.0: A Major Update to the DrugBank

Database for 2018. *Nucleic Acids Res.* **2017**, *46*, D1074–D1082.

(111) Selecting Diverse Sets Of Compounds. In *An Introduction To Chemoinformatics*; Leach AR, Gillet VJ, Eds.; Springer Netherlands: Dordrecht, **2007**; pp 119–139.

(112) *REAL DATABASE*. Enamine. https://enamine.net/compound-collections/real-compounds/real-database (accessed 2023-05-13).

(113) Mendez D, Gaulton A, Bento AP, Chambers J, De Veij M, Félix E, Magariños MP, Mosquera JF, Mutowo P, Nowotka M, Gordillo-Marañón M, Hunter F, Junco L, Mugumbate G, Rodriguez-Lopez M, Atkinson F, Bosc N, Radoux CJ, Segura-Cabrera A, Hersey A, Leach AR. ChEMBL: Towards Direct Deposition of Bioassay Data. *Nucleic Acids Res.* **2019**, *47*, D930–D940.

(114) Davies M, Nowotka M, Papadatos G, Dedman N, Gaulton A, Atkinson F, Bellis L, Overington JP. ChEMBL Web Services: Streamlining Access to Drug Discovery Data and Utilities. *Nucleic Acids Res.* **2015**, *43*, W612–W620.

(115) Capuzzi SJ, Muratov E, Tropsha A. *PAINS Killer: Popular Drug Screening Tool Has Serious Problems*. UNC Eshelman School of Pharmacy. https://pharmacy.unc.edu/2017/05/serious-problems-found-pains-alerts/ (accessed 2024-01-20).

(116) Baell JB, Holloway GA. 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.

(117) Baell J, Walters MA. Chemistry: Chemical Con Artists Foil Drug Discovery. *Nature* **2014**, *513*, 481–483.

(118) Chávez-Hernández AL, Sánchez-Cruz N, Medina-Franco JL. A Fragment Library of Natural Products and Its Comparative Chemoinformatic Characterization. *Mol. Inform.* **2020**, *39*, e2000050.

(119) Bemis GW, Murcko MA. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996**, *39*, 2887–2893.

(120) Schneider G, Clark DE. Automated DE Novo Drug Design: Are We Nearly There Yet? *Angew. Chem. Int. Ed Engl.* **2019**, *58*, 10792–10803.

(121) Ertl P, Schuhmann T. A Systematic Cheminformatics Analysis of Functional Groups Occurring in Natural Products. *J. Nat. Prod.* **2019**, *82*, 1258–1263.

(122) *FooDB*. https://foodb.ca/ (accessed 2023-04-20).

(123) Wassermann AM, Lounkine E, Hoepfner D, Le Goff G, King FJ, Studer C, Peltier JM, Grippo ML, Prindle V, Tao J, Schuffenhauer A, Wallace IM, Chen S, Krastel P, Cobos-Correa A, Parker CN, Davies JW, Glick M. Dark Chemical Matter as a Promising Starting Point for Drug Lead Discovery. *Nat. Chem. Biol.* **2015**, *11*, 958–966.

(124) *American Chemical Society: CAS COVID-19 Antiviral Candidate Compounds Dataset*. https://www.cas.org/covid-19-antiviral-compounds-dataset (accessed 2020-05-19).

(125) Tang B, He F, Liu D, He F, Wu T, Fang M, Niu Z, Wu Z, Xu D. AI-Aided Design of Novel Targeted Covalent Inhibitors against SARS-CoV-2. *Biomolecules* **2022**, *12*, 746.

(126) Chávez-Hernández AL, Sánchez-Cruz N, Medina-Franco JL. Fragment Library of Natural Products and Compound Databases for Drug Discovery. *Biomolecules* **2020**, *10*, 1518.

(127) Saldívar-González FI, Lenci E, Calugi L, Medina-Franco JL, Trabocchi A. Computational-Aided Design of a Library of Lactams through a Diversity-Oriented Synthesis Strategy. *Bioorg. Med. Chem.* **2020**, *28*, 115539.

(128) Laurini R. Geographic Relations. *Geographic Knowledge Infrastructure; Elsevier* **2017**.

(129) Ganesan A. The Impact of Natural Products upon Modern Drug Discovery. *Curr. Opin. Chem. Biol.* **2008**, *12*, 306–317.

(130) Tinworth CP, Young RJ. Facts, Patterns, and Principles in Drug Discovery: Appraising the Rule of 5 with Measured Physicochemical Data. *J. Med. Chem.* **2020**, *63*, 10091–10108.

(131) Chávez-Hernández AL, Medina-Franco JL. Natural Products Subsets: Generation and Characterization. *Artificial Intelligence in the Life Sciences* **2023**, *3*, 100066.

(132) Conery AR, Rocnik JL, Trojer P. Small Molecule Targeting of Chromatin Writers in Cancer. *Nat. Chem. Biol.* **2022**, *18*, 124–133.

(133) Prado-Romero DL, Medina-Franco JL. Advances in the Exploration of the Epigenetic Relevant Chemical Space. *ACS Omega* **2021**, *6*, 22478–22486.

(134) Chávez-Hernández AL, Juárez-Mercado KE, Saldívar-González FI, Medina-Franco JL. Towards the De Novo Design of HIV-1 Protease Inhibitors Based on Natural Products. *Biomolecules* **2021**, *11*, 1805.

XII. Apéndice

Tabla A1. Características atómicas que codifican los grafos moleculares.

Características	Descriptores moleculares	Número
Tipo de átomo	H, B, C, N, O, F, Si, P, S, Cl, Se, Br, I [Mg, Na, Ca, Fe, As, Al, 'I', B, V, K, Tl, Y, Sb, Sn, Ag, Pd, Co, Se, Ti, Zn, Li, Ge, Cu, Au, Ni, Cd, In, Mn, Zr, Cr, Pt, Hg, Pb][a]	13
Quiralidad	0:'CHI_UNSPECIFIED',1:'CHI_TETRAHEDRAL_CW', 2:'CHI_TETRAHEDRAL_CCW', 3: 'CHI_OTHER', 4:'CHI_TETRAHEDRAL', 5: 'CHI_ALLENE', 6: 'CHI_SQUAREPLANAR', 7: 'CHI_TRIGONALBIPYRAMIDAL', 8: 'CHI_OCTAHEDRAL'	9
Número de átomos vecinos	0, 1, 2, 3, 4, "Más que cuatro"	6
Hibridización	S, SP, SP2, SP3 [SP3D, SP3D2, OTHER]	4
Carga formal	-4,-3, -2, -1, 0, 1, 2, 3, 4, 5, 6	11
Sistema de anillos	El átomo está en un anillo: 0:No, 1:Sí	2
Aromaticidad	El átomo es parte de un sistema aromático: 0:No, 1:Sí	2
Hidrógenos	Número de hidrógenos vecinos	5

[a] Los átomos en color azul son metales.

Tabla A2. Características químicas de enlaces químicos que codifican los grafos moleculares.

Características	Descriptores moleculares	Número
Tipo de enlace	0:'UNSPECIFIED', 1: 'SINGLE', 2:'DOUBLE', 3:'TRIPLE', 4:'QUADRUPLE', 5:'QUINTUPLE', 6: 'HEXTUPLE', 7:'ONEANDAHALF', 8: 'TWOANDAHALF', 9: 'THREEANDAHALF', 10: 'FOURANDAHALF', 11: 'FIVEANDAHALF', 11: 'AROMATIC', 12:'IONIC', 13: 'HYDROGEN',14:'THREECENTER', 15: 'DATIVEONE', 16: 'DATIVE', 17: 'DATIVEL', 18: 'DATIVER', 19: 'OTHER', 20: 'ZERO'	21
Estereoquimica del enlace	0:'STEREONONE', 1: 'STEREOANY', 2: 'STEREOZ', 3: 'STEREOE', 4: 'STEREOCIS', 5: 'STEREOTRANS'	6
Enlace conjugado	0:No, 1:Sí	2

$$Puntuación\ de\ fragmentos\ moleculares\ = (Log\frac{Número\ de\ veces\ que\ se\ repite\ un\ fragmento\ en\ una\ molécula}{Fragmentos\ formados})$$

Ecuación A1. Puntuación de fragmentos moleculares.

$$Penalidad\ por\ complejidad\ =\ Complejidad\ de\ estereocentros\ +\ Penalidad\ de\ macrociclos\ +\ Penalidade\ de\ tamaño\ +\ Penalidad\ de\ anillos$$

$$Complejidad\ de\ estereocentros\ =\ log\ P(Número\ de\ estereocentros\ posibles\ +\ 1)$$

$$Penalidad\ de\ macrociclos\ =\ log(Número\ de\ macrocíclos\ +\ 1)$$

$$Penalidade\ de\ tamaño\ =\ Número\ de\ átomos(1.005\ -\ Número\ de\ átomos)$$

$$Penalidad\ de\ anillos\ =\ log(Número\ de\ átomos\ cabeza\ de\ puente\ +\ 1)\ +\ log(Número\ de\ átomos\ espiro\ +\ 1)$$

Ecuación A2. Penalidad por complejidad.

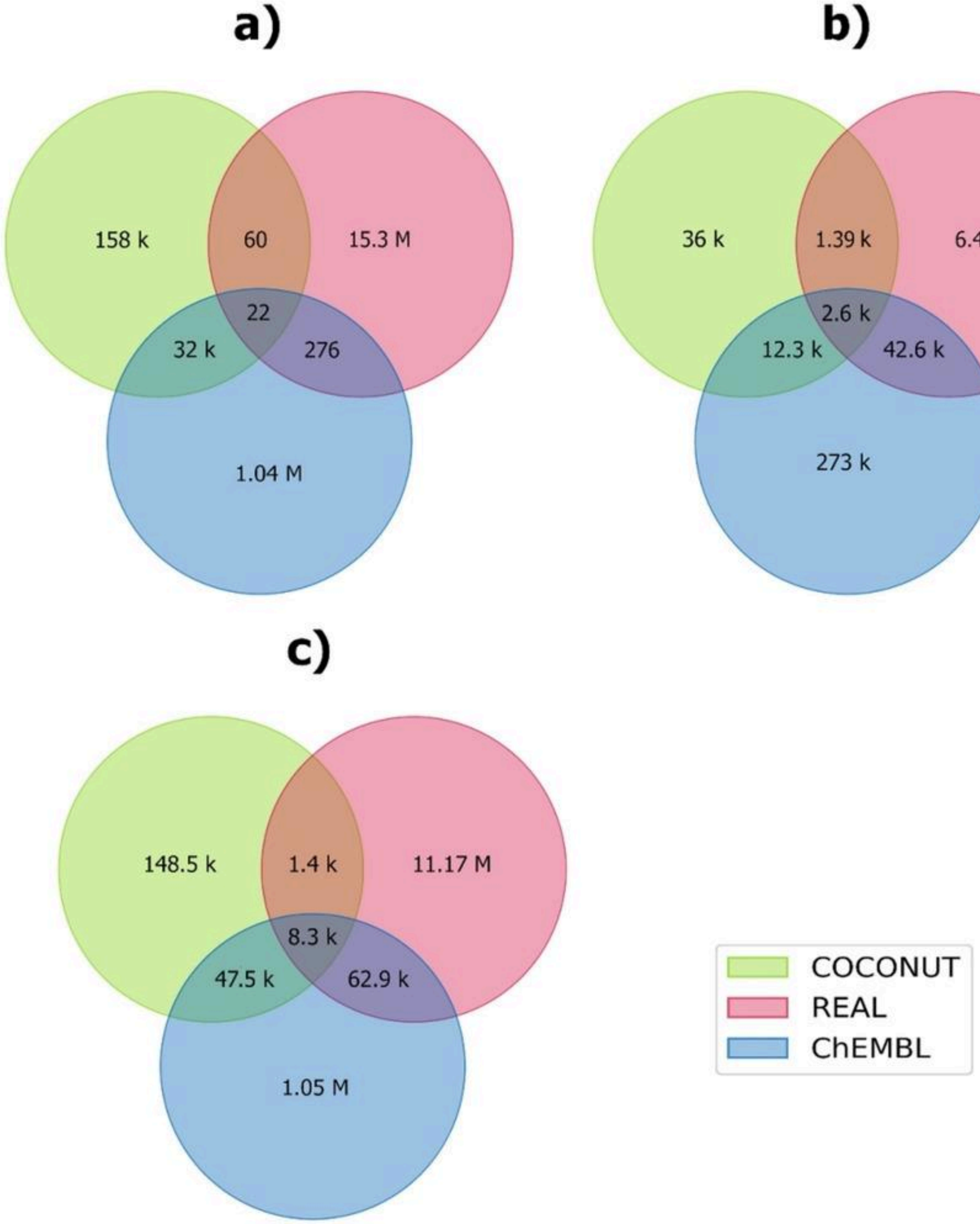


Figura A1. Estructuras únicas y en común entre COCONUT, ChEMBL y REAL analizadas en la Tabla 1. El contenido estructural fue analizado en terminos de a) compuestos, b) Núcleos estructural base y c) Fragmentos moleculares. Las bases de datos están representadas en diferentes colores como verde (COCONUIT), rojo (REAL) y azul (ChEMBL). La letra *k* representa miles de compuestos y la letra *M* representa millones de compuestos.

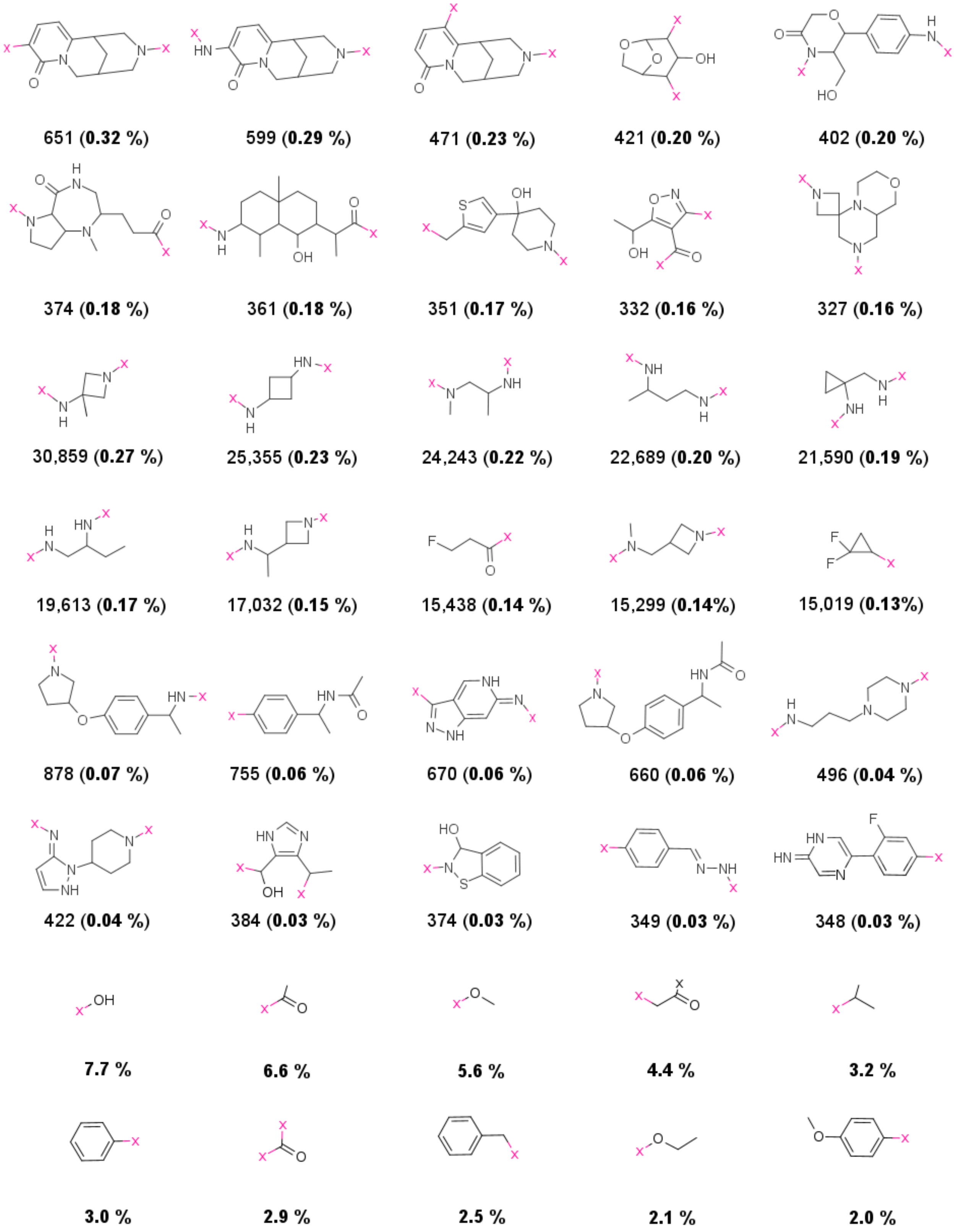


Figura A2. Los diez fragmentos moleculares más frecuentes y únicos de a) COCONUT, b) REAL y c) ChEMBL. El número de frecuencias en el que se encuentra el fragmento en cada base de datos se muestra en letra normal y el porcentaje en negritas.

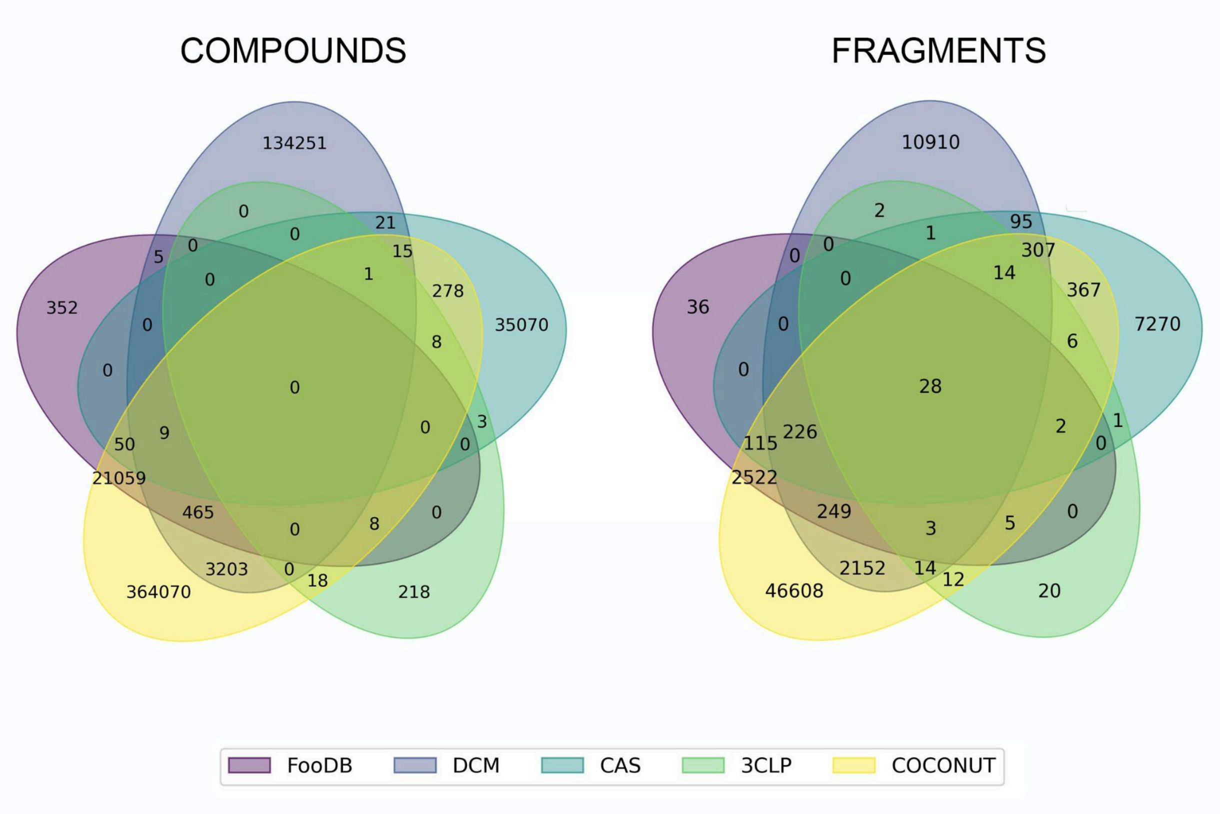


Figura A3. Compuestos y fragmentos en común de las bases de datos de COCONUT, FooDB, DCM, CAS y 3CLP. Los compuestos y fragmentos son representados en diferentes colores como amarillo (COCONUT), morado (FooDB), azul (DCM), verde (CAS) y lima (3CLP).

Tabla A3. Composición estructural de los compuestos de COCONUT, FooDB y bases de datos de referencia.[a]

Características estructurales	COCONUT	FooDB	DCM	CAS	3CLP
Átomos de carbono	25.640	26.563	18.059	22.496	25.828
Átomos de oxígeno	6.167	7.343	3.252	5.773	4.922
Átomos de nitrógeno	1.445	0.668	2.859	4.157	3.582
Átomos pesados	33.611	34.942	25.139	33.535	33.352
Fracción de carbonos sp^3	0.506	0.620	0.342	0.489	0.291
Fracción de carbonos quirales	0.154	0.152	0.028	0.145	0.069
Número de anillos	3.962	2.243	2.881	3.628	3.617
Número de anillos Alifáticos	2.250	1.426	0.791	1.372	0.645
Número de anillos aromáticos	1.712	0.817	2.089	2.256	2.973
Número de heterociclos	1.711	1.020	1.408	2.056	1.500
Número de heterociclos alifáticos	1.166	0.770	0.619	0.865	0.363
Número de heterociclos aromáticos	1.712	0.817	2.089	2.256	2.973
Átomos spiro	0.167	0.051	0.018	0.019	0.000
Átomos cabeza de puente	0.493	0.137	0.056	0.254	0.023

[a] Media de distribución.

Tabla A4. Resumen de composición estructural de los fragmentos de COCONUT, FooDB y bases de datos de referencia.[a]

Características estructurales	COCONUT	FooDB	DCM	CAS	3CLP	Fragmentos en común
Átomos de carbono	18.504	12.991	10.181	9.904	8.926	5.179
Átomos de oxígeno	3.524	3.173	1.748	3.678	1.556	1.107
Átomos de nitrógeno	0.795	0.394	1.475	0.883	0.713	0.107
Átomos pesados	23.034	16.760	14.057	15.532	11.537	6.464
Fracción de carbonos sp^3	0.557	0.615	0.330	0.656	0.298	0.318
Fracción de carbonos quirales	0.189	0.199	0.054	0.240	0.071	0.062
Número de anillos	2.999	1.739	1.686	1.496	1.398	0.571
Número de anillos Alifáticos	2.013	1.237	0.447	0.837	0.398	0.071
Número de anillos aromáticos	0.986	0.503	1.239	0.660	1.000	0.500
Número de heterociclos	1.087	0.577	0.899	0.787	0.574	0.179
Número de heterociclos alifáticos	0.751	0.390	0.313	0.573	0.176	0.036
Número de heterociclos aromáticos	0.986	0.503	1.239	0.660	1.000	0.500

[a]Media de distribución.

Tabla A4 (continuación). Resumen de composición estructural de los fragmentos de COCONUT, FooDB y bases de datos de referencia.[a]

Características estructurales	COCONUT	FooDB	DCM	CAS	3CLP	Fragmentos en común
Átomos spiro	0.190	0.085	0.013	0.010	0.000	0.000
Átomos cabeza de puente	0.507	0.288	0.043	0.109	0.056	0.000

[a]Media de distribución.

Tabla A5. Grupos funcionales usando SMARTS para filtrar fragmentos derivados de productos naturales.

Grupos funcionales	SMARTS
Alcohol alifático (ciclohexanol)	[#8;H1]-[#6]-1-[#6]-[#6]-[#6]-2-[#6](-[#6]-[#6]-[#6]-3-[#6]-4-[#6]-[#6]C5([#6]-[#6]-[#6]-[#6]5-[#6]-4-[#6]-[#6]-[#6]-2-3)[#6]([#8;H1])=O)-[#6]-1
ácido 2,2-dimetil succínico	[#6]C([#6])([#6]-[#6](-[#8])=O)[#6](-[#8])=O
Piperazina	[#6;H2;X4]1-[#6;H2;X4][#7;X3;!H1][#6;H2;X4]-[#6;H2;X4][#7;H1;X3]1
1,2-diaminoetano	[#7;H1;X3][#6;H2;X4][#6;H2;X4][#7;H2;X3]
1,3-diaminopropano	[#7;H1;X3][#6;H2;X4][#6;H2;X4][#6;H2;X4][#7;H2;X3]
Sistema cíclico derivado del ácido betulínico	[#6]1-[#6]-[#6]-[#6]2-[#6](-[#6]-1)-[#6]-[#6]-[#6]1-[#6]-2-[#6]-[#6]-[#6]2-[#6]3-[#6]-[#6]-[#6]-[#6]-3-[#6]-[#6]-[#6]-1-2

Tabla A6. SMIRKS usados para construir bibliotecas enfocadas de compuestos inhibidores de la proteasa viral del VIH-1 a partir de fragmentos de productos naturales.

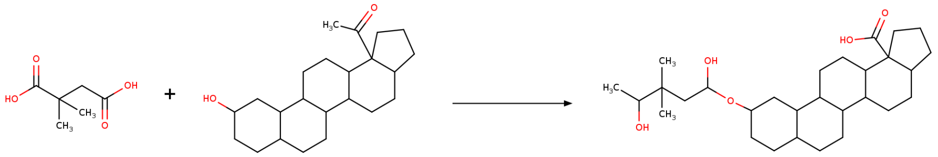
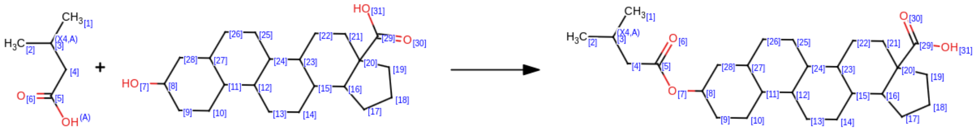
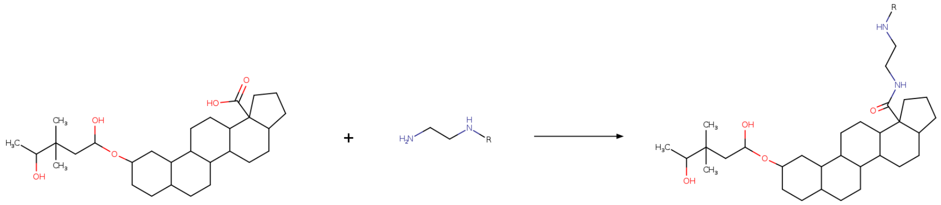
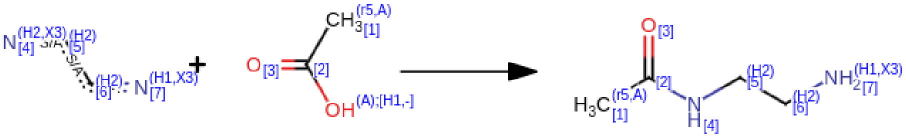
Descripción	Esquema de reacción
Reacción 1	
SMIRKS 1	 [#6:1][#6;A;X4:3]([#6:2])[#6:4]-[#6:5]([#8;A])=[O:6].[#8:7]-[#6:8]-1-[#6:9]-[#6:10]-[#6:11]-2-[#6:27](-[#6:26]-[#6:25]-[#6:24]-3-[#6:23]-4-[#6:22]-[#6:21][C:20]5([#6:19]-[#6:18]-[#6:17]-[#6:16]5-[#6:15]-4-[#6:14]-[#6:13]-[#6:12]-2-3)[#6:29](-[#8:31])=[O:30])-[#6:28]-1>>[#6:2][#6;A;X4:3]([#6:1])[#6:4]-[#6:5](=[O:6])-[#8:7]-[#6:8]-1-[#6:9]-[#6:10]-[#6:11]-2-[#6:27](-[#6:26]-[#6:25]-[#6:24]-3-[#6:23]-4-[#6:22]-[#6:21][C:20]5([#6:19]-[#6:18]-[#6:17]-[#6:16]5-[#6:15]-4-[#6:14]-[#6:13]-[#6:12]-2-3)[#6:29](-[#8:31])=[O:30])-[#6:28]-1
Reacción 2.1	
SMIRKS 2.1	 [#7;H1;X3:7][#6H2:6][#6;H2:5][#7;H2;X3:4].[#6;A;r5:1][#6:2]([#8;A;H1,-])=[O:3]>>[#6;A;r5:1][#6:2](=[O:3])-[#7:4]-[#6;H2:5]-[#6;H2:6]-[#7;H1;X3:7]

Tabla A6 (continuación). SMIRKS usados para construir nuevos compuestos derivados de fragmentos de productos naturales.

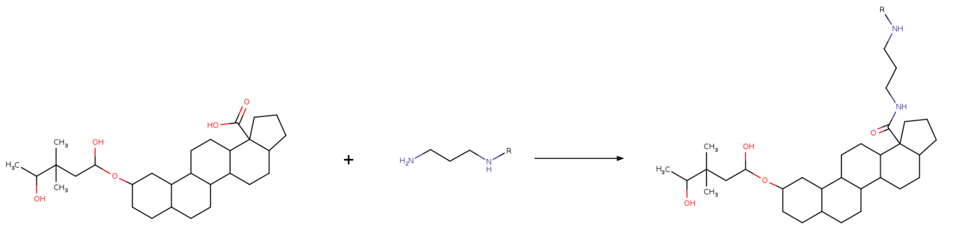
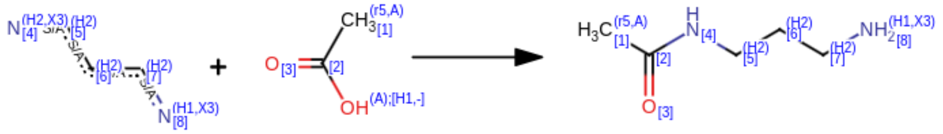
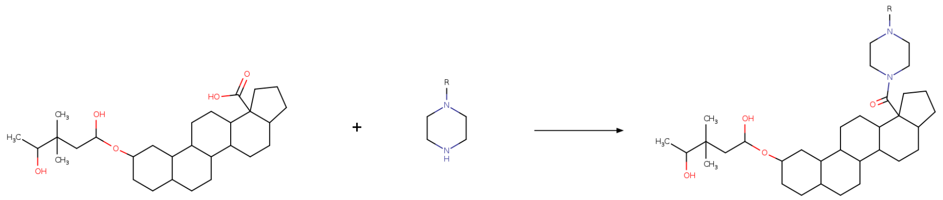
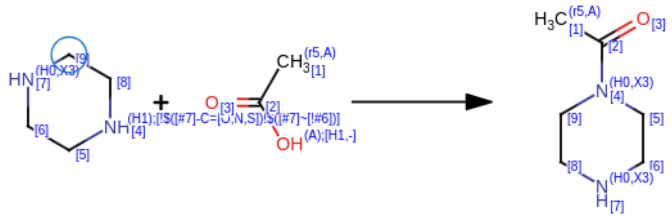
Descripción	Esquema de reacción
Reacción 2.2	
SMIRKS 2.2	 [#7;H1X3:8][#6H2:7][#6H2:6][#6H2:5][#7;H2X3:4].[#6;A;r5:1][#6:2]([#8;A;H1,-])=[O:3]>>[#6;A;r5:1][#6:2](=[O:3])-[#7:4]-[#6H2:5]-[#6H2:6]-[#6H2:7]-[#7;H1X3:8]
Reacción 2.3	
SMIRKS 2.3	 [#6:9]-1-[#6:8]-[#7H1;!$([#7]-C=[O,N,S])!$([#7]~[!#6]):4]-[#6:5]-[#6:6]-[#7;H0X3:7]-1.[#6;A;r5:1][#6:2]([#8;A;H1,-])=[O:3]>>[#6;A;r5:1][#6:2](=[O:3])-[#7;H0X3:4]-1-[#6:5]-[#6:6]-[#7;H0X3:7]-[#6:8]-[#6:9]-1

Tabla A7. Diversidad estructural basada en huellas digitales moleculares de compuestos generados a partir de fragmentos de COCONUT, ChemDiv y Enamine, y dos bases de datos de compuestos de referencia, fármacos aprobados por la FDA e inhibidores de la proteasa viral del VIH-1 aprobados por la FDA.

Base de datos	ECFP-4[a] (1024-bits)	MACCS Keys[a] (166-bits)
COCONUT	0.605	0.817
Enamine	0.682	0.823
ChemDiv	0.676	0.821
Fármacos aprobados por la FDA	0.096	0.293
Inhibidores de la proteasa viral del VIH-1 aprobados por la FDA	0.253	0.558

[a]Mediana de similitud.

Tabla A8. Propiedades de relevancia farmacéutica de compuestos inhibidores de la proteasa viral del VIH-1 aprobados por la FDA.

Molécula	log P	MW	HBD	HBA	TPSA	RB
Amprenavir[a]	2.4	505.22	4	9	131.19	11
Atazanavir[a]	4.21	704.39	5	13	171.22	14
Darunavir[a]	2.38	547.24	4	10	140.42	11
Fosamprenavir[a]	2.69	585.19	4	12	174.56	13
Indinavir[a]	2.87	613.36	4	9	118.03	11
Lopinavir[a]	4.33	628.36	4	9	120.00	15
Nelfinavir[a]	4.75	567.31	4	7	101.90	9
Ritonavir[a]	5.91	720.31	4	11	145.78	17
Saquinavir[a]	3.09	670.38	6	11	166.75	12
Tipranavir[a]	6.70	602.21	1	7	102.43	11
Mínimo	2.40	505.20	1	7	101.90	9
Máximo	6.70	720.30	6	13	174.60	17

[a]Valor mínimo y máximo para cada propiedad.

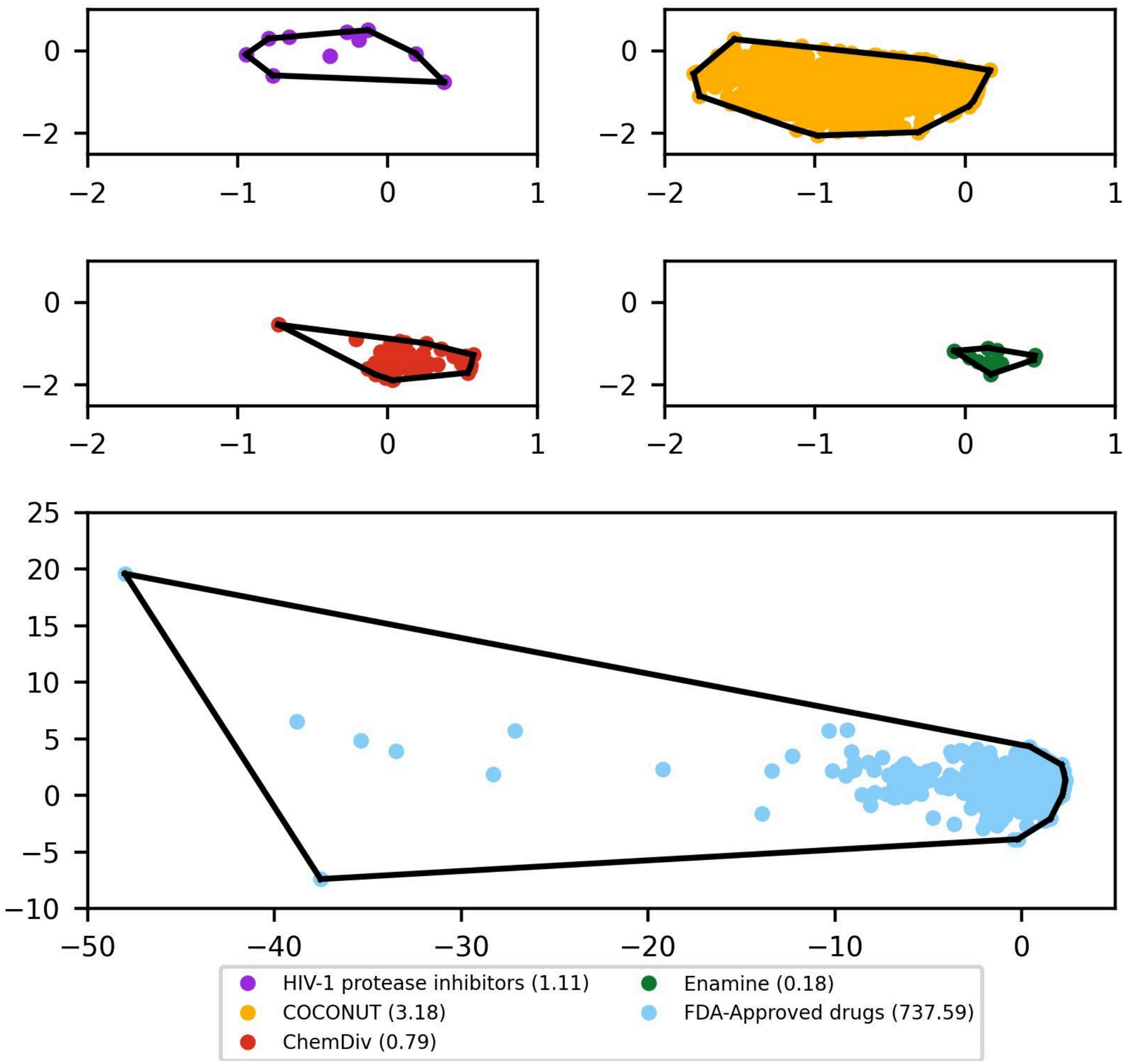


Figura A4. Análisis de *convex hull* derivado de una visualización de espacio químico basada en propiedades fisicoquímicas utilizando PCA de bibliotecas enfocadas a bibliotecas enfocadas de compuestos inhibidores de la proteasa viral del VIH-1 a partir de fragmentos de productos naturales de compuestos de referencia. Las bibliotecas enfocadas a bibliotecas enfocadas de compuestos inhibidores de la proteasa viral del VIH-1 a partir de fragmentos de productos naturales en color naranja (COCONUT), rojo (ChemDiv) y verde (Enamine). Las bibliotecas de compuestos de referencia se muestran en color azul (fármacos aprobados por la FDA), morado (inhibidores de la proteasa viral del VIH-1 aprobados por la FDA).

Artículos científicos

DOI: 10.1002/minf.202000050

A Fragment Library of Natural Products and its Comparative Chemoinformatic Characterization

Ana L. Chávez-Hernández+,[a] Norberto Sánchez-Cruz+,*[a] and José L. Medina-Franco*[a]

This manuscript is dedicated to all people affected directly or indirectly by the COVID-19 pandemic around the world.

Abstract: We report a comprehensive fragment library with 205,903 fragments derived from the recently published Collection of Open Natural Products (COCONUT) data set with more than 400,000 non-redundant natural products. The natural products-based fragment library was compared with other two fragment libraries herein generated from ChEMBL (biologically relevant compounds) and Enamine-REAL (a large on-demand collection of synthetic compounds), both used as reference data sets with relevance in drug discovery. It was found that there is a large diversity of unique fragments derived from natural products and that the entire structures and fragments derived from natural products are more diverse and structurally complex than the two reference compound collections. During this work we introduced a novel visual representation of the chemical space based on the recently published concept of statistical-based database fingerprint. The compounds and fragments libraries from natural products generated and analyzed in this work are freely available.

Keywords: ChEMBL · drug discovery · fingerprint · fragment · natural product

1 Introduction

Natural products (NPs) have been relevant in drug discovery pipelines since the beginning of the pharmaceutical era. They have inspired the synthesis of drugs such as aspirin from salicylic acid or ampicillin from penicillin[1] to such a degree that several drugs are NPs or derivatives thereof.[2] For instance, from the approved drugs between 1981 and 2014, 4% corresponds to unaltered NPs and 21% corresponds to NPs derivatives.[3] Also, since NPs have gone through an adaptation process, they represent attractive ligands for several biological targets.[4] Furthermore, in comparison with molecules obtained with combinatorial chemistry or other synthetic methods, NPs are structurally more diverse and complex thus contributing to their overall larger selectivity.[5–6] These reasons plus the broad use of NPs in traditional medicine, make them a fundamental part to inspire or be the starting point for developing new drugs.[4]

Otherwise, general downsides of NPs are the short amounts of them are obtained and their procurement procedures are costly and lengthy.[7] Despite these limitations, NPs, unlike synthetic molecules, possess unique functional groups, unique scaffolds, and unique characteristic structural fragments that could provide important information related to biological activity.[7] This could be used as the starting point for designing novel compounds. Thus, fragments obtained from NPs can be further used in traditional fragment-based or *de novo* drug design.[8] This is why it is desirable to generate fragment libraries from NPs[9] that can be used to build novel molecules such as the so-called “pseudo-NPs”.[8]

In this work, we report a novel and comprehensive database of fragments derived from NPs based on the COlleCtion of Open NatUral producTs (COCONUT),[10] a recently published database with more than 400,000 non-redundant compounds. The fragment library was characterized and compared with fragment libraries herein generated from two large reference compound data sets with relevance in drug discovery: ChEMBL as a source of biologically relevant compounds, and Enamine-REAL, a large on-demand collection of synthetic compounds. The newly developed fragment library from NP is freely accessible at https://doi.org/10.6084/m9.figshare.11997951

2 Methods

2.1 Data Sets

We selected three data sets with relevance for drug discovery: COCONUT (first version),[10] a data set assembled

[a] *A. L. Chávez-Hernández,+ N. Sánchez-Cruz,+ J. L. Medina-Franco*
Department of Pharmacy, School of Chemistry, Universidad Nacional Autónoma de México, Avenida Universidad 3000, Mexico City, 04510, Mexico
Phone: +5255-5622-3899. Ext. 44458
E-mail: norberto.sc90@gmail.com
medinajl@unam.mx

[+] *Both authors contributed equally to the work*

Table 1. Compound data sets analyzed in this work and summary statistics for diversity and complexity of the entire compounds.

Data sets*	Size (compounds)	Median similarity (Morgan2 – 1024 bits)	Median similarity (MACCS keys – 166 bits)	Mean fraction of sp^3 carbons	Mean fraction of chiral carbons	Reference
COCONUT	190,139	0.111	0.344	0.453	0.112	[10]
Enamine, REAL	15,297,437	0.123	0.420	0.526	0.068	[11]
ChEMBL	1,074,335	0.119	0.377	0.318	0.033	[12–13]

*Drug-like sets (see Section 2.2).

from 50 open-access databases containing 412,903 compounds and being the largest collection of NP available to this date; the REAL drug-like data set from Enamine[11] consisting of 15,547,017 Readily AccessibLe compounds representing the chemical space covered by synthetic molecules, and ChEMBL 25[12–13] as a representative example of the biologically tested chemical space with 1,844,434 compounds. The three datasets were curated using the same procedure outlined in Section 2.2 and are available at the Supporting Information.

2.2 Data Curation

SMILES strings with no stereochemistry information were selected as a molecular representation of compounds. Stereochemistry information was not considered in this work because not all compounds in the three data sets contain defined stereochemistry. The entire preparation process was performed with the open-source cheminformatics toolkit RDKit (http://www.rdkit.org), version 2019.09.1 and the functions Standardizer, LargestFragmentChoser, Uncharger, Reionizer and TautomerCanonicalizer implemented in the molecule validation and standardization tool MolVS.[14] Compounds were standardized and those consisting of multiple components were split and the largest component was retained. Compounds consisting of any element other than H, B, C, N, O, F, Si, P, S, Cl, Se, Br and I, as well as compounds with valence errors, were removed from the data set. The remaining compounds were neutralized and reionized to subsequently generate a canonical tautomer. Duplicated structures within each database were also removed. Six molecular properties were computed for each compound: averaged molecular weight (AMW), partition coefficient octanol/water (SlogP), number of hydrogen bond donors (HBD), number of hydrogen bond acceptors (HBA), number of rotatable bonds (RB), and topological polar surface area (TPSA). Only compounds complying with the "rule of 5" and Veber criteria ($AMW \leq 500$, $-1 \leq SlogP \leq 5$, $HBA \leq 10$, $HBD \leq 5$, $RB \leq 10$ and $TPSA \leq 140$) were preserved. Finally, pan-assay interference compounds were removed according to the substructures defined in RDKit. The three data sets used in this study after data curation are summarized in Table 1.

2.3 Fragment Generation

Fragment libraries for the three data sets described in Section 2.1 were generated by using the REtrosynthetic Combinatorial Analysis Procedure (RECAP) as implemented in RDKit. The RECAP algorithm is based on eleven cleavage rules derived from common chemical reactions.[15] In short, if a molecule contains any of eleven bounds (such as amide, ester, amine, urea, ether, olefin, quaternary nitrogen, aromatic nitrogen-aliphatic carbon, lactam nitrogen-aliphatic carbon, aromatic carbon-aromatic carbon, and sulphonamide) then it is cleaved into fragments. These rules only apply to acyclic bonds to leave residual rings intact. Each molecular fragment retains the atoms where a bond was cleaved to denote the atom environments from which it was obtained. Fragment libraries for the three data sets are available at the Supporting Information.

2.4 Data Sets Overlap

Overlap of COCONUT with the data sets selected as reference was assessed in terms of three different structural levels: compounds, scaffolds, and fragments. Compound and fragment overlap was determined in terms of canonical SMILES. For scaffold comparison, we use the definition proposed by Bemis and Murcko[16] as implemented in RDKit. For each structural level, we identified the unique structures belonging to each data set as well as those belonging to two or three of them.

2.5 Diversity and Complexity Analysis

One of the main goals to generate a general screening compound library is to have large diversity and cover as much chemical space as possible.[17] For this reason, the three original compound data sets, as well as the three fragment libraries derived from them, were analyzed in terms of structural diversity and complexity. Structural diversity was measured calculating the median value of the distribution of the pairwise similarity values calculated with the Tanimoto coefficient and both Molecular ACCes System (MACCS) keys (166-bits)[18] and Morgan fingerprint with radius 2 (Morgan2).[19] This was done for 10 random samples of 10,000 compounds and fragments, respectively. Structural complexity was measured as the mean fraction of chiral and sp^3 carbons.

In order to characterize the structural differences between the generated fragment databases, eleven descriptors were calculated being number of heavy atoms broken down into oxygen atoms, nitrogen atoms, bridgehead atoms and spiro atoms as well as the number of rings and number of heterocycles, both broken down into aromatic and aliphatic. The differences were analyzed in the context of the unique fragments from each data sets and the common fragments in all three of them, using as measure the mean values of the descriptors distributions.

2.6 Chemical Space Visualization Based on SB-DFP

To generate a two-dimensional representation of the chemical space covered by the analyzed data sets, we used the concept of Statistical-Based Database Fingerprint (SB-DFP),[20] a recently published approach to generate single fingerprint representations of compound data sets. A brief description for the construction of an SB-DFP is as follows: given a fingerprint representation of compounds in a data set, the frequency occurrence of each bit in the data set is compared to a reference in such a way that a bit is set to "1" in the final representation only if the frequency of such bit in the data set is statistically higher than in the reference otherwise, the bit is set to "0". In this work, we built two SB-DFPs: one to represent NPs and the other to represent synthetic compounds, in such a way that all compounds and fragments could be mapped according to its Tanimoto similarity to each of the generated SB-DFPs. To this end, we used a random sample of 60 % of compounds present exclusively in the prepared COCONUT or REAL data sets with 190,139 and 15,297,437 compounds (Table 1), respectively, using each as the reference for the other. The selected molecular representation was Morgan2. For the frequency comparisons, we employed a Z-test with a confidence level of 99 %, as described in the original work.[20] The remaining 40 % of compounds were used to compute the similarity values of compounds to the SB-DFPs and scale them to a range between 0 and 1. A visual representation of the chemical space covered by both compounds and fragments was generated based on their Tanimoto similarities to each of the generated SB-DFPs. SB-DFPs for COCONUT and REAL data sets are available at the Supporting Information.

3 Results and Discussion

3.1 Data Sets Overlap

We characterized the structural content of the three data sets summarized in Table 1 (COCONUT, REAL, and ChEMBL) in terms of unique compounds, molecular scaffolds, generated fragments and determined the overlap among them. Of note, from the data curation process described in Section 2.2 the COCONUT and ChEMBL analyzed herein are "drug-like" subsets from the initial sets and are comparable in properties to the "drug-like" REAL set. Figure 1 depicts Vehn diagrams showing the overlap among the compounds (Figure 1a), scaffolds using the Bemis-Murcko definition (Figure 1b), and fragments (Figure 1c).

Figure 1a indicated that there are 16,529,500 unique compounds among the three data sets. The largest overlap among them occurs for the intersection between COCONUT and ChEMBL, with a total of 32,053 compounds, from which only 22 were also shared with REAL. Overlaps involving the REAL data set are practically non-existing considering its size, being 60 and 276 compounds shared with COCONUT and ChEMBL data sets, respectively. It should be noted that despite the existing overlapping of the data sets, 99.8 of compounds are unique and belong only to a single set. In terms of each data set size, non-overlapping compounds represent 83.1 % of COCONUT, 97.0 % of ChEMBL, and more than 99.9 % of REAL.

In terms of scaffolds (Figure 1b), a total of 6,852,628 unique structures were identified, from which 99.1 are non-overlapping, representing 68.7 %, 82.6 % and 99.3 % of COCONUT, ChEMBL, and REAL data sets, respectively. While regarding fragments (Figure 1c), a total of 12,497,641 unique structures was obtained, 99.0 % of them being non-overlapping and corresponding to 72.2 %, 89.9 and 99.4 % of COCONUT, ChEMBL, and REAL data sets, respectively. These results are in agreement with the overall structural novelty associated with the drug-like data set from Enamine and suggest that the fragment space associated with NPs is not fully covered by those coming from synthetic or biologically tested compounds, supporting the idea that fragments of NPs can serve as building blocks for *de novo* design.[21] In addition, there is a broad diversity of unique fragments and scaffolds derived from NPs that could be used later in the development and discovery of new drugs.

3.2 Fragment Analysis

As described in Section 2.3, for all data sets, the RECAP fragmentation algorithm was useful to generate all fragments with common synthetic paths. Therefore, the fragments are delimited by the fragmentation algorithm used. Fragments were generated for 70.2 %, 87.3 %, and 97.0 % of compounds from COCONUT, ChEMBL, and REAL datasets, respectively. Given that RECAP is based on several cleavable bonds, this result shows that such bonds are more likely to be present in synthetic molecules. A total of 205,904 different fragments were obtained for COCONUT, from which 148,560 were unique for this collection. Figure 2a–c shows the chemical structures of the ten most frequent unique fragments from COCONUT, Enamine-REAL, and ChEMBL, respectively. Figure 2d shows the ten most

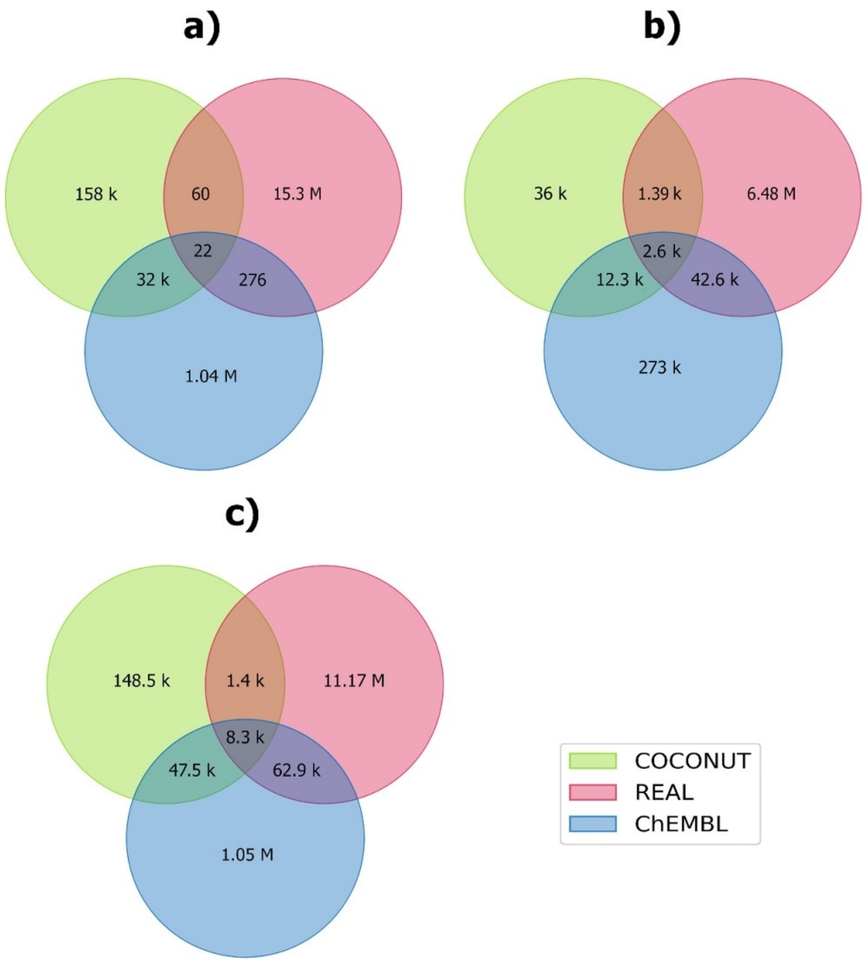


Figure 1. Unique and overlapping structures between COCONUT, ChEMBL and REAL data sets analyzed in this work (Table 1). Structural content was analyzed in terms of **a)** Compounds, **b)** Molecular scaffolds, and **c)** Fragments. The letter **k** represents thousands and the letter **M** represents millions.

Table 2. Summary of the structural diversity of unique and common fragments from COCONUT, Enamine-REAL, and ChEMBL.

Diversity structural	COCONUT*	Enamine-REAL*	ChEMBL*	Overlapping*
Heavy atoms	20.922	19.583	19.784	10.788
Oxygen atoms	3.793	2.080	2.130	1.300
Nitrogen atoms	0.847	3.006	2.562	1.119
Bridgehead atoms	0.282	0.108	0.052	0.020
Spiro atoms	0.110	0.053	0.022	0.001
Rings	2.479	2.377	2.504	1.172
Aromatic rings	0.957	1.341	1.857	0.920
Aliphatic rings	1.522	1.036	0.647	0.252
Heterocycles	1.077	1.556	1.371	0.538
Aromatic heterocycles	0.369	0.862	0.884	0.354
Aliphatic heterocycles	0.707	0.694	0.487	0.184

*Mean of the distribution

common overlapping fragments in all three datasets. The number and frequency of all fragments in each of the three data sets analyzed in this work are included as a separate file in the Supporting Information. Comparison of the chemical structures of unique and common fragments among data sets in Figure 2 and Table 2 indicate that COCONUT fragments (Figure 2a) had the most number of oxygen atoms (hydroxyl, epoxide), chiral centers, aliphatic

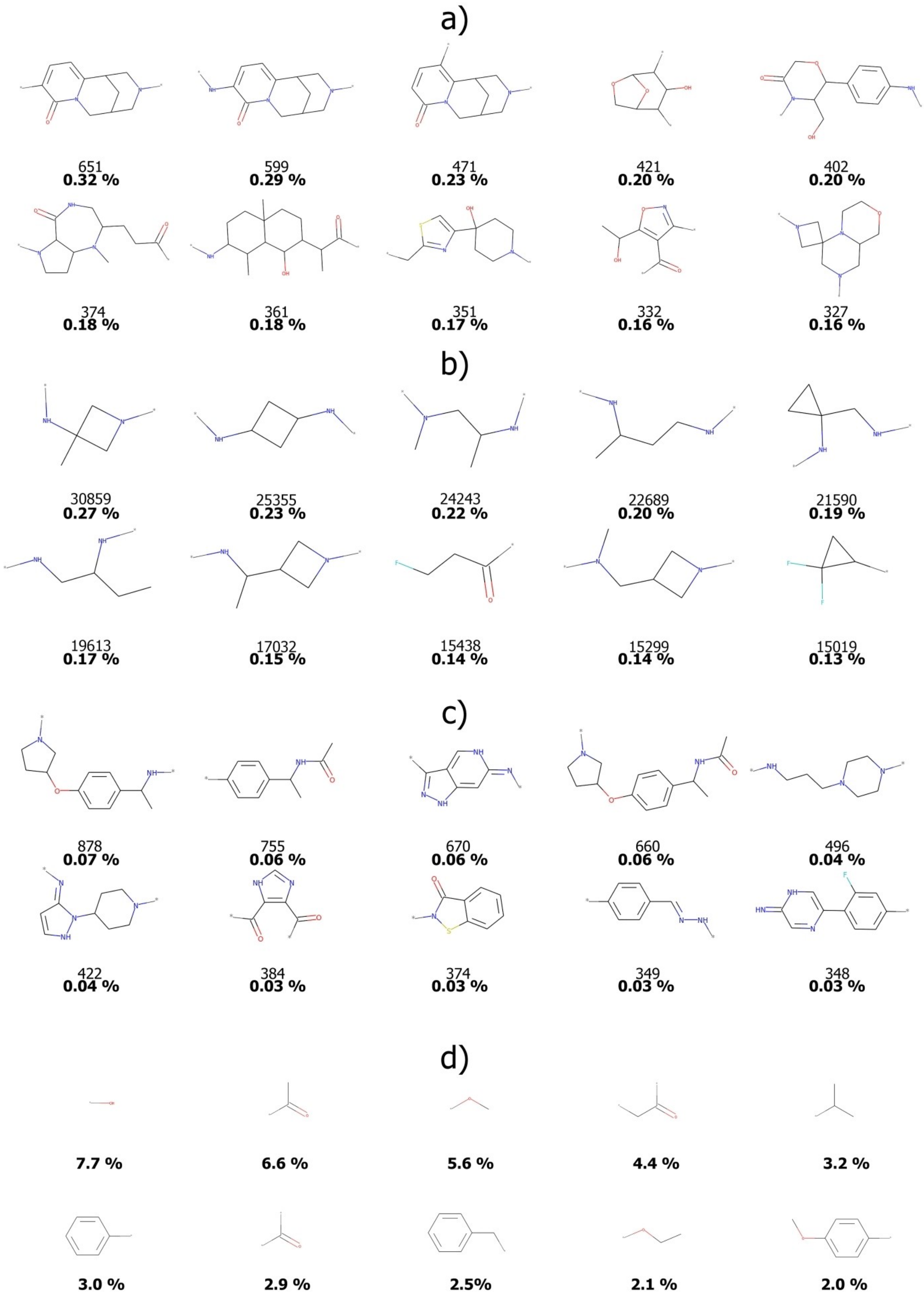


Figure 2. Ten most frequent unique fragments from **a)** COCONUT, **b)** Enamine-REAL, and **c)** ChEMBL.**d)** Ten most frequent common fragments in all three data sets. Occurrences in the data set are indicated in regular letter and percentage in bold.

Table 3. Summary of the diversity and complexity measures of the three fragment data sets.

Fragment data sets*	Size (fragments)	Median similarity (Morgan2 - 1024 bits)	Median similarity (MACCS keys - 166 bits)	Mean fraction of sp^3 carbons	Mean fraction of chiral carbons
COCONUT	205,904	0.117	0.314	0.518	0.175
Enamine, REAL	11,243,078	0.134	0.408	0.516	0.074
ChEMBL	1,177,361	0.122	0.334	0.335	0.046

*Drug-like sets (see Section 2.2).

rings, and bicycles (according to the number of bridgehead and spiro atoms) compared to ChEMBL fragments (Figure 2c), and REAL fragments (Figure 2b). However COCONUT fragments had fewer aromatic rings compared to ChEMBL fragments and REAL fragments. Furthermore, REAL fragments had the most number of nitrogen atoms (e.g. amine) and aromatic heterocycles followed by ChEMBL fragments (e.g. amide). Usually, NPs contain functional groups like oxygen atoms (e.g. hydroxyl, epoxide rings, ester, and peroxide) while synthetic molecules have nitrogen-containing and more easily accessible functional groups like amide, urea, sulfone, imida functionalities, and substituents such as fluoro.[9] This latter observed in REAL fragments and ChEMBL fragments since they contain fluorine substituents (Figure 2b–c). Nevertheless common fragments were characterized by a lower number of aliphatic rings, aromatic rings, bicycles (according to the number of bridgehead and spiro atoms), and relatively greater number of oxygen atoms relative to the number of nitrogen atoms as exemplified in Figure 2d and Table 2. In general, the common fragments to all three data sets (Figure 2d) are smaller in size and less structurally diverse relative to the unique fragments of each data set.

3.3 Diversity and Complexity Analysis

To compare the structural diversity of fragments generated from COCONUT with those generated from the two reference data sets, we computed the median similarity of the pairwise similarity matrix on 10 random sets of 10,000 fragments taken from each dataset,[22] using two molecular fingerprints: MACCs Keys (166-bits) and Morgan2 (1024-bits). The similarity was computed with the Tanimoto coefficient. On the other hand, for comparison of the structural complexity among the data sets, we selected two properties that are relevant in drug discovery,[23] the mean fraction of sp^3 and chiral carbons, computed over the whole fragment libraries. As a reference, we performed the same calculations over the compound data sets. Tables 1 and 3 summarize the statistics of these analyses for the compound and fragment data sets, respectively.

Regarding the structural diversity of the fragment libraries, it was found that COCONUT was the most diverse data set in terms of both MACCS keys and Morgan 2 fingerprints (0.314, 0.117), followed by ChEMBL (0.334, 0.122), and REAL (0.408, 0.134). The same tendency was observed when comparing the compound datasets.

For the measures of the structural complexity of the fragment libraries, COCONUT was found to be the most complex data set in terms of the mean fraction of sp^3 carbons and the mean fraction of chiral carbons (0.518, 0.175), followed by REAL (0.516, 0.074) and ChEMBL (0.335, 0.046), this was determined via a *t*-test with a 99% of confidence. For the compound data sets, the same trend was observed for the mean fraction of chiral carbons, while for the mean fraction of sp^3 carbons the positions of COCONUT and REAL were slightly inverted (0.518, 0.516, Table 3) that can be due to the increased complexity of fragments from NPs. This shows that fragments derived from NPs are structurally more diverse and complex than those obtained from synthetic compounds, preserving the differences associated with the source compounds with complete chemical structures.[5]

3.4 Chemical Space Visualization Based on SB-DFP

As mentioned in section 2.6, two SB-DFPs were built for subsets of NPs and synthetically available compounds derived from COCONUT and REAL, respectively. Different subsets not used in the elaboration of the single fingerprint representations were used to scale the similarity values of compounds to the SB-DFPs and to generate a visual representation of the chemical space covered by compounds. Figure 3 shows the visualization of the chemical space based on SB-DFPs similarities. In the graph, each structure is plotted according to its scaled similarity value to the reference SB-DFPs. The SB-DFPs, as well as the scaling parameters for the similarity values, are included as Supporting Information. To better illustrate the unique structures present in COCONUT and REAL, Figure 3a,b and Figure 3d,e shows unique structures in those data sets, while Figure 3c,g shows all structures from ChEMBL. In each plot of Figure 3, the number of compounds is represented with a continuous color scale from yellow (highly populated regions) to purple (less populated regions). The chemical space visualization of compounds shows that NPs tend to occupy a space closer to the COCONUT SB-DFP (Figure 3a), while synthetically available compounds are closer to the REAL SB-DFP (Figure 3b). Compounds from ChEMBL share space with compounds from both COCONUT and REAL data

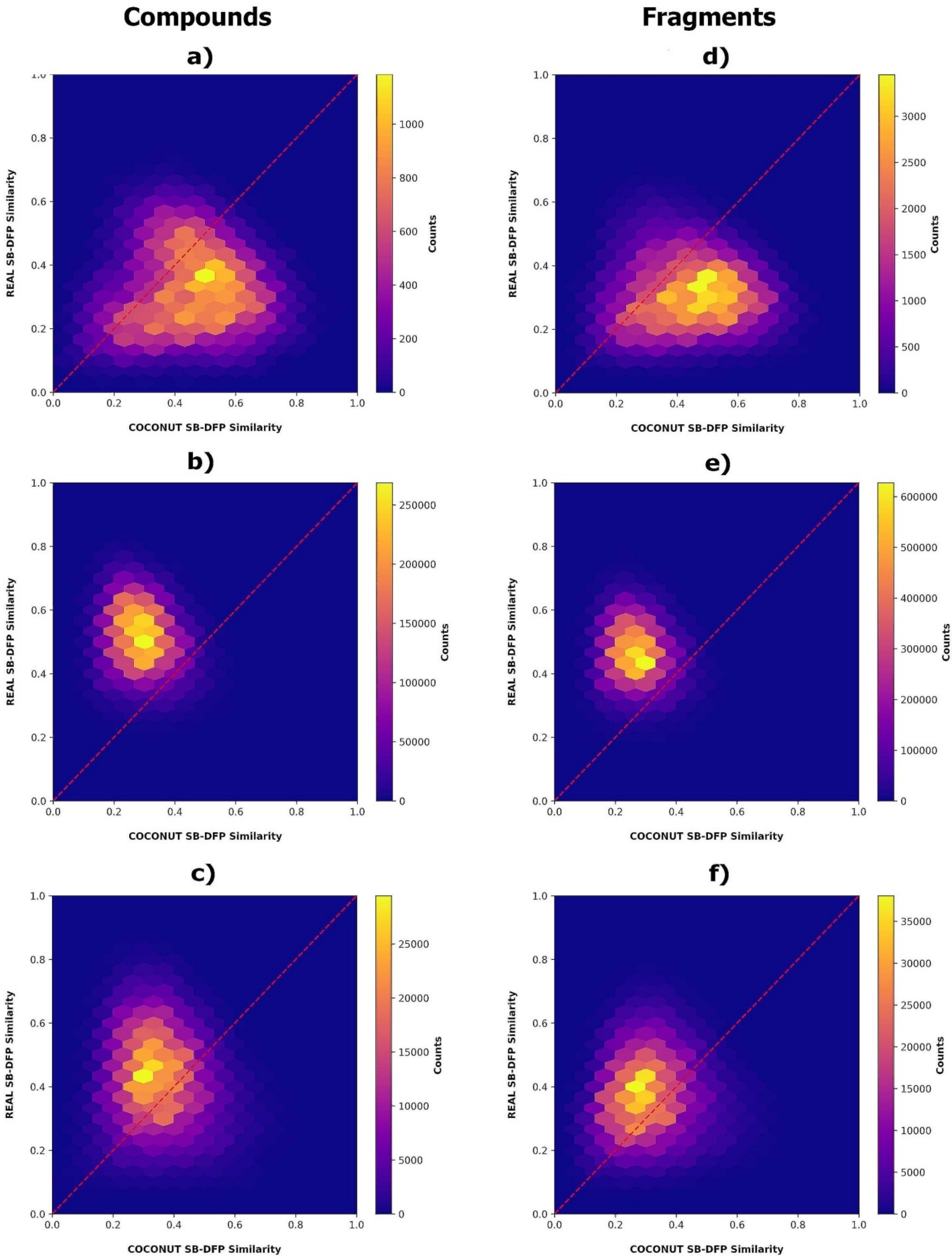


Figure 3. Visual representation of the chemical space for compounds and fragments of natural products, synthetics compounds, and biologically relevant compounds. The number of compounds is represented with a continuous color scale. Compounds data sets used: **a)** COCONUT, **b)** Enamine-REAL, **c)** ChEMBL. Fragment data sets used: **d)** COCONUT, **e)** Enamine-REAL, and **f)** ChEMBL.

sets, being generally closer to the seconds (Figure 3c). Fragments derived from NPs, synthetic compounds, and biologically tested compounds follow the same trend as their source data sets, supporting the idea that the obtained fragments preserve the structural properties of the original compounds from which they were originated.

4 Conclusions

Herein we generated and made publicly available a database of fragments derived from a large collection of drug-like NPs. The NPs-based fragment library was compared with two herein generated fragment libraries obtained from large collections of compounds relevant in drug discovery; one with more than 1 million drug-like compounds tested for biological activity (as presented by ChEMBL), and the second with more than 15 million synthetically accessible yet novel molecules (as represented by the drug-like set of Enamine-REAL). The comparison of the unique and overlapping fragment of NPs with other reference collections revealed that there is a large diversity of unique fragments derived from NPs that could be used as building blocks for the *de novo* design and synthesis of novel compounds. It was also concluded that both the entire structures and fragments derived from NPs are more diverse and structurally complex than the two reference compound collections.

As part of this work, we introduced a novel visual representation of the chemical space based on SB-DFPs. It was concluded that the SB-DFPs developed for NPs and synthetically accessible compounds, respectively, are consistent in that NPs were more similar to the fingerprint generated for COCONUT-SB-DFP and the synthetic compounds were more similar to the REAL-SB-DFP. In this representation of chemical space was concluded that, overall, ChEMBL compounds had higher similarity to the REAL-SB-DFP further emphasizing the opportunity to increase the number of NPs tested for biological activity (e.g., enrich ChEMBL with drug-like compounds available in COCONUT).

Supporting Information

Structure files of all curated data sets and fragment libraries used in this work, as well as the SB-DFPs used for the chemical space visualization are available at https://doi.org/10.6084/m9.figshare.11997951. The Supporting information contains the following:

COCONUT_Compounds.sdf, ChEMBL_Compounds.csv and REAL_Compounds.csv contain the curated structures of drug-like subsets from those major compound data sets. All files contain the following information for each compound: identification number (ID), simplified molecular input line entry system (Smiles), Average Molecular Weight (AMW), partition coefficient octanol/water (SlogP), number of hydrogen bond donors (HBD), number of hydrogen bond acceptors (HBA), number of rotatable bonds (RB), topological polar surface area (TPSA), fraction of sp^3 carbons (FractionCSP3), fraction of chiral carbons (FractionCC), number of generated fragments (NFragments) and a list of the fragments obtained if any (LFragments).

COCONUT_Fragments.sdf, ChEMBL_Fragments.csv and REAL Fragments.csv contain the structures generated from the respective compound data sets. All files include the following information for each fragment: identification number (ID), source collection (Data Set), simplified molecular input line entry system (Fragment), belonging to one (Unique) or the three data sets (Overlapped), number of compounds containing that fragment in the data set (Counts) and fraction of them (Proportion), fraction of sp3 carbons (FractionCSP3), fraction of chiral carbons (FractionCC), number of heavy atoms (NumHeavyAtoms), number of oxygen atoms (NumO), number of nitrogen atoms (NumN), number of bridgehead atoms (NumBridgeHead), number of spiro atoms (NumSpiro), number of rings (NumRings), number of aromatic rings (NumArRings), number of aliphatic rings (NumAlRings), number of heterocycles (NumHet), number of aromatic heterocycles (NumArHet) and number of aliphatic heterocycles (NumAlHet).

SB-DFPs.csv contains the Statistical-Based Database Fingerprints for COCONUT and REAL data sets. The file includes the value for each bit for a Morgan fingerprint of radius 2 (1024-bits) according to RDKit algorithm as well as the empirical minimum and maximum Tanimoto similarity values used for scaling of the data (MinSimilarity and MaxSimilarity).

Conflict of Interest

None declared.

Acknowledgments

NS-C is thankful to CONACyT for the granted scholarship number 335997. We thank *Dirección General de Cómputo y de Tecnologías de Información y Comunicación* (DGTIC), UNAM, for the computational resources to use Miztli supercomputer and UNAM, project LANCAD-UNAM-DGTIC-335.

References

[1] G. M. Rishton, *Am. J. Cardiol.* **2008**, *101*, S43–S49.
[2] D. J. Newman, G. M. Cragg, K. M. Snader, *J. Nat. Prod.* **2003**, *66*, 1022–1037.
[3] D. J. Newman, G. M. Cragg, *J. Nat. Prod.* **2016**, *79*, 629–661.
[4] A. Pahl, H. Waldmann, K. Kumar, *Chimia* **2017**, *71*, 653–660.

[5] M. Feher, J. M. Schmidt, *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 218–227.

[6] F. López-Vallejo, M. A. Giulianotti, R. A. Houghten, J. L. Medina-Franco, *Drug Discovery Today* **2012**, *17*, 718–726.

[7] K. S. Lam, *Trends Microbiol.* **2007**, *15*, 279–289.

[8] A. Christoforow, J. Wilke, A. Binici, A. Pahl, C. Ostermann, S. Sievers, H. Waldmann, *Angew. Chem. Int. Ed.* **2019**, *58*, 14715–14723.

[9] P. Ertl, T. Schuhmann, *J. Nat. Prod.* **2019**, *82*, 1258–1263.

[10] M. Sorokina, C. Steinbeck, *Preprints* **2019**, 2019120332.

[11] Enamine REAL Database. https://enamine.net/library-synthesis/real-compounds/real-database (accessed April 1, 2020).

[12] D. Mendez, A. Gaulton, A. P. Bento, J. Chambers, M. De Veij, E. Félix, M. P. Magariños, J. F. Mosquera, P. Mutowo, M. Nowotka, M. Gordillo-Marañón, F. Hunter, L. Junco, G. Mugumbate, M. Rodriguez-Lopez, F. Atkinson, N. Bosc, C. J. Radoux, A. Segura-Cabrera, A. Hersey, A. R. Leach, *Nucleic Acids Res.* **2018**, *47*, D930-D940.

[13] M. Davies, M. Nowotka, G. Papadatos, N. Dedman, A. Gaulton, F. Atkinson, L. Bellis, J. P. Overington, *Nucleic Acids Res.* **2015**, *43*, W612-W620.

[14] MolVS. https://molvs.readthedocs.io/en/latest/ (accessed April 1, 2020).

[15] X. Q. Lewell, D. B. Judd, S. P. Watson, M. M. Hann, *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 511–522.

[16] G. W. Bemis, M. A. Murcko, *J. Med. Chem.* **1996**, *39*, 2887–2893.

[17] G. M. Keserű, D. A. Erlanson, G. G. Ferenczy, M. M. Hann, C. W. Murray, S. D. Pickett, *J. Med. Chem.* **2016**, *59*, 8189–8206.

[18] J. L. Durant, B. A. Leland, D. R. Henry, J. G. Nourse, *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1273–1280.

[19] D. Rogers, M. Hahn, *J. Chem. Inf. Model.* **2010**, *50*, 742–754.

[20] N. Sánchez-Cruz, J. L. Medina-Franco, https://doi.org/10.1186/s13321-018-0311-x.

[21] G. Schneider, D. E. Clark, *Angew. Chem. Int. Ed.* **2019**, *58*, 10792–10803.

[22] D. K. Agrafiotis, *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 159–167.

[23] O. Méndez-Lucio, J. L. Medina-Franco, *Drug Discovery Today* **2017**, *22*, 120–126.

Received: March 18, 2020
Accepted: April 17, 2020
Published online on April 29, 2020

biomolecules

MDPI

Article

Fragment Library of Natural Products and Compound Databases for Drug Discovery †

Ana L. Chávez-Hernández, Norberto Sánchez-Cruz and José L. Medina-Franco *

DIFACQUIM Research Group, Department of Pharmacy, School of Chemistry,
Universidad Nacional Autónoma de México, Avenida Universidad 3000, Mexico City 04510, Mexico;
anachavez3026@gmail.com (A.L.C.-H.); norberto.sc90@gmail.com (N.S.-C.)
* Correspondence: medinajl@unam.mx; Tel.: +52-55-5622-3899

† This work is dedicated to the memory of José Juan Hernández Hernández.

Received: 19 October 2020; Accepted: 4 November 2020; Published: 6 November 2020

Abstract: Natural products and semi-synthetic compounds continue to be a significant source of drug candidates for a broad range of diseases, including coronavirus disease 2019 (COVID-19), which is causing the current pandemic. Besides being attractive sources of bioactive compounds for further development or optimization, natural products are excellent substrates of unique substructures for fragment-based drug discovery. To this end, fragment libraries should be incorporated into automated drug design pipelines. However, public fragment libraries based on extensive collections of natural products are still limited. Herein, we report the generation and analysis of a fragment library of natural products derived from a database with more than 400,000 compounds. We also report fragment libraries of a large food chemical database and other compound datasets of interest in drug discovery, including compound libraries relevant for COVID-19 drug discovery. The fragment libraries were characterized in terms of content and diversity.

Keywords: chemoinformatics; COVID-19; drug discovery; drug design; fingerprint; food chemicals; natural products fragments; SARS-CoV-2

1. Introduction

Natural products (NP) have long been studied and used in medicine and chemistry, starting from ancient civilizations throughout history. Natural sources were the basis of early research in medicinal chemistry and drug discovery and have yielded valuable therapeutic agents still in use today [1]. A recent review reveals that 3.8% of drugs approved between 1981 and 2019 are NP, and 18.9% are NP derivatives [2].

The unique and complex chemical structures of NP make them unique sources to explore novel areas of the chemical space [3]. However, considering the structural complexity of NP, it is a challenge to produce them in large quantities, which is typically required during drug development. Therefore, in recent years novel methods and synthetic strategies have been developed to obtain diverse and semi-synthetic compounds libraries based on NP [4]. Similarly, NP are becoming attractive starting points to conduct fragment-based drug design and build the so-called “pseudo-NPs” [5].

The increasing use of NP in modern drug discovery has promoted the application of chemoinformatic methods for natural product-based drug discovery. One such contribution is the generation and development of compound databases [6–8]. The development of compound databases of NP and synthetic analogs has been recently reviewed [8,9]. A recent notable example is the COlleCtion of Open NatUral producTs (COCONUT), a compendium of 50 open-access databases collecting more than 400,000 compounds. These and other public collections of food chemicals are important sources to generate fragment libraries of compounds of natural origin. The authors recently reported and made public a library with 205,903 fragments derived from a drug-like subset of the first version of COCONUT [10]. In that work, a total of 190,139 molecules were analyzed. Recently COCONUT was updated, and a fragment library based on its full comprehensive collection has not been reported.

The goal of this work was to generate a fragment library of the complete and most recent version of COCONUT that contains 432,706 compounds. We also expanded the analysis to generate fragment libraries of large public collections of 23,883 food chemicals that have a close association with NP [11] and are part of the increasing research field of *foodinformatics* [12]. The fragment libraries were characterized using chemoinformatic methods and compared with reference fragment libraries generated from molecules in the Dark Chemical Matter (DCM). DCM is a collection of 139,352 compounds that showed no activity when tested in at least 100 screening assays but that have recently led to the identification of bioactive compounds [13]. In light of the current coronavirus disease 2019 (COVID-19) pandemic, we also included in this study two large reference libraries with relevance in drug discovery in relation to this disease [14]. Of note, food chemicals and DCM compounds analyzed in this work were recently screened in silico to identify potential inhibitors of the main protease of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), one of the main promising molecular targets for the treatment of COVID-19 [15].

2. Materials and Methods

2.1. Compound Databases

In this work, we generated fragment libraries of five compound databases of interest in drug discovery, summarized in Table 1 and listed here: COCONUT, the largest database, with a total of 423,706 unique molecules [16], Food Database (FooDB) with 23,883 food chemicals [17], and a database with 139,352 small molecules, classified as DCM [13]. We also analyzed a focused public library relevant to COVID-19 research assembled by the Chemical Abstract Service (CAS) with 48,876 compounds [18] and 280 inhibitors of the main protease of SARS-CoV-2 (3CLP) [15].

Table 1. Compound data sets analyzed in this work.

Dataset	Original Compounds	Processed Compounds	Generated Fragments	Reference
COCONUT	432,706	382,248	52,630	[16]
FooDB	23,883	21,319	3186	[17]
Dark Chemical Matter (DCM)	139,352	139,326	14,001	[13]
Chemical Abstract Service (CAS) set focused on COVID-19	48,876	44,692	8432	[18]
Inhibitors of the main protease of SARS-CoV-2 (3CLP)	280	256	108	[15]

COCONUT, COlleCtion of Open NatUral producTs, FooDB, Food Database (FooDB, COVID-19, coronavirus disease 2019, SARS-CoV-2, severe acute respiratory syndrome coronavirus 2.

2.2. Data Curation

Similar to our previous work [10], the preparation of the five datasets was performed with the open-source cheminformatics toolkit RDKit [19], (version 2020.03.2.0, RDKit, San Francisco, CA, USA) and the functions Standardizer, LargestFragmentChoser, Uncharger, Reionizer, and TautomerCanonicalizer implemented in the molecule validation and standardization tool MolVS [20]. SMILES strings [21], with no stereochemistry information, were generated because not all compounds in the datasets have a defined stereochemistry. Compounds with valence errors or any chemical element other than H, B, C, N, O, F, Si, P, S, Cl, Se, Br, and I were removed. With the chemical compounds retained, neutralized, and reionized, a canonical tautomer was generated. The average molecular weight (AMW) was calculated, and all compounds with AMW ≤ 1300 were retained. Table 1 summarizes the number of compounds used for the fragmentation analysis and the number of unique fragments generated.

2.3. Generation of Unique Fragments Using the RECAP Algorithm

Fragment libraries were produced with the Retrosynthetic Combinatorial Analysis Procedure (RECAP) as implemented in RDKit (version 2020.03.2.0, RDKit, San Francisco, LA, USA). The RECAP algorithm is based on 11 cleavage rules derived from chemical reactions [22]. A molecule is cleaved into fragments if it contains any of the following bonds: amide, ester, amine, urea, ether, olefin, quaternary nitrogen, aromatic nitrogen–aliphatic carbon, lactam nitrogen–aliphatic carbon, aromatics carbon–aromatic carbon, and sulphonamide. For this study, only terminal fragments were generated.

All curated datasets and fragments libraries used in this work are available at https://doi.org/10.6084/m9.figshare.13064231.v1. Datasets contain the curated structures and the following information: identification number (ID), simplified molecular input line entry system (Smiles), Average Molecular Weight (AMW), number of carbons, oxygens, nitrogens, heavy atoms, aliphatic rings, aromatic rings, heterocycles and bridgehead atoms, fraction of sp^3 carbon atoms and chiral carbons, and a list of fragments generated from each compound. Fragment libraries contain structures generated (Fragments) from each compound library (Dataset) and the following information: number of compounds that contain that fragment in a dataset (Count) and fraction of them (Proportion), Average Molecular Weight (AMW), number of carbons, oxygens, nitrogens, heavy atoms, aliphatic rings, aromatic rings, heterocycles and bridgehead atoms, fraction of sp^3 carbon atoms and chiral carbons.

2.4. Structural Diversity and Complexity

The structural diversity of the compounds and fragment datasets was evaluated by calculating the median value of the distribution of the pairwise similarity values generated with the Tanimoto coefficient for both Morgan fingerprint with radius 2 (Morgan2, 1024-bits) [23] and Molecular ACCes System (MACCS) keys (166-bits) [24]. For 4 sets of entire compounds (except 3CLP), the calculation was done for 10 random samples of 10,000 compounds each, and the medians were then averaged. For 3CLP, all 256 molecules were used. For the fragment datasets, all fragments were employed for the calculation, except for COCONUT, for which 10 random samples of 10,000 fragments were used. It has been shown that for large datasets, several random samples of 1000 compounds each are a reasonable approach to quantify the pairwise fingerprint-based diversity of the entire datasets [25].

The structural differences between compound and fragment datasets were evaluated, calculating 14 molecular descriptors, namely, number of carbon, oxygen, nitrogen, and heavy atoms, the number of rings and heterocycles—both aliphatic and aromatic—spiro atoms, bridgehead atoms, the fraction of sp^3 carbons, and chiral carbons.

2.5. Chemical Space Visualization

Morgan fingerprints with radius 2 (Morgan2, 1024-bits) were generated for each compound and fragment data set. To generate a visual representation of the chemical space, we used the recently developed algorithm TMAP (Tree MAP). This method allows the visual representation of many molecules that are difficult to visualize using other standard methods such as principal component analysis. Basically, TMAP allows the visualization of large data sets (such as the ones studied in this work—Table 1) through the distance between the clusters and the cluster's detailed structure through branches and sub-branches [26,27]. Fingerprints for each data set (input data) were indexed in a local sensitive hashing (LSH) forest data structure, enabling c-approximate k-nearest neighbor (k-NN). Fingerprints were encoded using the MinHash algorithm. An undirected weighted c-approximate k-nearest neighbor graph (c-k-NNG) is constructed from the data points indexed in the LSH forest. This graph takes two arguments, k, the number of nearest-neighbors, and kc, the factor used by the augmented query algorithm. In this work, we used k = 50 and kc =10. Further details of the TMAP approach are published elsewhere [28].

3. Results and Discussion

3.1. Overlapping Fragments and Compounds

Figure 1 shows the number of unique and overlapping compounds and fragments. We found 533,961 unique compounds among all datasets which comprising 364,070 COCONUT compounds (93.54%), 352 from FooDB (1.6%), 134,251 from DCM (96.35%), 35,070 from CAS (78.31%), and 218 from 3CLP (85.15%). The largest compound overlap occurred between COCONUT and FooDB (21,591 (98.37%) FooDB compounds in COCONUT). The second largest overlap was between COCONUT and 3CLP (concerning 35 (13.67%) 3CLP compounds), followed by the overlaps between COCONUT and DCM (concerning 3693 (2.65%) DCM compounds) and COCONUT and CAS (concerning 361 (0.26%) CAS compounds).

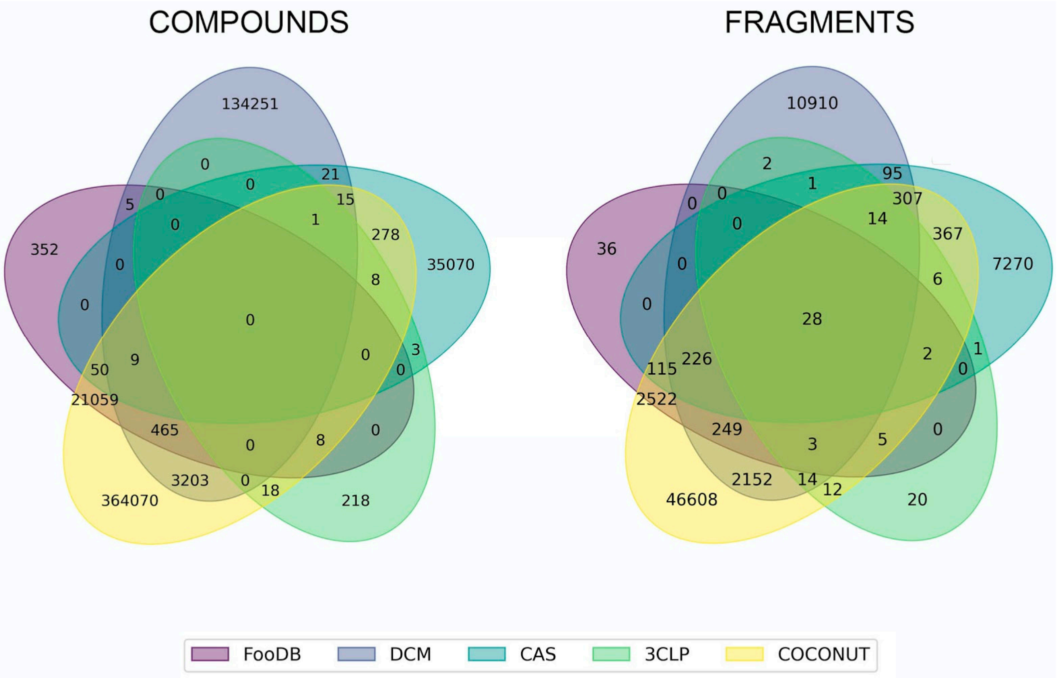


Figure 1. Unique and overlapping compounds and fragments from COCONUT, FooDB, DCM, CAS, and 3CLP. Compounds and fragments are represented with colors: yellow (COCONUT), violet (FooDB), purple (DCM), green (CAS), and lime (3CLP).

Regarding the fragments, Figure 1 indicates that there we identified 64,844 unique fragments among all datasets, including 46,608 COCONUT fragments, 36 FooDB fragments (1.12%), 10,910 DCM fragments (77.92%), 7270 CAS fragments (86.21%), and 20 3CLP fragments (18.51%). The largest fragment overlap occurred for 3150 FooDB fragments (98.87%) overlapped with COCONUT fragments, followed by 84 3CLP fragments (77.77%), 2993 DCM fragments, and 1065 CAS fragments. We also found that 28 fragments were shared by all fragment libraries (Figure 1).

It should be noted that around 13% of 3CLP inhibitors are found within a global dataset of NP. Likewise, 77% of 3CLP fragments can be obtained from NP. This observation reinforces our hypothesis that previously isolated and characterized NP are potential sources of compounds against COVID-19. In turn, this is also in agreement with several reports of virtual screenings of NP databases aimed to identify compounds with activity against 3CLP [29,30].

3.2. Fragment Analysis

As described in the Methods Section 2.3, molecular fragments (terminal fragments only) were obtained from the five compound datasets. The NP fragments in COCONUT and the food chemicals in FooDB were compared with molecules of three reference datasets: small molecules with no biological activity despite having been exhaustively tested in high-throughput screening (HTS) and two collections for COVID-19 drug discovery. Table 1 summarizes the results. The largest number of different fragments was generated for COCONUT (52,630), while the smallest number of fragments was calculated for 3CLP (108). Figures 2–6 show the chemical structures of the 10 most frequent and unique fragments in the 5 databases studied. The figure indicates the frequency and percentage of each fragment in the corresponding dataset.

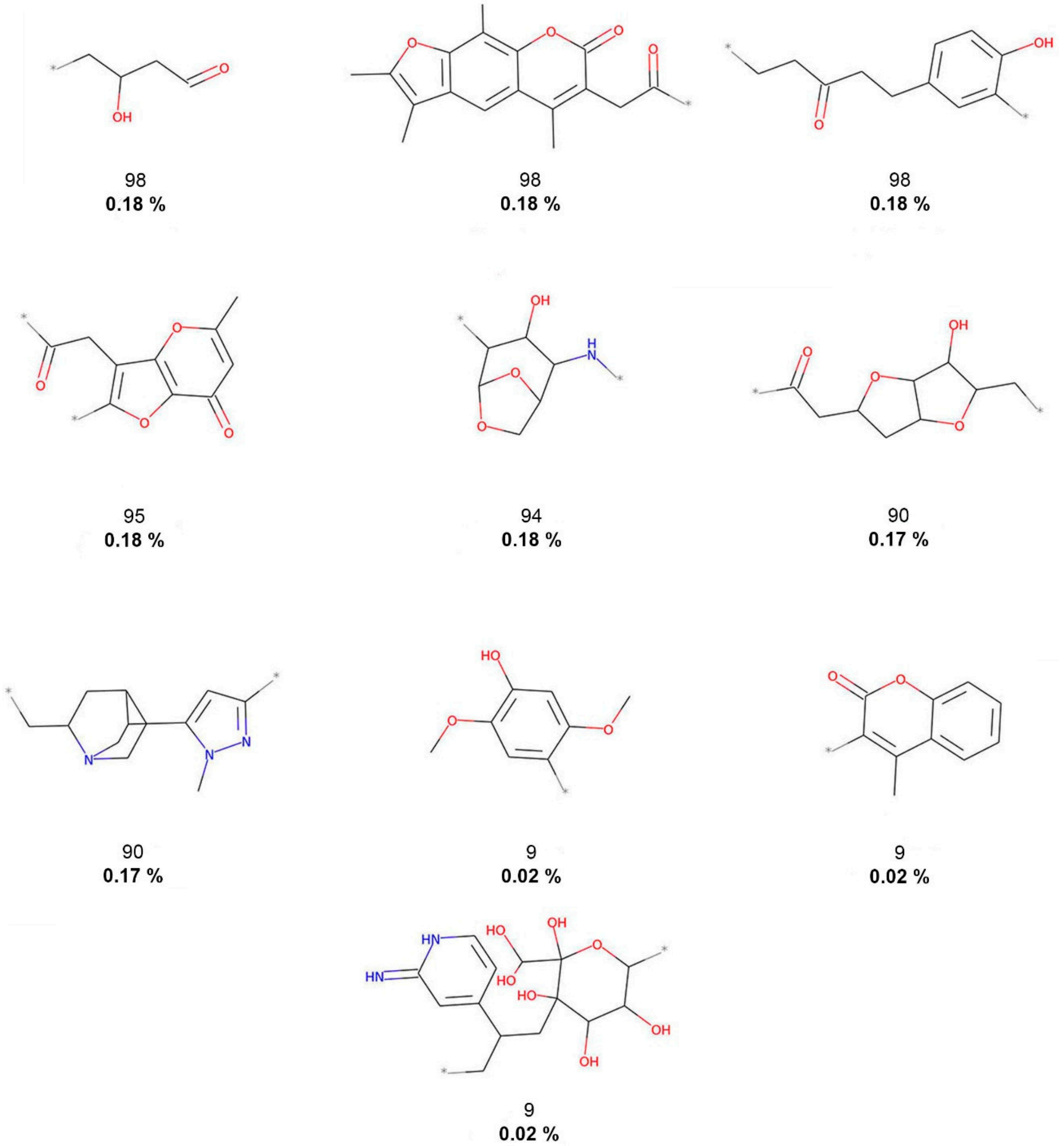


Figure 2. The 10 most frequent and unique COCONUT fragments. Frequency (regular font) and proportion (bold font) are listed below the chemical structures.

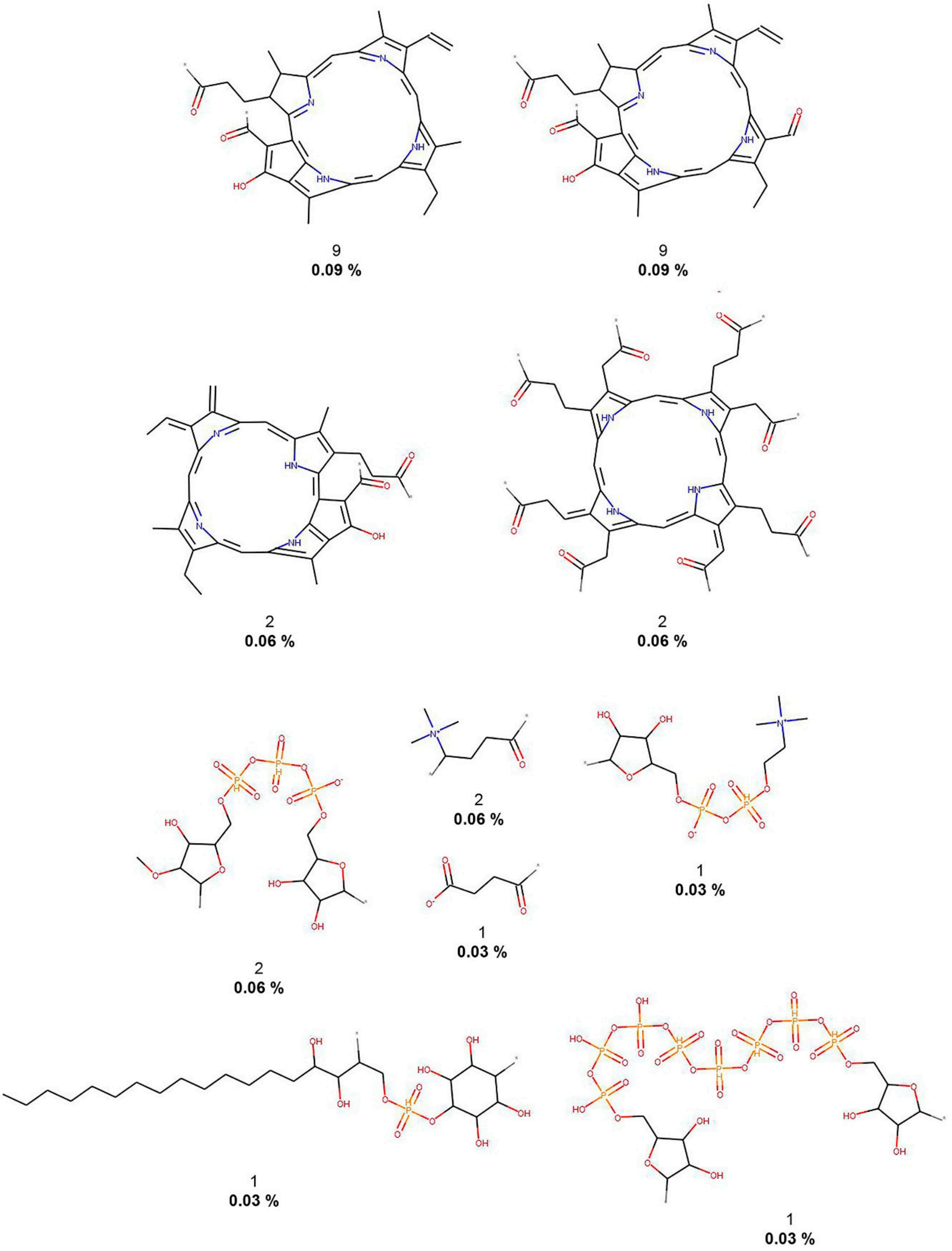


Figure 3. The 10 most frequent and unique FooDB fragments. Frequency (regular font) and proportion (bold font) are listed below the chemical structures.

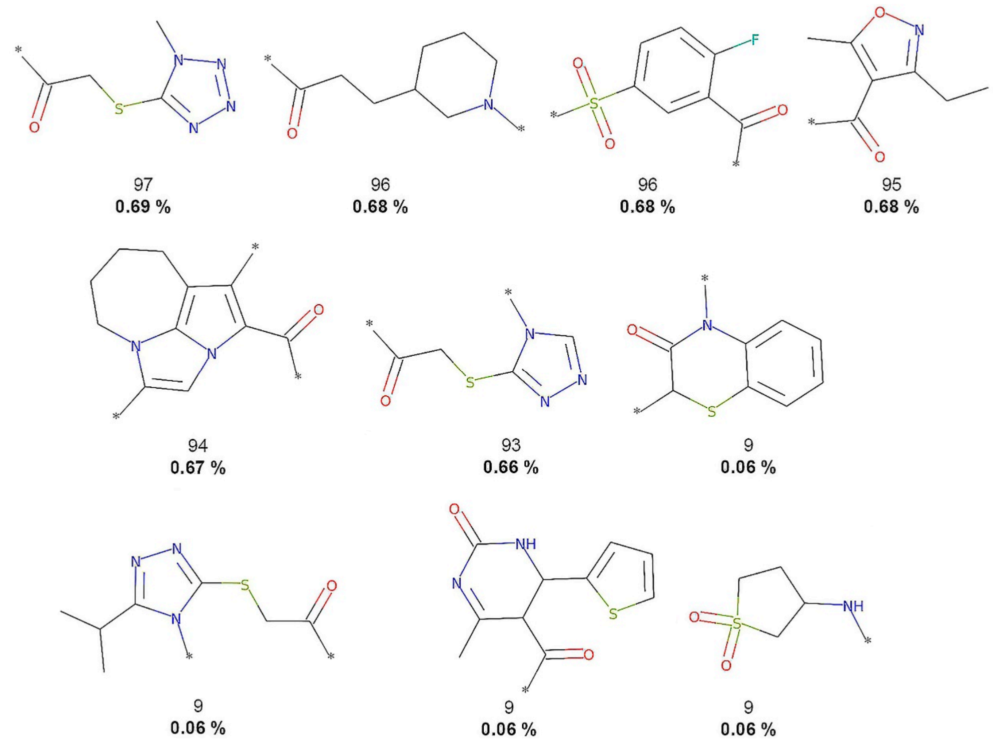


Figure 4. The 10 most frequent and unique DCM fragments. Frequency (regular font) and proportion (bold font) are listed below the chemical structures.

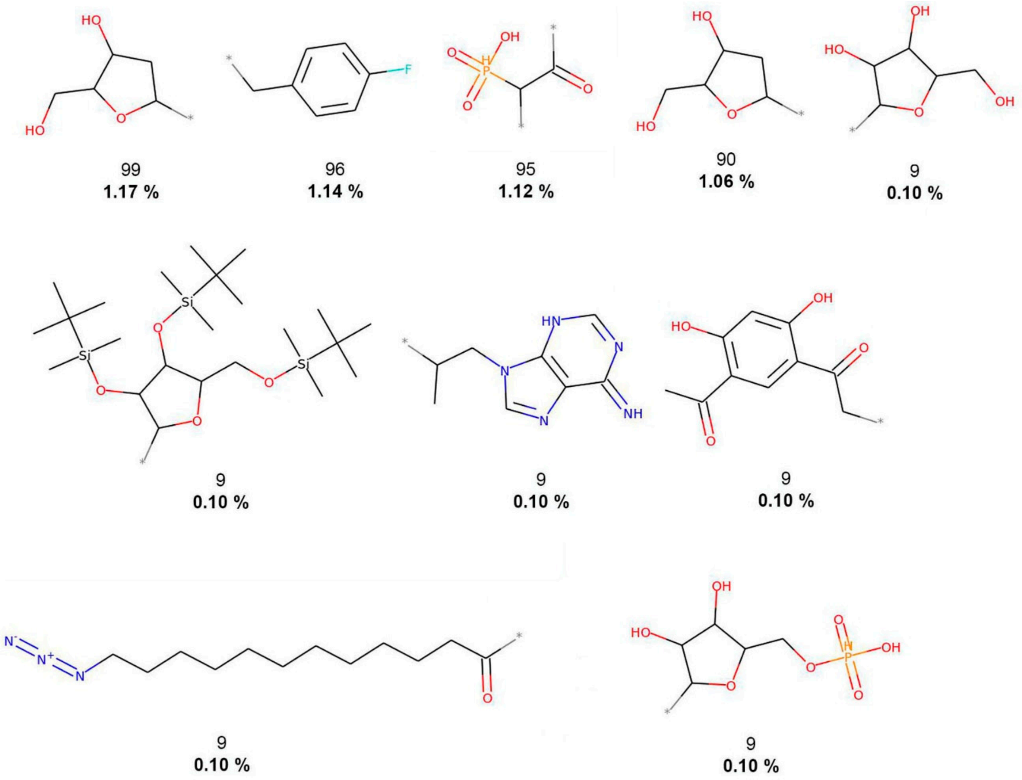


Figure 5. The 10 most frequent and unique CAS fragments. Frequency (regular bond) and proportion (bold font) are listed below the chemical structures.

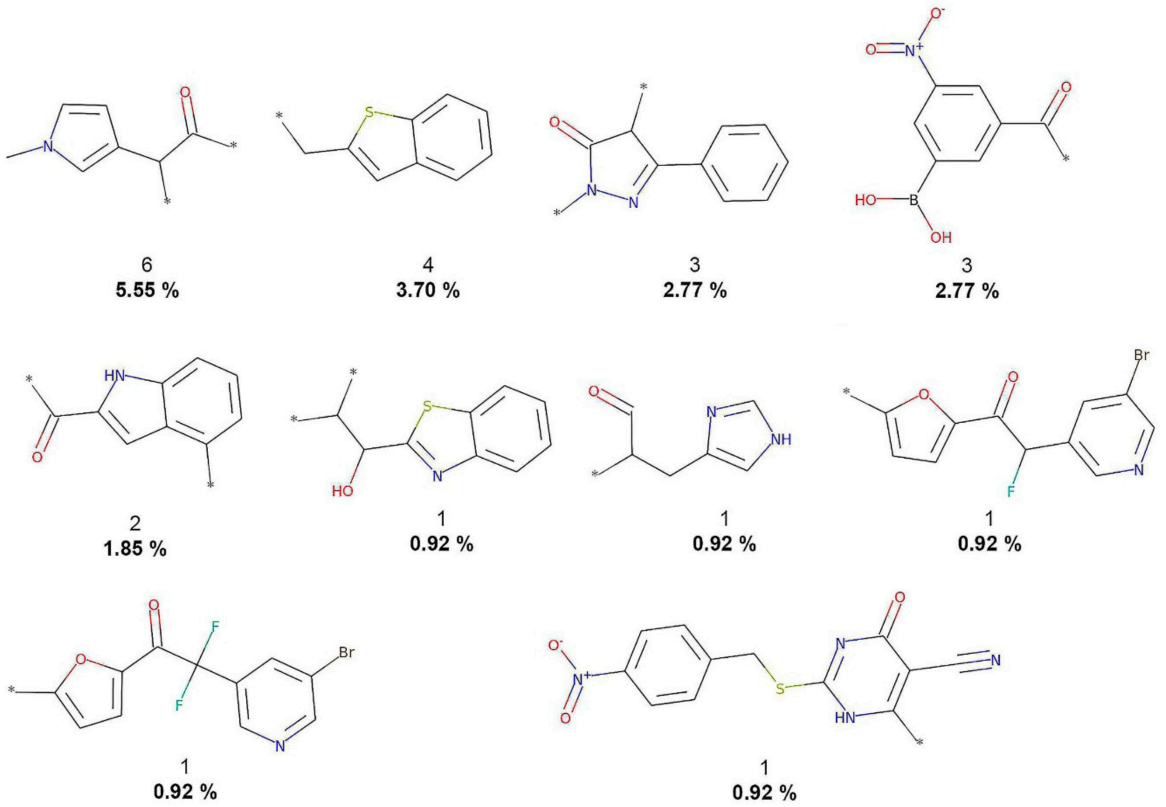


Figure 6. The 10 most frequent and unique 3CLP fragments. Frequency (regular bond) and proportion (bold font) are listed below the chemical structures.

Figure 2 shows that COCONUT fragments contain the largest number of oxygen atoms (carbonyls, alcohols, and aldehydes), aliphatic rings, like tetrahydrofurans and pyranones, and other oxygen-containing heterocycles. FooDB fragments are characterized by having macrocycles (porphyrin rings) and triphosphates groups (Figure 3). In contrast, fragments from CAS, 3CLP, and DCM have larger numbers of nitrogen atoms and aromatic rings than fragments from COCONUT and FooDB as shown in Figures 4–6. The most frequent DCM fragments contain various triazole and pyrimidine rings, and 3CLP fragments comprise pyrrole, imidazole, and pyrazole rings.

The chemical structures of the 28 fragments common (overlap) to all five data sets (Figure 1) are represented in Figure 7, which shows the sum of frequencies of each fragment in all databases and the cleavage bonds in gray color (also marked with *). Relevant overlapping fragments include acetophenones (5642, 1377, and 647), 2-acetylfuran (2156), cyclopropyl methyl ketone (12,223), benzylacetone (493, 419), 2-acetylthiophene (1101 and 11), 2-aminohexane-2,5-dione (98), 2-aminoacetophenone (74), 2-acetylindole (57).

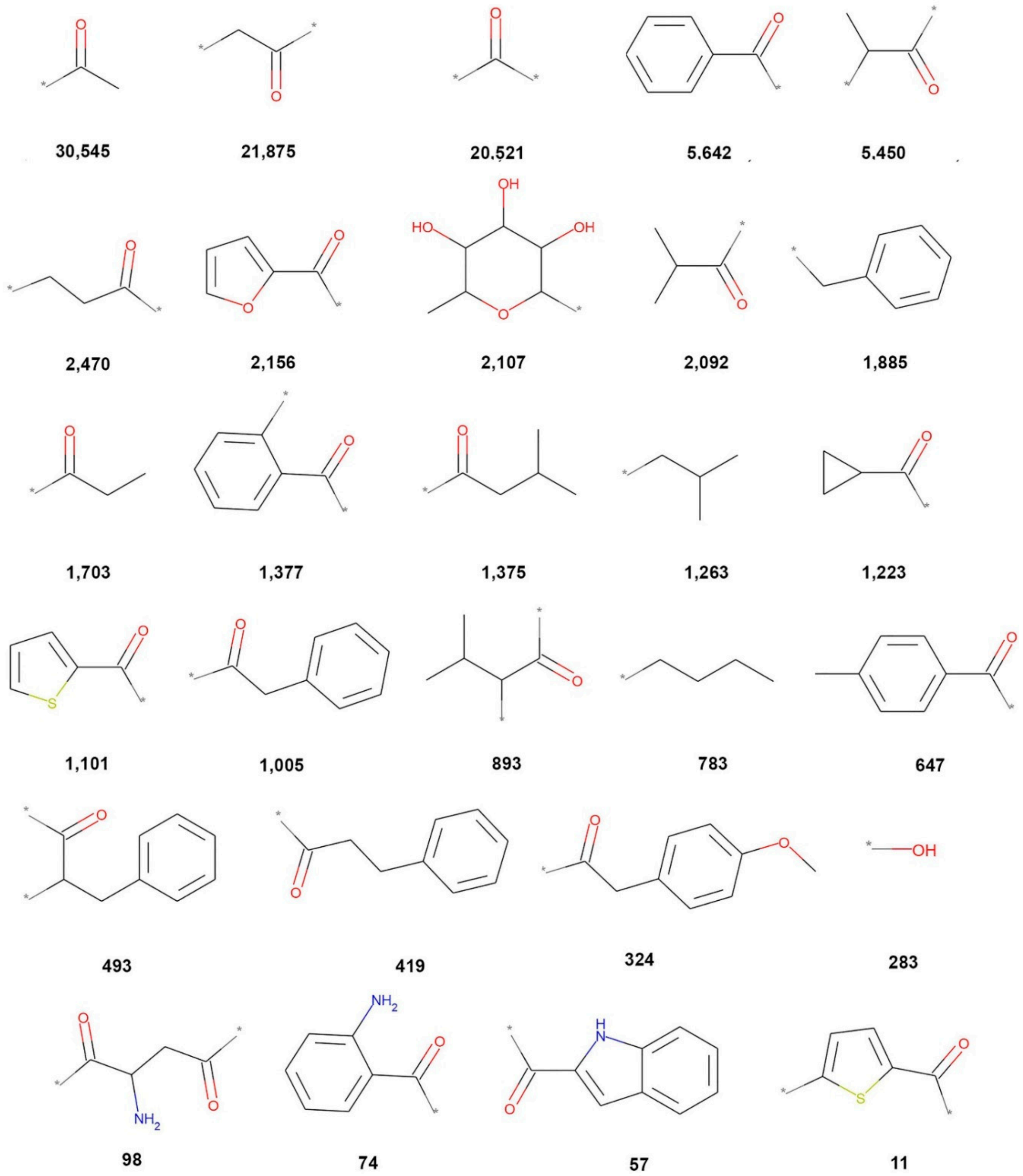


Figure 7. Overlapping fragments between COCONUT, FooDB, DCM, CAS, and 3CLP. The sum of frequencies of each fragment in all databases is indicated in bold font.

Tables 2 and 3 summarize the distribution of carbon, oxygen, nitrogen, and heavy atoms for the entire compounds and fragment datasets, respectively. The tables also summarize the fraction of sp^3 carbon atoms and chiral carbons as representative structural complexity measures. Finally, both tables indicate the distribution of the number of rings (total number, aliphatic, and aromatic) and other important structural features of the compound and fragment datasets. Table 2 shows that compounds from COCONUT and FooDB have the highest mean fraction of sp^3 carbons, 0.506 and 0.620, respectively, whose values range from 0.45 and 0.59 for NPs [31]. CAS, DCM, and 3CLP show the largest number of aromatic rings and aromatic heterocycles, which are characteristic of drugs and synthetic compounds [32]. Compounds in COCONUT and FooDB have the largest number of carbon and oxygen atoms, fraction of chiral carbons, and number of aliphatic rings and bridgehead atoms, a trend that is preserved for their respective fragments (see Table 3). However, fragments from COCONUT and FooDB overlapping with those from CAS, DCM, and 3CLP have the lowest number of carbon, oxygen, and aliphatic rings, compared to unique fragments (Table 3).

Table 2. Summary of the structural composition of compounds from COCONUT, FooDB, and reference datasets [a].

Structural Feature	COCONUT	FooDB	DCM	CAS	3CLP
Carbon atoms	25.640	26.563	18.059	22.496	25.828
Oxygen atoms	6.167	7.343	3.252	5.773	4.922
Nitrogen atoms	1.445	0.668	2.859	4.157	3.582
Heavy atoms	33.611	34.942	25.139	33.535	35.352
Fraction of sp^3 carbons	0.506	0.620	0.342	0.489	0.291
Fraction of chiral carbons	0.154	0.152	0.028	0.145	0.069
Rings	3.962	2.243	2.881	3.628	3.617
Aliphatic rings	2.250	1.426	0.791	1.372	0.645
Aromatic rings	1.712	0.817	2.089	2.256	2.973
Heterocycles	1.711	1.020	1.408	2.056	1.500
Aliphatic heterocycles	1.166	0.770	0.619	0.865	0.363
Aromatic heterocycles	1.712	0.817	2.089	2.256	2.973
Spiro atoms	0.167	0.051	0.018	0.019	0.000
Bridgehead atoms	0.493	0.137	0.056	0.254	0.023

[a] Mean of the distribution.

Table 3. Summary of the structural composition of fragments from COCONUT, FooDB, CAS, DCM, and 3CLP and overlapping fragments [a].

Structural Feature	COCONUT	FooDB	DCM	CAS	3CLP	Overlapping Fragments
Carbon atoms	18.504	12.991	10.181	9.904	8.926	5.179
Oxygen atoms	3.524	3.173	1.748	3.678	1.556	1.107
Nitrogen atoms	0.795	0.394	1.475	0.883	0.713	0.107
Heavy atoms	23.034	16.760	14.057	15.532	11.537	6.464
Fraction of sp^3 carbons	0.557	0.615	0.330	0.656	0.298	0.318
Fraction of chiral carbons	0.189	0.199	0.054	0.240	0.071	0.062
Rings	2.999	1.739	1.686	1.496	1.398	0.571
Aliphatic rings	2.013	1.237	0.447	0.837	0.398	0.071
Aromatic rings	0.986	0.503	1.239	0.660	1.000	0.500
Heterocycles	1.087	0.577	0.899	0.787	0.574	0.179
Aliphatic heterocycles	0.751	0.390	0.313	0.573	0.176	0.036
Aromatic heterocycles	0.986	0.503	1.239	0.660	1.000	0.500
Spiro atoms	0.190	0.085	0.013	0.010	0.000	0.000
Bridgehead atoms	0.507	0.288	0.043	0.109	0.056	0.000

[a] Mean of the distribution.

In general, NPs have been reported to have a higher fraction of sp^3 carbons (associated with a greater structural complexity) and number of oxygen atoms and a lower number of nitrogen atoms and aromatics rings as well as NP fragments [31,33]. Therefore, the fragments from COCONUT and FooDB are also attractive as building blocks for designing drug candidates.

3.3. Structural Diversity and Complexity

The fingerprint-based structural diversity was measured as the median value of the distribution of the pairwise similarity values calculated with the Tanimoto Coefficient, both MACCS keys and Morgan2 (see Methods, Section 2.4). The results are summarized in Tables 4 and 5. Regarding the diversity of the compound libraries, FooDB was the most diverse in terms of Morgan2 and MACCS keys fingerprints (median similarity of 0.092, 0.322), followed by COCONUT (0.107, 0.380) (Table 4). The structural diversity of the most recent version of COCONUT (studied in this work) is similar to the fingerprint diversity calculated for a drug-like subset of COCONUT (0.117, 0.314) computed recently [10]. CAS appeared to be one of the least diverse sets, which is consistent because the datasets were selected by focusing on COVID-19 research (vide supra).

Table 4. Summary of the fingerprint-based structural diversity of the entire compounds.

Dataset	Morgan2 [a] (1024-bits)	MACCS Keys [a] (166-bits)
COCONUT	0.107	0.380
FooDB	0.092	0.322
DCM	0.136	0.407
CAS	0.117	0.473
3CLP inhibitors	0.127	0.403

[a] Median similarity.

Table 5. Summary of the fingerprint-based structural diversity of the fragment datasets.

Dataset of Fragments	Morgan2 [a] (1024-bits)	MACCS Keys [a] (166-bits)
COCONUT	0.111	0.300
FooDB	0.106	0.241
DCM	0.125	0.243
CAS	0.095	0.222
3CLP inhibitors	0.147	0.214

[a] Median similarity.

Regarding the fingerprint-based diversity of the fragment datasets (Table 5), in general, all fragment libraries showed a larger diversity than their parent compounds. Specifically, the CAS fragments were the most diverse according to both molecular fingerprints (0.094, 0.222), followed by FooDB (0.106, 0.241) and COCONUT (0.111 only for Morgan2). Possibly, the difference in the diversity of the fragments from NP in COCONUT and food chemicals in FooDB is associated with the fragmentation algorithm (i.e., the RECAP fragmentation algorithm terminal fragments only as compared to our previous work [10]). This result means that the diversity of fragments appears in the intermediate compounds generated throughout the fragmentation process.

3.4. Chemical Space Visualization

A visual representation of the chemical space of the entire compounds and fragments was explored using the TMAP approach, as described in Methods, Section 2.5. Of note, TMAPs facilitate the visualization of very large datasets (e.g., more than 380,000 molecules from COCONUT, Table 1). The visual representation of the chemical space for the entire compounds and fragments is shown in Figures 8 and 9, respectively. The figures display the chemical space of all compounds and fragments using the same coordinates. To improve the visualization's clarity, each set of unique compounds and fragments from the five datasets is shown individually. The figures also present three panels showing direct comparisons of COCONUT with the other datasets, highlighting in different colors the compounds that are in common, i.e., COCONUT–FooDB (purple); COCONUT–CAS (black); COCONUT–DCM (green), and COCONUT–3CLP (magenta).

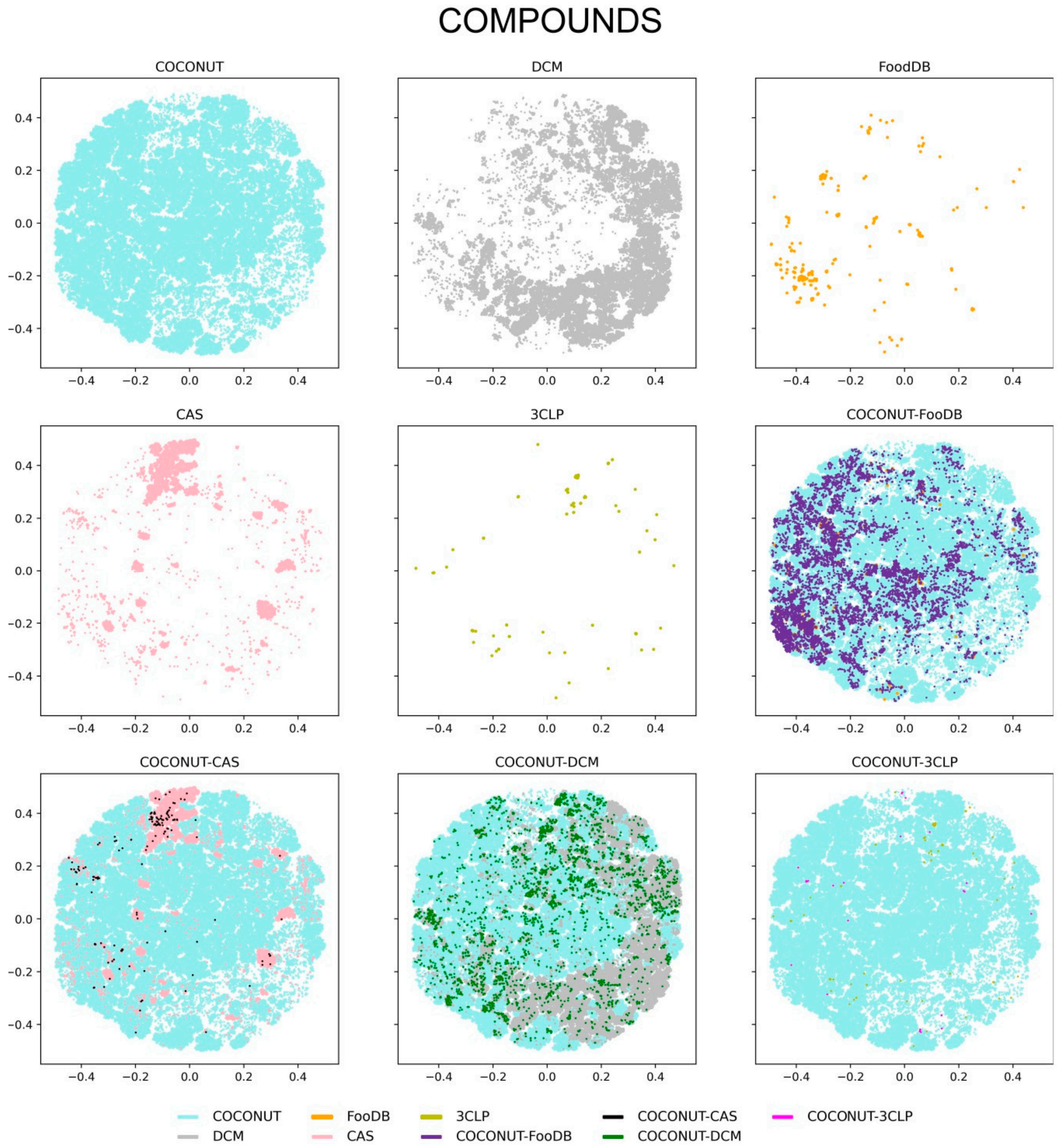


Figure 8. Visualization of the chemical space of the compound datasets generated with Tree Maps. Datasets are represented with colors: COCONUT (cyan), DCM (gray), FooDB (orange), CAS (pink), and inhibitors of the main protease of SARS-CoV-2, 3CLP, (olive). Overlapping compounds in COCONUT–FooDB (purple), COCONUT–CAS (black), COCONUT–DCM (green), and COCONUT–3CLP (magenta) are indicated.

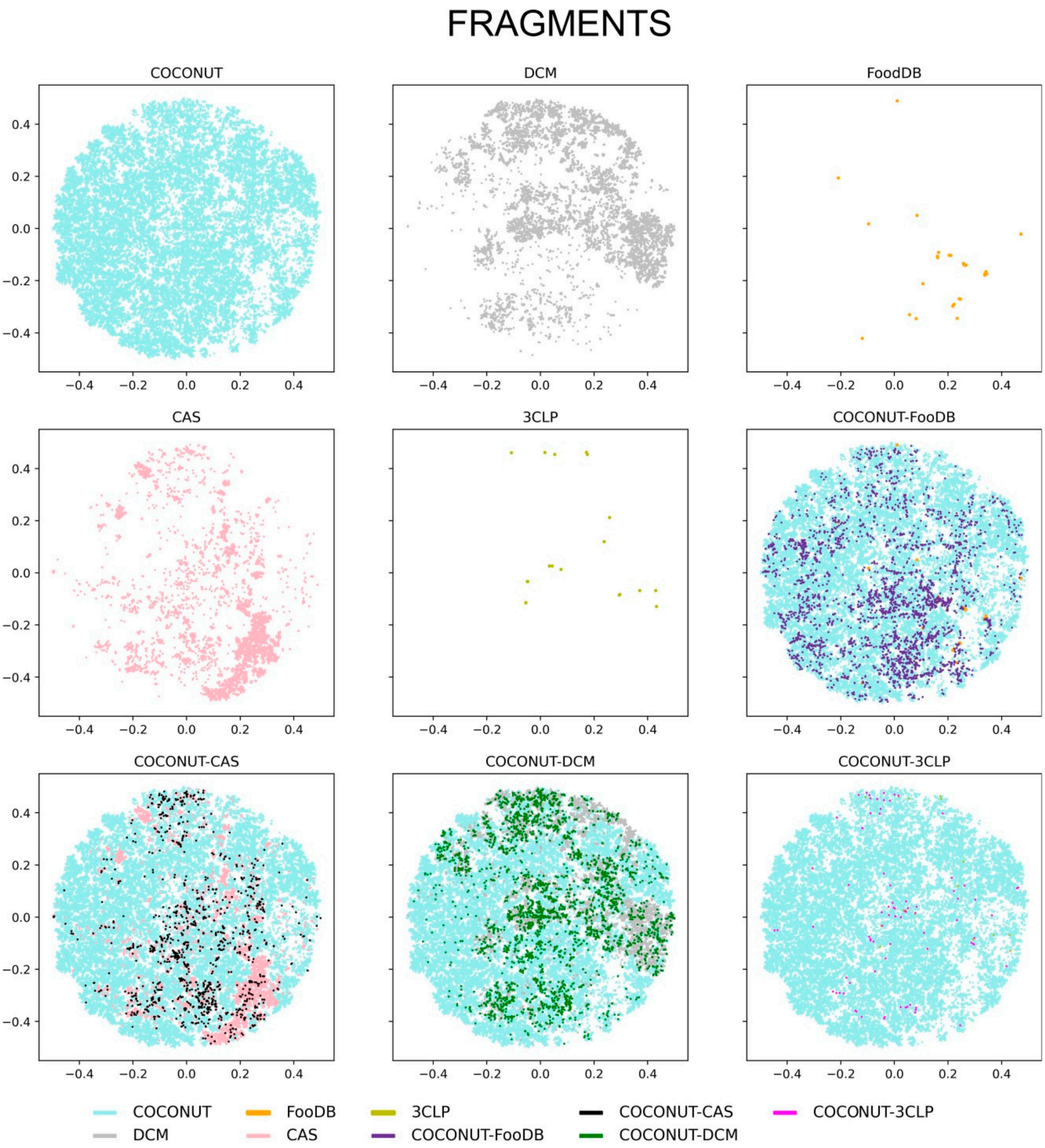


Figure 9. Visualization of the chemical space of fragments generated with Tree Maps. Datasets are represented with colors: COCONUT (cyan), DCM (gray), FooDB (orange), CAS (pink), and inhibitors of the main protease of SARS-CoV-2, 3CLP, (olive). Overlapping fragments in COCONUT–FooDB (purple), COCONUT–CAS (black), COCONUT–DCM (green), and COCONUT–3CLP (magenta) are indicated.

Figure 8 shows that all compound datasets converged in the chemical space largely defined by COCONUT, followed by that of DCM. The density distribution of the compounds appeared concentrated between COCONUT and FooDB, in association with the large (98%) overlap between FooDB and COCONUT compounds (vide supra, Figure 1); a lower density was evidenced for DCM, CAS, and 3CLP. Figure 9 shows that the chemical space of the fragments was mostly defined by COCONUT fragments. Nevertheless, FooDB fragments presented a lower density compared to FooDB compounds, whereas a higher density was found for DCM fragments and CAS fragments concentrating in the chemical space covered by COCONUT fragments.

On the other hand, small molecules with scarce biological activity, like DCM, still converged in a large portion of chemical space covered by NPs (COCONUT) and CAS datasets. To further illustrate this point, Figures 10 and 11 show a direct comparison of DCM, CAS, and the overlapping compounds and fragments. DCM compounds and CAS compounds hardly converged on chemical space, while CAS fragments and DCM fragments appeared to cover a large area of chemical space. For this reason, DCM fragments showed a significant larger overlap with CAS fragments in comparison with the original compounds. This observation suggests that fragments generated from DCM can be used as building blocks in de novo design of bioactive molecules, despite the source compounds' lack of biological activity.

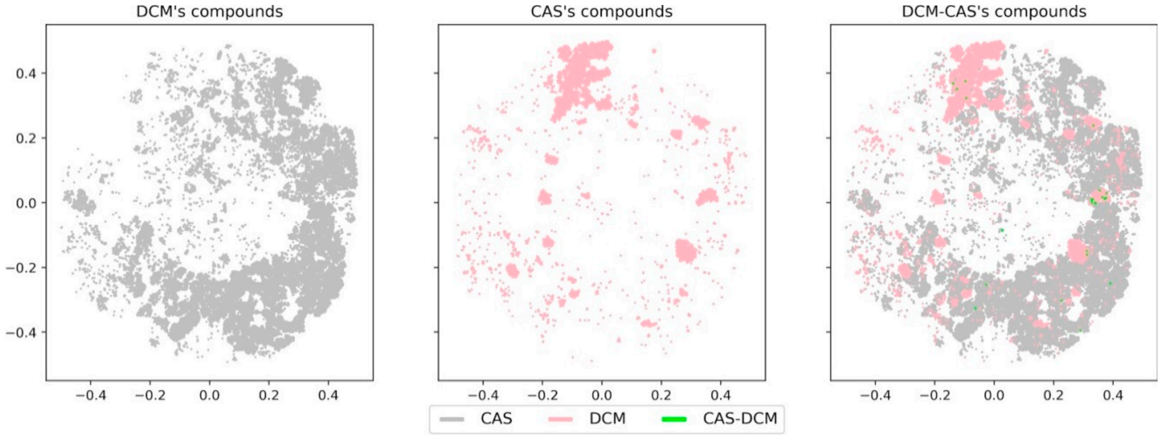


Figure 10. Visualization of the chemical space from CAS compounds (pink), DCM compounds (gray), and overlapping DCM-CAS compounds (green).

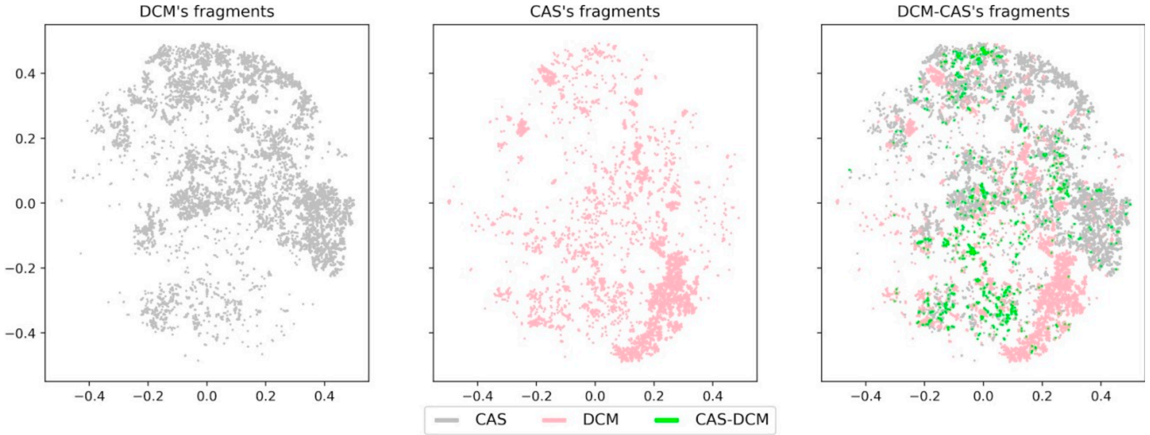


Figure 11. Visualization of the chemical space from CAS fragments (pink), DCM fragments (gray), and overlapping DCM-CAS fragments (green).

4. Conclusions

Herein, we generated, analyzed the composition, and made publicly available a fragment library obtained from an extensive collection of NP. The source compounds and fragment libraries were compared to herein assembled fragment libraries of compounds of interest in drug discovery, including molecules with significance in COVID-19 research. It was concluded that, in general, the fragments generated retained the structural characteristics of the source compounds (COCONUT, FooDB, CAS, DCM, and 3CLP). This analysis found that compounds from NP and food chemicals were structurally more diverse and complex than compounds from CAS, DCM, and 3CLP. Fragments generated from COCONUT and FooDB were more diverse than those from DCM and 3CLP and less diverse than those of the CAS fragments. It was also concluded that fragments from DCM overlapped with bioactive compounds like those of the CAS subset studied in this work. This reinforces previous observations of DCM as a source of building blocks for designing bioactive molecules.

Similarly, fragments of NP from COCONUT and FooDB appear to be important and valuable building blocks for the future de novo design of bioactive compounds. The fragment libraries of the reference databases generated in this work and focused on COVID-19 research (CAS and 3CLP) can be used to identify novel compounds of medical interest and are not currently available in commercial libraries. The fragment libraries for COCONUT and FooDB and the reference libraries DCM, CAS, and 3CLP that we developed in this work are publicly available at https://doi.org/10.6084/m9.figshare.13064231.v1.

Author Contributions: Designed and supervised the project, J.L.M.-F.; wrote the paper, A.L.C.-H., J.L.M.-F.; methodology development, A.L.C.-H., J.L.M.-F., N.S.-C.; data curation and fragments generation, A.L.C.-H., N.S.-C.; data visualization A.L.C.-H., N.S.-C.; formal analysis, A.L.C.-H., J.L.M.-F. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Programa de Apoyo a la Investigación y el Posgrado (PAIP) grant 5000-9163, Facultad de Química, UNAM. We also thank Dirección General de Cómputo y de Tecnologías de Información y Comunicación (DGTIC), UNAM, for the computational resources to use Miztli supercomputer and UNAM, project LANCAD-UNAM-DGTIC-335.

Acknowledgments: A.L.-C.H. and N.S.-C. are thankful to CONACyT for the granted scholarship number 847870 and 335997, respectively.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, and in the decision to publish the results.

References

1. Prieto-Martínez, F.D.; Norinder, U.; Medina-Franco, J.L. Cheminformatics explorations of natural products BT. In *Progress in the Chemistry of Organic Natural Products 110: Cheminformatics in Natural Product Research*; Kinghorn, A.D., Falk, H., Gibbons, S., Kobayashi, J., Asakawa, Y., Liu, J.-K., Eds.; Springer International Publishing: Cham, Switzerland, 2019; pp. 1–35. ISBN 978-3-030-14632-0.
2. Newman, D.J.; Cragg, G.M. Natural products as sources of new drugs over the nearly four decades from 01/1981 to 09/2019. *J. Nat. Prod.* **2020**, *83*, 770–803. [CrossRef]
3. López-Vallejo, F.; Giulianotti, M.A.; Houghten, R.A.; Medina-Franco, J.L. Expanding the medicinally relevant chemical space with compound libraries. *Drug Discov. Today* **2012**, *17*, 718–726. [CrossRef] [PubMed]
4. Ganesan, A. Natural products as a hunting ground for combinatorial chemistry. *Curr. Opin. Biotechnol.* **2004**, *15*, 584–590. [CrossRef] [PubMed]
5. Christoforow, A.; Wilke, J.; Binici, A.; Pahl, A.; Ostermann, C.; Sievers, S.; Waldmann, H. Design, synthesis, and phenotypic profiling of pyrano-furo-pyridone pseudo natural products. *Angew. Chemie Int. Ed.* **2019**, *58*, 14715–14723. [CrossRef] [PubMed]
6. Medina-Franco, J.L. Chapter 21—Discovery and development of lead compounds from natural sources using computational approaches. In *Evidence-Based Validation of Herbal Medicine*; Mukherjee, P.K., Harwansh, R.K., Bahadur, S., Banerjee, S., Kar, A., Eds.; Elsevier: Boston, MA, USA, 2015; pp. 455–475. ISBN 978-0-12-800874-4.
7. 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. [CrossRef] [PubMed]
8. Chen, Y.; Kirchmair, J. Cheminformatics in natural product-based drug discovery. *Mol. Inf.* **2020**. [CrossRef]
9. Medina-Franco, J.L. Towards a unified Latin American natural products database: LANaPD. *Futur. Sci. OA* **2020**, *6*, FSO468. [CrossRef]
10. Chávez-Hernández, A.L.; Sánchez-Cruz, N.; Medina-Franco, J.L. A fragment library of natural products and its comparative chemoinformatic characterization. *Mol. Inf.* **2020**. [CrossRef]
11. Santini, A.; Cicero, N. Development of food chemistry, natural products, and nutrition research: Targeting new frontiers. *Foods* **2020**, *9*, 482. [CrossRef]
12. Martinez-Mayorga, K.; Medina-Franco, J.L. *Foodinformatics: Applications of Chemical Information to Food Chemistry*; Springer: Berlin/Heidelberg, Germany, 2014; ISBN 3319102265.
13. Wassermann, A.M.; Lounkine, E.; Hoepfner, D.; Le Goff, G.; King, F.J.; Studer, C.; Peltier, J.M.; Grippo, M.L.; Prindle, V.; Tao, J.; et al. Dark chemical matter as a promising starting point for drug lead discovery. *Nat. Chem. Biol.* **2015**, *11*, 958–966. [CrossRef]

14. Santibáñez-Morán, M.G.; López-López, E.; Prieto-Martínez, F.D.; Sánchez-Cruz, N.; Medina-Franco, J.L. Consensus virtual screening of dark chemical matter and food chemicals uncover potential inhibitors of SARS-CoV-2 main protease. *RSC Adv.* **2020**, *10*, 25089–25099. [CrossRef]
15. Tang, B.; He, F.; Liu, D.; Fang, M.; Wu, Z.; Xu, D. AI-aided design of novel targeted covalent inhibitors against SARS-CoV-2. *bioRxiv* **2020**. [CrossRef]
16. Sorokina, M.; Steinbeck, C. Review on natural products databases: Where to find data in 2020. *J. Cheminform.* **2020**, *12*, 20. [CrossRef]
17. The Metabolomics Innovation Centre. The Metabolomics Innovation Centre: FooDB (Version 1). Available online: https://foodb.ca/ (accessed on 19 May 2020).
18. American Chemical Society: CAS COVID-19 Antiviral Candidate Compounds Dataset. Available online: https://www.cas.org/covid-19-antiviral-compounds-dataset (accessed on 19 May 2020).
19. Toolkit RDKit. Available online: http://rdkit.org (accessed on 21 May 2020).
20. MolVS. Available online: https://molvs.readthedocs.io/en/latest/ (accessed on 21 May 2020).
21. Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31–36. [CrossRef]
22. Lewell, X.Q.; Judd, D.B.; Watson, S.P.; Hann, M.M. RECAPRetrosynthetic combinatorial analysis procedure: A powerful new technique for identifying privileged molecular fragments with useful applications in combinatorial chemistry. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 511–522. [CrossRef]
23. Rogers, D.; Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742–754. [CrossRef]
24. Durant, J.L.; Leland, B.A.; Henry, D.R.; Nourse, J.G. Reoptimization of MDL Keys for use in drug discovery. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1273–1280. [CrossRef]
25. Agrafiotis, D.K. A constant time algorithm for estimating the diversity of large chemical libraries. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 159–167. [CrossRef] [PubMed]
26. Probst, D.; Reymond, J.-L. Visualization of very large high-dimensional data sets as minimum spanning trees. *J. Cheminform.* **2020**, *12*, 12. [CrossRef]
27. TMAP. Available online: https://tmap.gdb.tools/ (accessed on 18 August 2020).
28. Sánchez-Cruz, N.; Pilón-Jiménez, B.A.; Medina-Franco, J.L. Functional group and diversity analysis of BIOFACQUIM: A Mexican natural product database. *F1000Research* **2020**, *8*. [CrossRef]
29. Sayed, A.M.; Khattab, A.R.; AboulMagd, A.M.; Hassan, H.M.; Rateb, M.E.; Zaid, H.; Abdelmohsen, U.R. Nature as a treasure trove of potential anti-SARS-CoV drug leads: A structural/mechanistic rationale. *RSC Adv.* **2020**, *10*, 19790–19802. [CrossRef]
30. Gentile, D.; Patamia, V.; Scala, A.; Sciortino, M.T.; Piperno, A.; Rescifina, A. Putative inhibitors of SARS-CoV-2 main protease from a library of marine natural products: A virtual screening and molecular modeling study. *Mar. Drugs* **2020**, *18*, 225. [CrossRef]
31. Chen, Y.; de Lomana, G.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]
32. Feher, M.; Schmidt, J.M. Property distributions: Differences between drugs, natural products, and molecules from combinatorial chemistry. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 218–227. [CrossRef] [PubMed]
33. Cremosnik, G.S.; Liu, J.; Waldmann, H. Guided by evolution: From biology oriented synthesis to pseudo natural products. *Nat. Prod. Rep.* **2020**. [CrossRef]

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

biomolecules

MDPI

Article

Towards the De Novo Design of HIV-1 Protease Inhibitors Based on Natural Products

Ana L. Chávez-Hernández, K. Eurídice Juárez-Mercado , Fernanda I. Saldívar-González and José L. Medina-Franco *

DIFACQUIM Research Group, Department of Pharmacy, School of Chemistry, Universidad Nacional Autónoma de México, Avenida Universidad 3000, Mexico City 04510, Mexico; anachavez3026@gmail.com (A.L.C.-H.); kaeuridice@gmail.com (K.E.J.-M.); fer.saldivarg@gmail.com (F.I.S.-G.)
* Correspondence: medinajl@unam.mx; Tel.: +52-55-5622-3899

Abstract: Acquired immunodeficiency syndrome (AIDS) caused by the human immunodeficiency virus (HIV) continues to be a public health problem. In 2020, 680,000 people died from HIV-related causes, and 1.5 million people were infected. Antiretrovirals are a way to control HIV infection but not to cure AIDS. As such, effective treatment must be developed to control AIDS. Developing a drug is not an easy task, and there is an enormous amount of work and economic resources invested. For this reason, it is highly convenient to employ computer-aided drug design methods, which can help generate and identify novel molecules. Using the de novo design, novel molecules can be developed using fragments as building blocks. In this work, we develop a virtual focused compound library of HIV-1 viral protease inhibitors from natural product fragments. Natural products are characterized by a large diversity of functional groups, many sp^3 atoms, and chiral centers. Pseudo-natural products are a combination of natural products fragments that keep the desired structural characteristics from different natural products. An interactive version of chemical space visualization of virtual compounds focused on HIV-1 viral protease inhibitors from natural product fragments is freely available in the supplementary material.

Keywords: artificial intelligence; de novo design; fragment-based drug discovery; HIV-1 inhibitors; pseudo natural products

check for updates

Citation: Chávez-Hernández, A.L.; Juárez-Mercado, K.E.; Saldívar-González, F.I.; Medina-Franco, J.L. Towards the De Novo Design of HIV-1 Protease Inhibitors Based on Natural Products. *Biomolecules* **2021**, *11*, 1805. https://doi.org/10.3390/biom11121805

Academic Editors: Antonio Lavecchia, Carmen Cerchia and Dongqing Wei

Received: 5 November 2021
Accepted: 29 November 2021
Published: 1 December 2021

1. Introduction

The acquired immunodeficiency syndrome (AIDS) caused by the human immunodeficiency virus (HIV) is a major global public health concern. In 2020, the World Health Organization (WHO) reported that approximately 37.7 million people live with HIV out of 24.5 million from the African region. In 2020, 680,000 people died from HIV-related causes and 1.5 million people acquired it [1]. There is no definite treatment for AIDS. Therefore, it is necessary to collaborate to develop a treatment since the antiretroviral drugs currently approved by Food and Drug Administration (FDA) to clinical use only control AIDS and prevent HIV-1 transmission between individuals (Figure 1 and Table 1) [2–4].

Drug design and development demand many years of hard work and economic investment. Most drug candidates are prone to fail [5]. From 25,000 compounds that start in the laboratory, only 25 make it through preclinical testing to human testing, and just five of those reach the actual clinical use [6]. Computer-aided drug design (CADD) has contributed to yielding several drugs into the clinic, yet it has several challenges ahead [7]. Among the CADD methods, de novo design has gained relevance due to the diversity of structures generated by optimizing the algorithms used. From a methodological point of view, artificial intelligence as boosted the development and application of de novo design [5,8,9]. Notably, de novo design is a structure-based drug design method that benefits from the experimental information available of the binding sites of molecular targets.

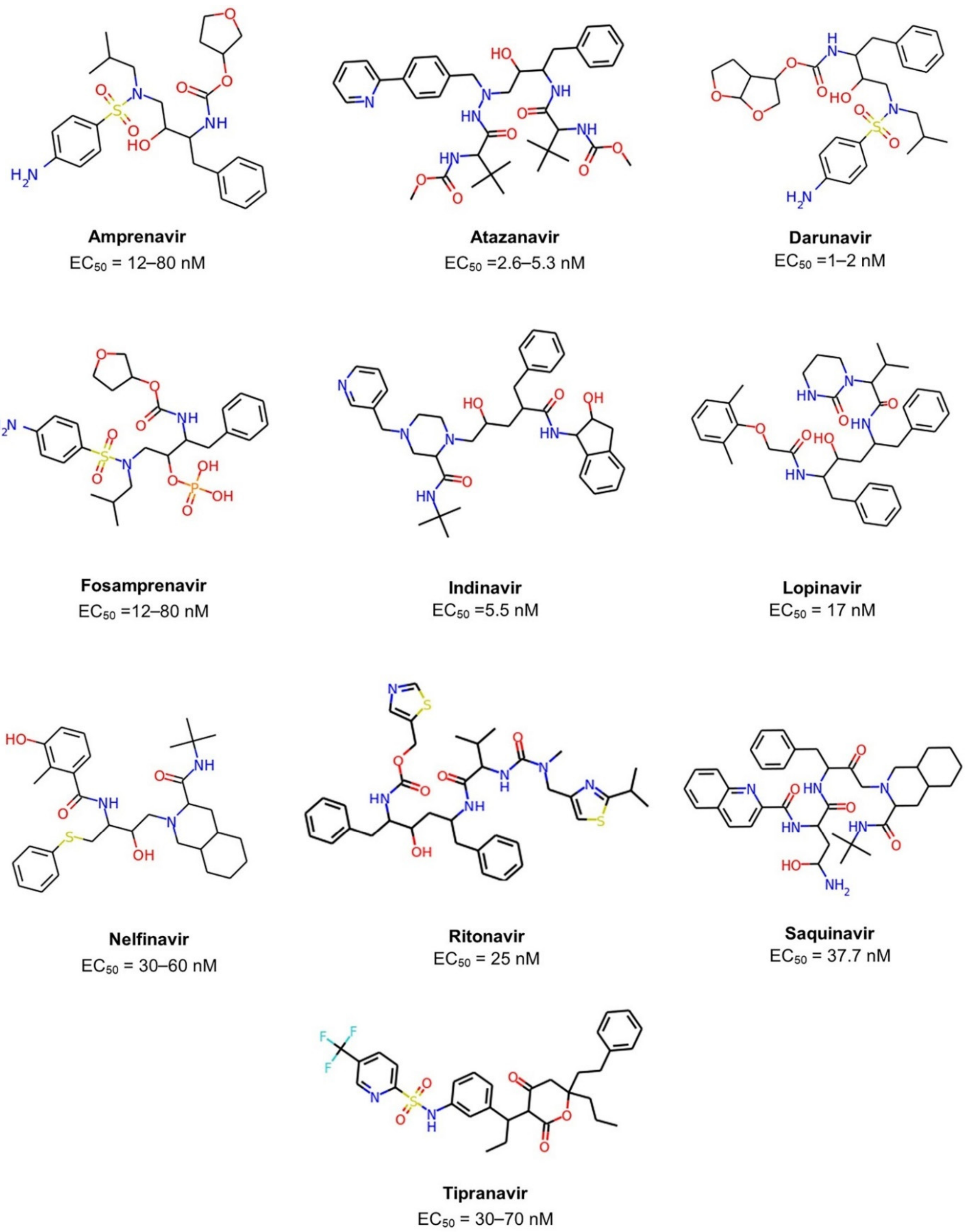


Figure 1. Chemical structures of ten FDA-approved HIV-1 protease inhibitors (Amprenavir, Atazanavir, Darunavir, Fosamprenavir, Indinavir, Lopinavir, Nelfinavir, Ritonavir, Saquinavir, Tipranavir). The EC50 is the concentration of drug required to produce 50% of the maximum possible effect.

Table 1. FDA-approved HIV-1 protease inhibitors which will be used as a reference for the de novo design of the new chemical compounds. [a] Fosamprenavir is the phosphate ester prodrug of amprenavir.

Generic Name	**Brand Name**	**EC_{50} [3]**	**FDA Approval**
Amprenavir	Agenerase	12–80 nM	1999
Atazanavir	Reyataz	2.6–5.3 nM	2003
Darunavir	Prezista	1–2 nM	2006
Fosamprenavir [a]	Lexiva	12–80 nM	2003
Indinavir	Crixivan	5.5 nM	1996
Lopinavir	Kaletra	17 nM	2000
Nelfinavir	Viracept	30–60 nM	1997
Ritonavir	Norvir	25 nM	1996
Saquinavir	Invirase	37.7 nM	1995
Tipranavir	Aptivus	30–70 nM	2005

The main goal of de novo design is to suggest novel molecular structures from scratch with desired activity on a pharmacological target and desired properties [10]. The new structures can be made using two general approaches: fragment-based and atom-based. The advantage of the fragment-based approach is that it narrows down the search in chemical space and maintains good chemical structure diversity [11–13]. Additionally, fragments form fewer interactions that should be able to bind to a greater number of sites on a greater number of proteins. Fragments are small (less than 20 heavy atoms) and typically soluble; they are likely to have better pharmaceutical properties as well as the new chemical compounds generated from them [14]. Over the last 20 years, four drugs from fragment-based drug discovery (FBDD) have been approved, and 40 compounds are currently in clinical trials [15].

Recently, de novo design and artificial intelligence have been combined to propose novel molecules for the treatment of SARS-CoV-2 based on HIV-1 protease and the approved drugs that inhibit this viral protease [8]. Another successful example of de novo design focusing on HIV research led to four molecules from a new compound library generated from the ZINC database [16]. Other approaches de novo design was based on enumerating libraries using chemical reactions [17,18] and are also promising to expand the epigenetic relevant chemical space [19].

The development of new chemical compounds using de novo design can begin from natural product-derived fragments. Natural products have been attractive chemical compounds because they are characterized by a larger number of sp^3 carbon atoms, chiral centers (associated with structural complexity), the larger scaffold diversity, and functional groups, hence their relevance for use as building-blocks [20,21]. Indeed, larger structural complexity of small organic molecules has been associated with increased selectivity and drug-likeness. In previous studies, we showed that natural products cover regions of chemical space that have not yet been explored by synthetically accessible compounds and those with biological activity [22]. For this reason, natural products could be used as building-blocks to develop novel synthetic molecules or pseudo-natural products which combine the desired structural characteristics from different natural products [23].

The goal of this work was to develop a virtual focused compound library of HIV-1 protease inhibitors from natural products fragments through de novo design. The focused library was compared with two virtual libraries of HIV-1 protease inhibitors developed from commercially available fragment libraries that were used as reference. The commercial reference libraries were 4063 ChemDiv's fragments (enriched with sp^3 carbons) [24], and 4150 natural product fragments from Enamine [25]. The natural product fragments were built from the COlleCtion of Open NatUral producTs (COCONUT), the currently largest accessible database of natural products with more than 400,000 non-redundant compounds [26]. Of note, the novel chemoinformatics protocol presented herein is general and can be adapted to generate the compound libraries using de novo design, different molecular templates and molecular targets. Herein we focus on HIV-1 protease because of its current relevance in public health. Thus, we aim that the present work will contribute towards the research that leads to effective HIV treatments.

2. Materials and Methods

The virtual focused compound libraries of HIV-1 viral protease inhibitors from natural product fragments and two commercially available fragments libraries were developed using the protocol outlined in Figure 2.

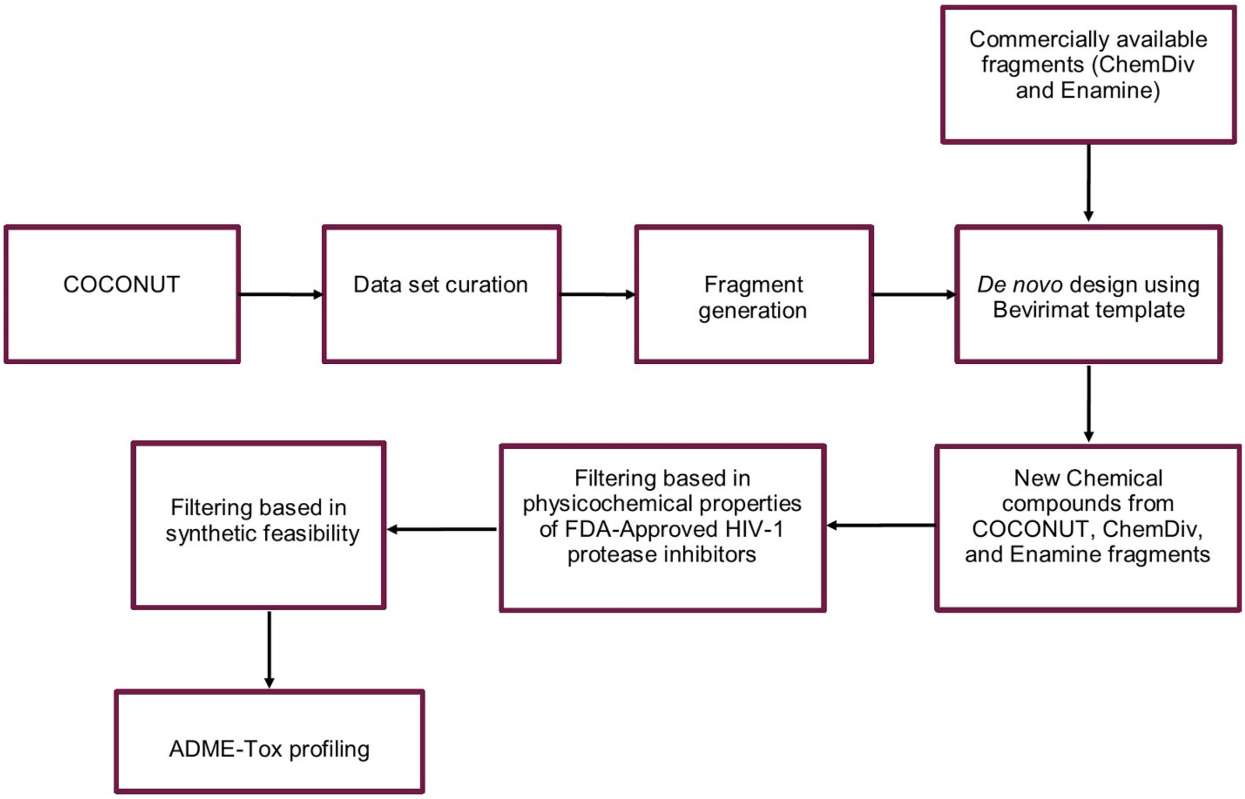

Commercially available fragments (ChemDiv and Enamine)
COCONUT
Data set curation
Fragment generation
De novo design using Bevirimat template
New Chemical compounds from COCONUT, ChemDiv, and Enamine fragments
Filtering based in physicochemical properties of FDA-Approved HIV-1 protease inhibitors
Filtering based in synthetic feasibility
ADME-Tox profiling

Figure 2. De novo design of the virtual focused compound libraries of HIV-1 viral protease inhibitors from natural product fragments (COCONUT) and commercially available fragments (ChemDiv and Enamine).

2.1. Dataset Curation

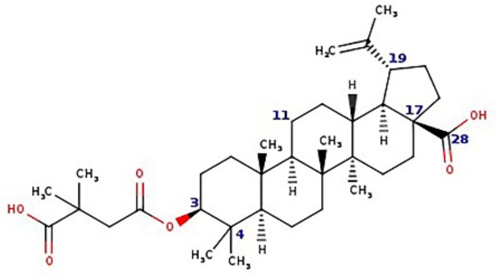
The preparation of compounds, encoded in Simplified Molecular Input Line System (SMILES) [27], was performed using the open-source cheminformatics toolkit RDKit version 2021.03.3 [28], tool MolVS version 0.1.1 [29], and python programming language, version 3.7.10. Compounds with valence errors or any chemical element other than H, B, C, N, O, F, Si, P, S, Cl, Se, Br, and I were deleted. Stereochemistry information was removed because not all compounds in datasets have it defined. Compounds with multiple components were split, and the largest component was retained. The remaining compounds were neutralized and reionized to subsequently generate a canonical tautomer. Repeated compounds were deleted. To narrow down the search chemical space, physicochemical properties were computed: hydrogen bond donors (HBD), hydrogen bond acceptors (HBA), topological polar surface area (TPSA), number of rotatable bonds (RB), molecular weight (MW), and partition coefficient octanol/water (SlogP). Molecular compounds with the "rule of five" [30] and Veber [31] (MW $\leq$ 500, HBD $\leq$ 5, HBA $\leq$ 10, SlogP $\leq$ 5, TPSA $\leq$ 140, RB $\leq$ 10) were retained. Of note, despite the fact some of the fragments used in this work are generated from natural products (as illustrated in Figure 2), the type of molecules designed are small organic drug-like molecules.

2.2. Generation of Unique Fragments Using Retrosynthetic Rules

Fragment libraries were produced with the Retrosynthetic Combinatorial Analysis Procedure (RECAP) as implemented in RDKit. The RECAP algorithm [32] cleaves a molecule into fragments if this had any of the following bonds: amide, ester, amine, urea, ether, olefin, quaternary nitrogen, aromatic nitrogen–aliphatic carbon, lactam nitrogen–aliphatic carbon, aromatics carbon–aromatic carbon, and sulphonamide.

2.3. De Novo Design

The new chemical structures were built based on the template previously proposed by Zhao et al. developed from the structure-activity relationship (SAR) analysis for the optimization of bevirimat (Figure 3), a compound derived from betulinic acid (Figure 4) [33]. Bevirimat [34,35] is a compound in clinical trials that targets the Gag polyprotein inhibiting the action of HIV protease at its the last cleavage event of the capsid protein and spacer peptide 1 (CA-SP1) [36,37]. The template proposed for building new chemical compounds related to bevirimat is shown in Figure 5.



Bevirimat

Figure 3. Chemical structure of bevirimat.

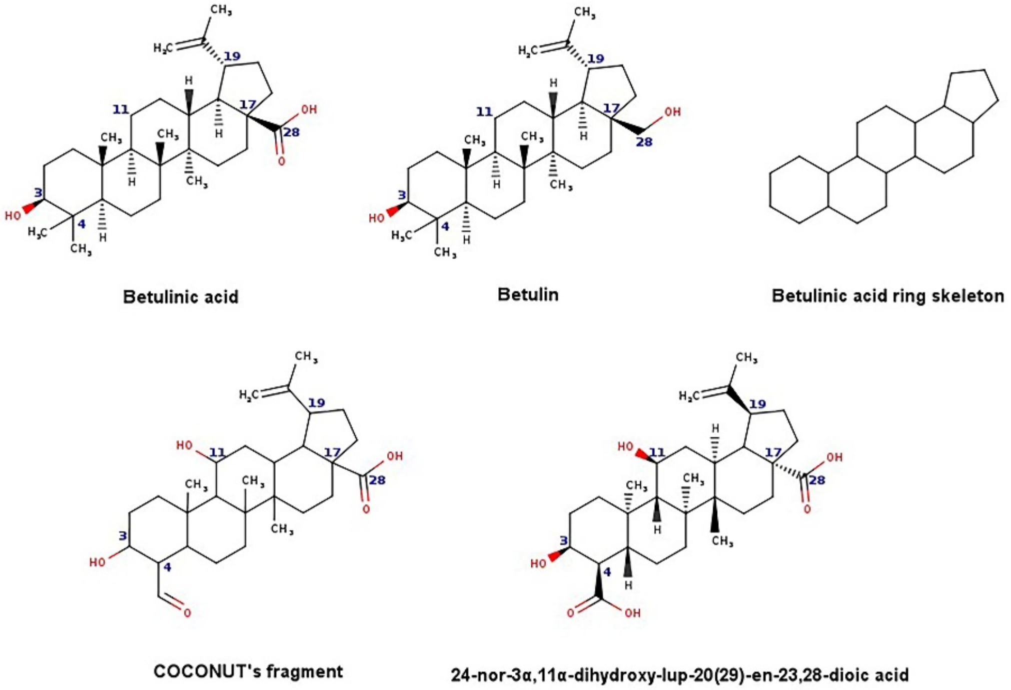

Betulinic acid
Betulin
Betulinic acid ring skeleton
COCONUT's fragment
24-nor-3α,11α-dihydroxy-lup-20(29)-en-23,28-dioic acid

Figure 4. Chemical structures of betulinic acid, betulin, cyclic system skeleton derived from betulinic acid, COCONUT's fragment with betulinic acid ring skeleton derived from the 24-nor-3α,11α-dihydroxy-lup-20(29)-en-23,28-dioic acid.

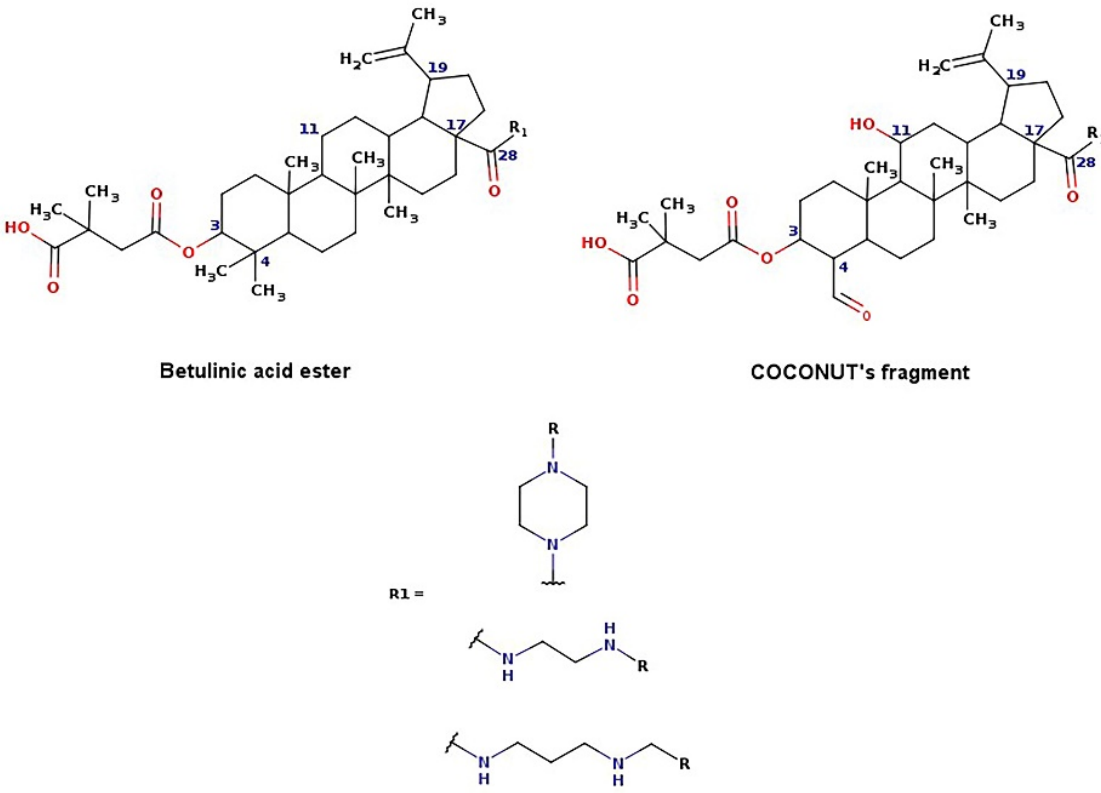


Figure 5. Template for building new chemical compounds similar to bevirimat using the ester of betulinic acid to ChemDiv fragments and Enamine fragments, and the ester of COCONUT's fragment derived from 24-nor-3α,11α-dihydroxy-lup-20(29)-en-23,28-dioic acid.

New molecules were generated using the Python programming language and the toolkit RDKit [28], following the protocol described for Saldívar-González et al. to enumerate chemical libraries [18]. We used COCONUT fragments with a cyclic system skeleton similar to betulinic acid, a hydroxyl group attached to carbon 3, and a carboxylic acid group attached to carbon 17, as shown in Figure 4. The COCONUT's fragment selected was derived from 24-nor-3α,11α-dihydroxy-lup-20(29)-en-23,28-dioic acid (COCONUT ID: CNP0243494 or Reaxys ID: 6547020). Betulinic acid was used to build new chemical compounds from ChemDiv fragments and Enamine fragments because there were no fragments of cyclic system skeleton derived from betulinic acid or analogous triterpenes.

Chemical reactions were represented in SMIRKS, a hybrid notation of SMILES and SMARTS (SMILES Arbitrary Target Specification). Reaction 1, esterification, was made between triterpene alcohol and 2,2-dimethyl succinic acid using SMIRKS 1, as shown in Table 2. Reaction 2, amidation, was built from the carboxyl group attached to carbon 17 as shown in Figure 4 using fragments attached to piperazine, 1,3-diaminoethane, and 1,3-diaminopropane find in COCONUT fragments, ChemDiv fragments, and Enamine fragments. The SMIRKS 2.1–2.3 were used in reaction 2 and shown in Table 2. The compounds and fragments were selected using the functional groups in SMARTS notation described in Table 3. Newly generated chemical structures with valence errors were removed. Canonical SMILES were generated, and duplicate molecules were deleted.

Table 2. SMIRKS used for building the new chemical compounds from natural products fragments.

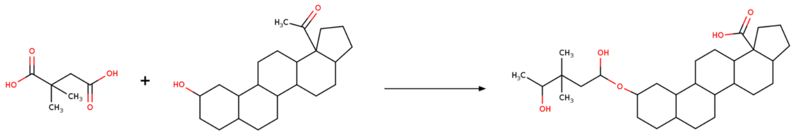
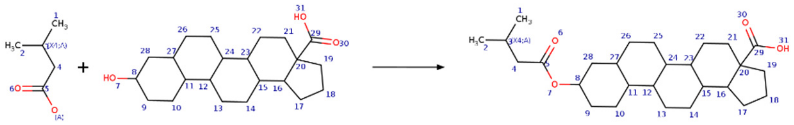
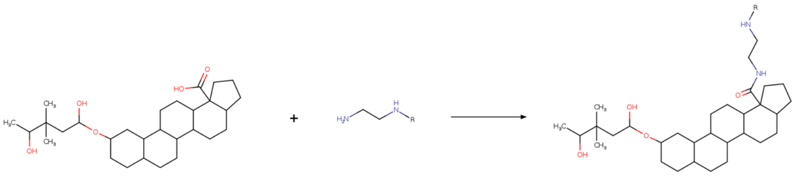
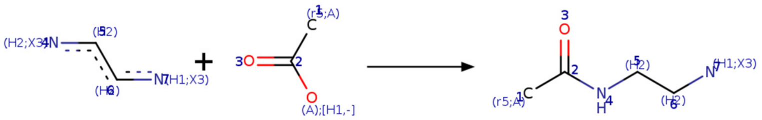
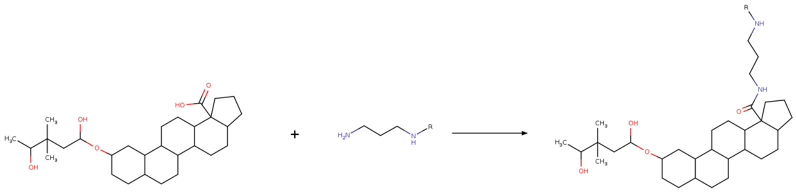
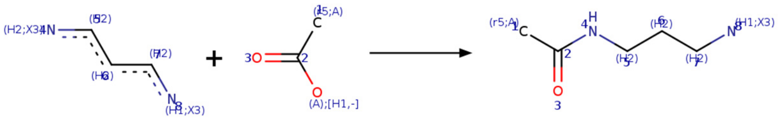
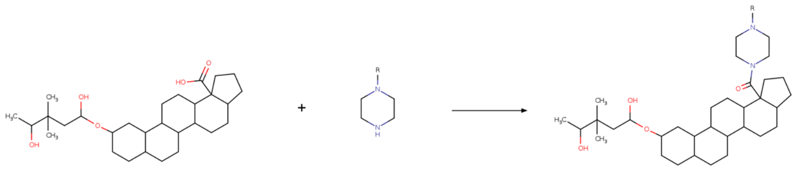
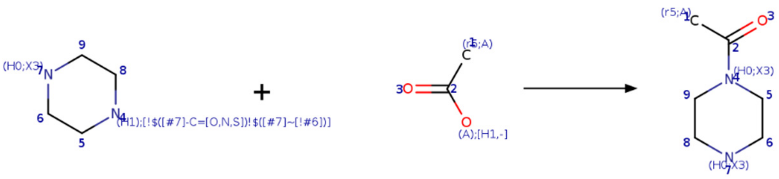
Description	Scheme
Reaction 1	
SMIRKS 1	 [#6:1][#6;A;X4:3]([#6:2])[#6:4]-[#6:5]([#8;A])=[O:6].[#8:7]-[#6:8]-1-[#6:9]-[#6:10]-[#6:11]-2-[#6:27](-[#6:26]-[#6:25]-[#6:24]-3-[#6:23]-4-[#6:22]-[#6:21][C:20]5([#6:19]-[#6:18]-[#6:17]-[#6:16]5-[#6:15]-4-[#6:14]-[#6:13]-[#6:12]-2-3)[#6:29](-[#8:31])=[O:30])-[#6:28]-1>>[#6:2][#6;A;X4:3]([#6:1])[#6:4]-[#6:5](=[O:6])-[#8:7]-[#6:8]-1-[#6:9]-[#6:10]-[#6:11]-2-[#6:27](-[#6:26]-[#6:25]-[#6:24]-3-[#6:23]-4-[#6:22]-[#6:21][C:20]5([#6:19]-[#6:18]-[#6:17]-[#6:16]5-[#6:15]-4-[#6:14]-[#6:13]-[#6:12]-2-3)[#6:29](-[#8:31])=[O:30])-[#6:28]-1
Reaction 2.1	
SMIRKS 2.1	 [#7;H1;X3:7][#6H2:6][#6;H2:5][#7;H2;X3:4].[#6;A;r5:1][#6:2]([#8;A;H1,-])=[O:3]>>[#6;A;r5:1][#6:2](=[O:3])-[#7:4]-[#6;H2:5]-[#6;H2:6]-[#7;H1;X3:7]
Reaction 2.2	
SMIRKS 2.2	 [#7;H1X3:8][#6H2:7][#6H2:6][#6H2:5][#7;H2X3:4].[#6;A;r5:1][#6:2]([#8;A;H1,-])=[O:3]>>[#6;A;r5:1][#6:2](=[O:3])-[#7:4]-[#6H2:5]-[#6H2:6]-[#6H2:7]-[#7;H1X3:8]
Reaction 2.3	
SMIRKS 2.3	 [#6:9]-1-[#6:8]-[#7H1;!$([#7]-C=[O,N,S])!$([#7]~[!#6]):4]-[#6:5]-[#6:6]-[#7;H0X3:7]-1.[#6;A;r5:1][#6:2]([#8;A;H1,-])=[O:3]>>[#6;A;r5:1][#6:2](=[O:3])-[#7;H0X3:4]-1-[#6:5]-[#6:6]-[#7;H0X3:7]-[#6:8]-[#6:9]-1

Table 3. Functional groups using SMARTS notation to filter fragments from natural products.

Functional Groups	SMARTS
Aliphatic alcohol (cyclohexanol)	[#8;H1]-[#6]-1-[#6]-[#6]-[#6]-2-[#6](-[#6]-[#6]-[#6]-3-[#6]-4-[#6]-[#6]C5([#6]-[#6]-[#6]-[#6]5-[#6]-4-[#6]-[#6]-[#6]-2-3)[#6]([#8;H1])=O)-[#6]-1
2,2-dimethyl succinic acid	[#6]C([#6])([#6]-[#6](-[#8])=O)[#6](-[#8])=O
piperazine	[#6;H2;X4]1-[#6;H2;X4][#7;X3;!H1][#6;H2;X4]-[#6;H2;X4][#7;H1;X3]1
1,2-diaminoethane	[#7;H1;X3][#6;H2;X4][#6;H2;X4][#7;H2;X3]
1,3-diaminopropane	[#7;H1;X3][#6;H2;X4][#6;H2;X4][#6;H2;X4][#7;H2;X3]
Cyclic system skeleton derived from betulinic acid	[#6]1-[#6]-[#6]-[#6]2-[#6](-[#6]-1)-[#6]-[#6]-[#6]1-[#6]-2-[#6]-[#6]-[#6]2-[#6]3-[#6]-[#6]-[#6]-[#6]-3-[#6]-[#6]-[#6]-1-2

2.4. Structural Diversity and Complexity

The structural diversity of the new chemical compounds generated was evaluated to compute the median value of the distribution of the pairwise similarity values generated with the Tanimoto coefficient for Morgan fingerprint with radius 2 (Morgan2, 1024-bits) [38] and Molecular ACCes System (MACCS) Keys (166-bits) [39].

2.5. Chemical Space Visualization

The chemical space visualization was done using two methods, principal component analysis (PCA) based on physicochemical properties and the Tree MAP (TMAP) algorithm based on molecular fingerprints [40,41].

PCA is a linear dimensionality reduction technique to transform data with many dimensions into a lower dimensional space and preserve the different relationships between the data points as much as possible [42]. PCA was generated from six physicochemical properties (MW, HB, HBA, SlogP, TPSA, and RB).

TMAP allows the visual representation of many chemical compounds through the distance between the clusters and the cluster's detailed structure through Local Sensitive Hashing (LSH) forest data structure, enabling c-approximate k-nearest neighbors (k-NN). Morgan fingerprints for chemical compounds were encoded using the MinHash algorithm. The number of nearest-neighbors, k = 50, and the factor used by the augmented query algorithm, kc = 10, were used to develop the TMAP graphs. Morgan fingerprints with radius 2 (Morgan2, 1024-bits) were generated to generate TMAP graphs [38]. Applications of TMAP for chemical space visualization of other compound datasets have been reported [43,44].

2.6. Filtering of the New Chemical Compounds Generated

To narrow down the search in chemical space and set the conditions for the newly generated compounds, physicochemical properties were computed for libraries generated and FDA-approved HIV-1 protease inhibitors (Table 1 and Figure 1). The maximum values of the physicochemical properties obtained from the HIV-1 protease inhibitors was HBD ≤ 6, HBA ≤ 13, SlogP ≤ 6.7, MW ≤ 720.30, TPSA ≤ 174.60, and RB ≤ 17 (Table 4). Molecules with at least four rules were retained. SlogP strictly must be complied. These sets of properties and values were used as a heuristic rule that is slightly less stringent than the Lipinski and Veber rules [30,31].

Table 4. Properties of pharmaceutical relevance of FDA-approved HIV-1 protease inhibitors.

Parent Molecule	SlogP	MW	HBD	HBA	TPSA	RB
Amprenavir	2.40	505.22	4	9	131.19	11
Atazanavir	4.21	704.39	5	13	171.22	14
Darunavir	2.38	547.24	4	10	140.42	11
Fosamprenavir [a]	2.69	585.19	4	12	174.56	13
Indinavir	2.87	613.36	4	9	118.03	11
Lopinavir	4.33	628.36	4	9	120.00	15
Nelfinavir	4.75	567.31	4	7	101.90	9
Ritonavir	5.91	720.31	4	11	145.78	17
Saquinavir	3.09	670.38	6	11	166.75	12
Tipranavir	6.70	602.21	1	7	102.43	11
Minimum [a]	2.40	505.20	1	7	101.90	9
Maximum [a]	6.70	720.30	6	13	174.60	17

[a] Maximum and minimum values for each property.

2.7. Synthetic Feasibility

The complexity of the compounds generated was estimated using the synthetic accessibility score (SAscore) previously reported [45]. The SAscore implemented in this work is the difference between fragment score and complexity penalty. The fragment score captures common structural features in a large number of already synthesized molecules (934,046 representative molecules from the PubChem). Molecules are fragmented using extended connectivity fragments (ECFP_4# fragments), and the fragment score is calculated as a sum of contributions of all fragments in the molecule divided by the number of fragments in the molecule. The fragment frequency is related to their synthetic accessibility, and hence easy-to-prepare substructures are present in molecules quite often. The complexity score is calculated as the sum of ring complexity (ring bridge atoms and spiro atoms), the number of stereocenters, large rings (ring size greater than eight, molecular complexity increases), and molecule size. The SAscore was calculated for the virtual focused libraries of HIV-1 viral protease inhibitors generated, and two reference datasets of FDA-approved drugs, and FDA-approved HIV-1 protease inhibitors [46]. The SAscore was calculated using the Python script published by Ertl and Schuffenhauer [45].

2.8. ADME-Tox Profiling

Absorption, distribution, metabolism, excretion, and toxicity (ADME-Tox) properties of virtual focused libraries of HIV-1 viral protease inhibitors generated were calculated using the SwissADME server [47] and the pkCSM-pharmacokinetics server [48]. The ADME-Tox properties of FDA-approved drugs were also computed as reference. The SwissADME server was used to compute descriptors associated with absorption and metabolism. The pkCSM-pharmacokinetics server was used to compute descriptors associated with absorption, distribution, excretion, and toxicity. The evaluation of descriptors related to ADME-Tox properties was computed as previously described [49]. The descriptors calculated were absorption broken down into solubility, Silico-IT LogSw; lipophilicity, consensus LogPo/w, and human intestinal absorption (HIA). The blood-brain barrier (BBB) permeability, P-glycoprotein substrate, P-glycoprotein I inhibitor, and P-glycoprotein II (take binary values: yes/no) for distribution. Inhibition of five main cytochrome enzymes (CYP-1A2, CYP-2C19, CYP-2C9, CYP-2D6, CYP-3A4) for metabolism (take binary values: yes/no). Total clearance log (mL/min/kg) to excretion. The hERG I/II inhibition, AMES toxicity, and hepatotoxicity to toxicity (take binary values: yes/no).

3. Results and Discussion

As mentioned in the Introduction and Methods sections, new chemical compounds were built from two commercially available libraries: 4063 ChemDiv fragments enriched with sp^3 carbons, 4160 Enamine natural products fragments, and 184,769 COCONUT fragments computationally generated in house. The total number of molecules generated were: 1534 from COCONUT's fragments, 62 molecules from ChemDiv fragments, and 11 molecules from Enamine fragments. Fragments attached to 1,3-diaminopropane were not found in ChemDiv and Enamine's fragment collections. Similarly, fragments attached to 1,2-diaminoethane were not found in Enamine fragments.

3.1. Structural Diversity

The median of similarity generated using Morgan2 and MACCS keys fingerprints are shown in brackets, respectively, and described in Table S1 in the supplementary material. FDA-approved drugs (0.096, 0293) and FDA-approved HIV-1 protease inhibitors (0.253, 0.558) were the most diverse datasets, following by compounds derived from COCONUT fragments (0.605, 0.817), ChemDiv fragments (0.676, 0.821), and Enamine fragments (0.682, 0.823). Compounds computationally generated from fragment datasets were less diverse because these datasets are focused on bevirimat-like compounds.

3.2. Chemical Space Visualization

A visual representation of the chemical space based on physicochemical properties (MW, HB, HBA, SlogP, TPSA, and RB, as stated in the Methods Section 2.5) using PCA is shown in Figure 6. Principal component 1 recovered 73.6% of the variance, and principal component 2 recovered 21.2% of the variance. The accumulated variance recovered by the first two principal components represented in Figure 6 was 94.8%. In this chemical space visualization, the compounds generated from the three fragment libraries are within the space of physicochemical properties of FDA-approved drugs. Likewise, some compounds generated from COCONUT fragments had physicochemical properties similar to FDA-approved HIV-1 protease.

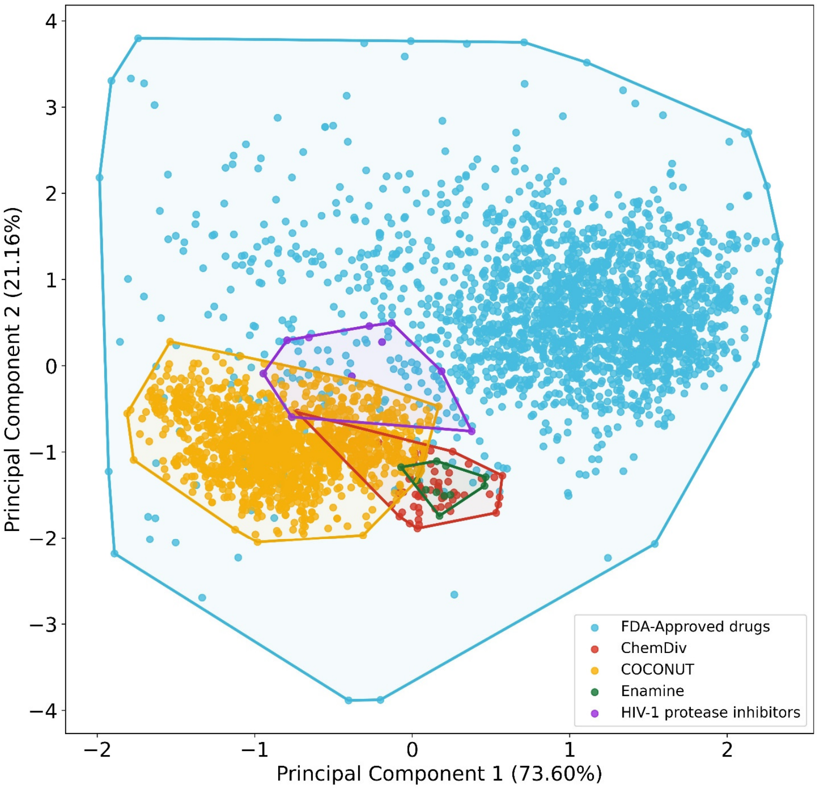


Figure 6. Chemical space visualization of the virtual focused compound library of HIV-1 viral protease inhibitors from natural product fragments and two compound reference libraries using PCA based on physicochemical properties. Compound reference libraries represented in colors: FDA-approved drugs (blue) and FDA-approved HIV-1 protease inhibitors (purple). Likewise. for new chemical compounds generated from COCONUT (orange), ChemDiv (red), and Enamine (green) fragment libraries.

To quantitatively define which dataset is the most diverse, coverage space obtained by convex hull analysis derived from PCA was computed for each dataset (Figure S1). The convex hull is defined as the minimum convex polygon so that the point set is either inside this polygon or at its border [50,51]. The convex hull area computed were for FDA-approved drugs (737.59), HIV-1 protease inhibitors (1.11), compounds from COCONUT's fragments (3.18), compounds from ChemDiv's fragments (0.79), and compounds from Enamine fragments (0.18). The outcome of this analysis was similar to the results of the structural diversity analysis based on fingerprints (Section 3.1): reference datasets were more diverse than the new chemical compounds generated from fragments datasets. The new chemical compounds derived from COCONUT fragments were the most diverse, followed by new chemical compounds derived from ChemDiv and Enamine fragments.

The visual representation of the chemical space based on molecular fingerprint using the TMAP algorithm is shown in Figure 7. An interactive version of the TMAP is available at https://figshare.com/s/ceb58d58e8f5585ce67e (accessed on 5 November 2021). The chemical structures of new chemical compounds generated were very different in comparison with FDA-approved drugs and FDA-HIV-1 protease inhibitors. The chemical structures of the new compounds generated from ChemDiv and Enamine fragments were very similar compared to compounds derived from COCONUT fragments. In some cases, the chemical structures of compounds generated from COCONUT's fragments were very similar to some FDA-approved drugs, for instance, palbociclib and pipecuronium. In these cases where there are not commercially available fragments like COCONUT's fragments could be used palbociclib and pipecuronium.

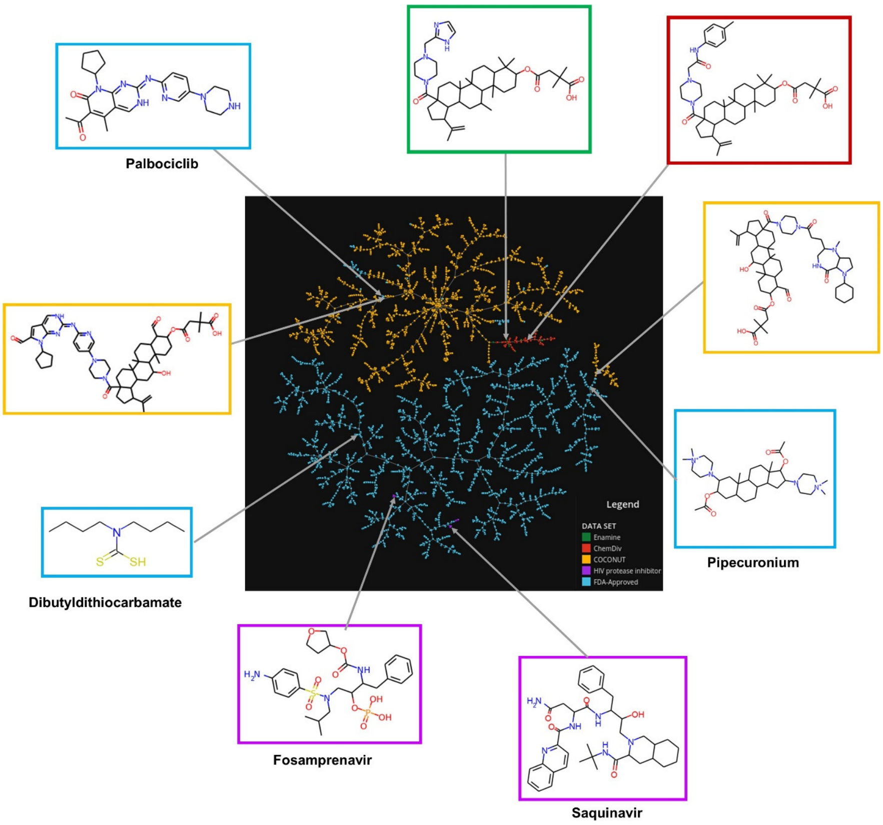


Figure 7. Chemical space visualization of the virtual focused compound library of HIV-1 viral protease inhibitors from natural product fragments and two compound reference libraries using TMAP based on molecular fingerprints. Compounds reference libraries represented in colors: FDA-approved drugs (blue), and FDA-approved HIV-1 protease inhibitors (purple). Likewise, for new chemical compounds generated from COCONUT (orange), ChemDiv (red), and Enamine (green) fragment libraries. The interactive version is available at https://figshare.com/s/ceb58d58e8f5585ce67e (accessed on 5 November 2021).

3.3. Compound Filtering Based on Physicochemical Properties

Figure 8 shows box-whisker plots of physicochemical properties after applying the empirical rules proposed (Section 2.6). The summary of descriptive statistics is shown in Tables S2–S7 in the supplementary material. 352 compounds generated from COCONUT fragments (20%) and 1 compound generated from ChemDiv fragments were retained (2%), and compounds generated from Enamine fragments were not retained (0%). Based on the properties' distribution shown in the box-whisker plots, the physicochemical properties of compounds generated from COCONUT fragments, ChemDiv fragments, and Enamine fragments were different regarding FDA-approved HIV-1 protease inhibitors and FDA-approved drugs.

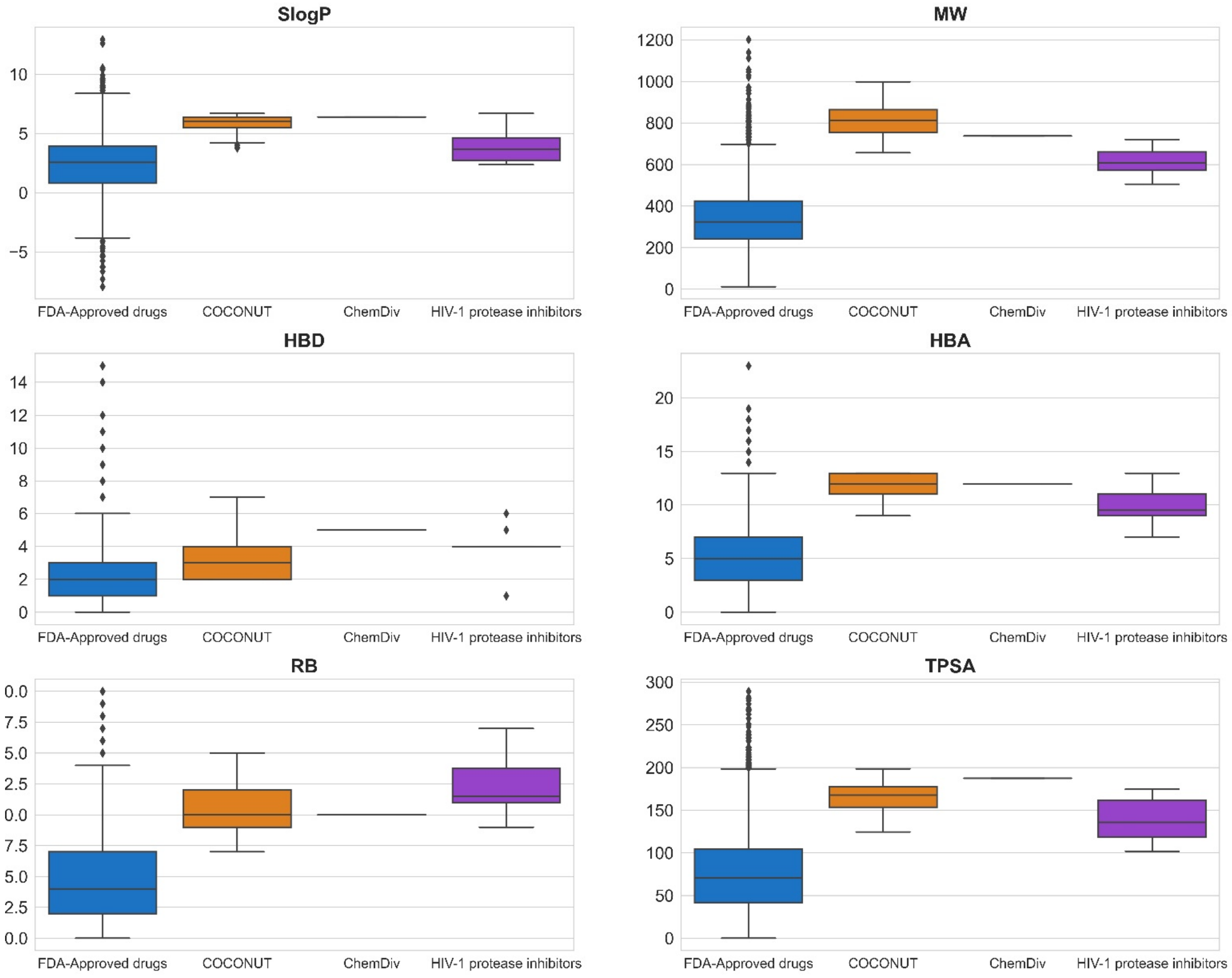


Figure 8. Box-whisker plots of physicochemical properties of FDA-approved drugs (blue), FDA-approved HIV-1 protease inhibitors (purple), and new chemical compounds generated from COCONUT (orange) and ChemDiv (red) fragment libraries after applying physicochemical properties filtering. Black diamonds represent outliers.

The physicochemical properties calculated for datasets were: SlogP $\leq$ 12.94, MW $\leq$ 1201.84, RB $\leq$ 20, TPSA $\leq$ 286.50, HBA $\leq$ 23, HBD $\leq$ 15 for FDA-approved drugs; SlogP $\leq$ 6.70, MW $\leq$ 720.31, RB $\leq$ 17, TPSA $\leq$ 174.56, HBA $\leq$ 13, HBD $\leq$ 6 for FDA-approved HIV-1 protease inhibitors; SlogP $\leq$ 6.69, MW $\leq$ 998.63, RB $\leq$ 15, TPSA $\leq$ 198.54, HBA $\leq$ 13, HBD $\leq$ 7 for compounds generated from COCONUT fragments, and SlogP = 6.4, MW = 737.47, RB = 10, TPSA = 187.47, HBA = 12, HBD = 5 for the compound generated from ChemDiv's fragments. The SlogP, RB, and HBA values of compounds generated from COCONUT fragments and ChemDiv fragments were less than FDA-approved HIV-1 protease inhibitors. HBA values were equal or less than FDA-approved HIV-1 protease inhibitors. The SlogP values of compounds derived from Enamine fragments were larger than FDA-approved HIV-1 protease inhibitors as shown in Figure S2; accordingly, no compound was retained. The MW, TPSA, and HBD values of compounds generated from COCONUT fragments were larger than for FDA-approved HIV-1 protease inhibitors and less than for FDA-approved drugs. As mentioned above Ganesan [52], natural products that violate the Lipinsky rules remain largely compliant in terms of log P and HBD. He considers that "nature has learned to maintain low hydrophobicity and intermolecular H-bond donating potential when it needs to make biologically active compounds with high molecular weight and a large number of rotatable bonds". In drugs, the molecules that exceed HBD 5 or HBA 10 the majority are natural product-related [53].

3.4. Filtering Based on Synthetic Feasibility

The synthetic feasibility was computed for FDA-approved drugs, FDA-approved HIV-1 protease inhibitors, and compounds generated from COCONUT and ChemDiv fragments with physicochemical properties like FDA-approved HIV-1 protease inhibitors. Figure 9 summarizes the results of synthetic feasibility. Molecules with a low SAscore value < 6 are easily synthetically accessible [45]. A total of 97% FDA-approved drugs had SAscore < 6, and FDA-approved HIV-1 protease inhibitors had SAscore $\leq$ 4.24. Similarly, 75% of compounds generated from COCONUT fragments had SAscore $\leq$ 6.03 and the compound generated from ChemDiv had SAscore = 5.54. Although, compounds generated from COCONUT fragments had 5.50 $\leq$ SAscore $\leq$ 6.03, still in recommended range so that can be synthetically accessible; moreover, the high SAscore, in compounds generated regarding FDA-approved HIV-1 protease inhibitors, was influenced by the ten stereocenters of betulinic acid and 24-nor-3α,11α-dihydroxy-lup-20(29)-en-23,28-dioic acid. Considering that these stereocenters do not have to be generated within the organic synthesis, the SAscore value would be lower.

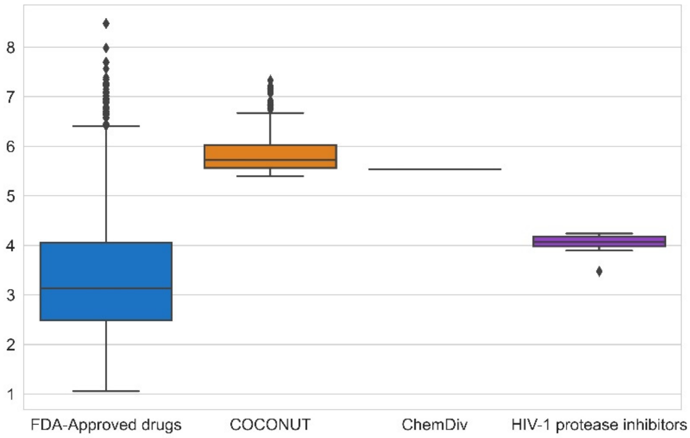


Figure 9. Box-whisker plot of synthetic feasibility calculated for FDA-approved drugs (blue), FDA-approved HIV-1 protease inhibitors (purple), and new chemical compounds generated from COCONUT fragments (orange) and ChemDiv (red) fragments with physicochemical properties like FDA-approved HIV-1 protease inhibitors. Black diamonds represent outliers.

3.5. ADME-Tox Profiling

The ADME-Tox profiling was computed for 251 compounds generated from COCONUT fragments and 1 compound generated from ChemDiv fragments with physicochemical properties like FDA-approved HIV-1 protease inhibitors and estimated as easy synthesizable (i.e., SAscore $\leq$ 6). Similarly, ADME-Tox profiling was computed for FDA-approved drugs and FDA-approved HIV-1 protease inhibitors.

3.5.1. Absorption

Solubility, lipophilicity, and HIA are summarized in Figure 10 and Tables S9–S11 in the supplementary material. Solubility was expressed by Silicos-IT LogSw and lipophilicity was expressed by consensus LogP. Silicos-IT LogSw and consensus LogP were computed with the SwissADME server. Percentage of HIA was computed with the pkCSM-pharmacokinetics server.

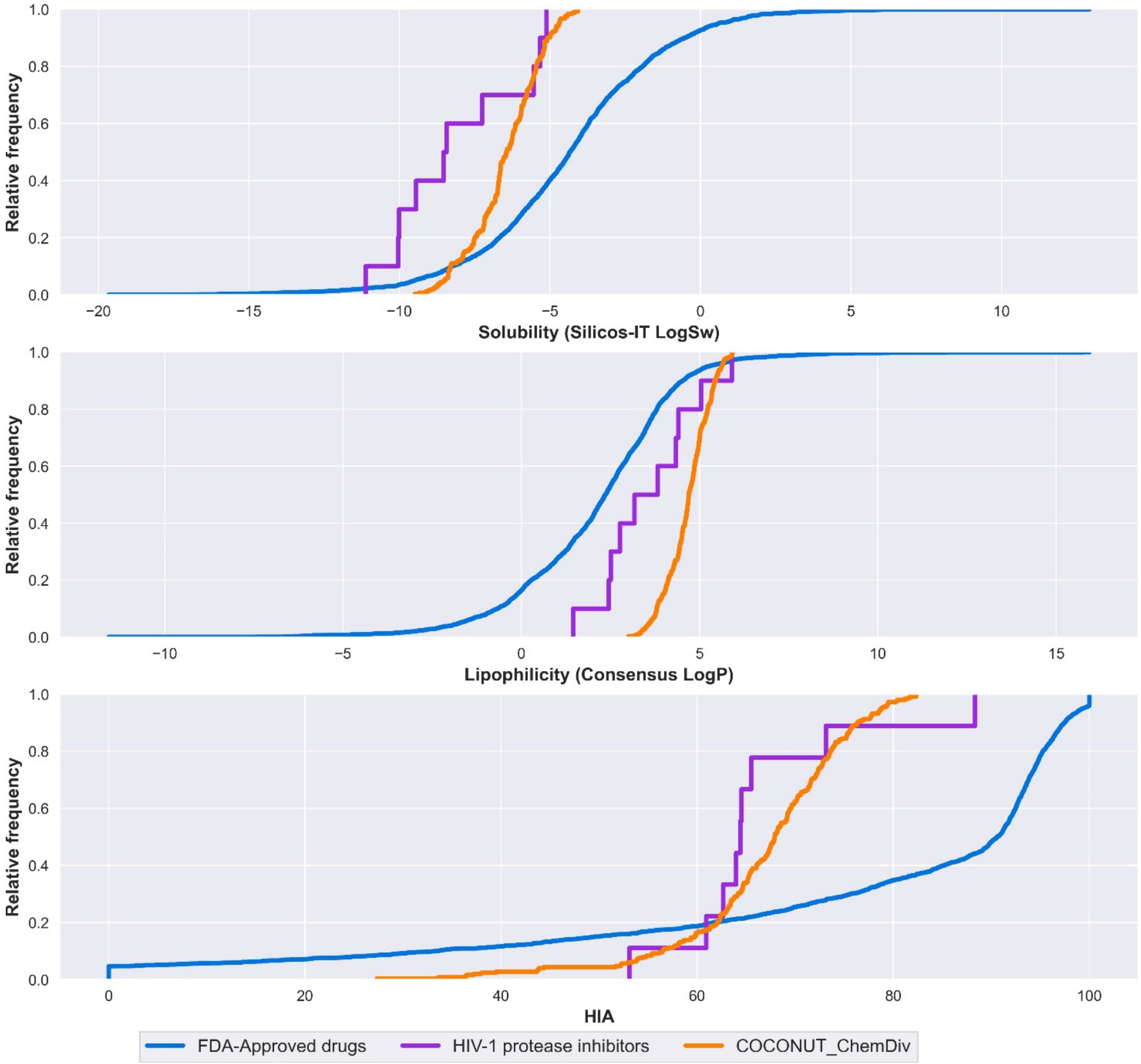


Figure 10. Distribution curve of solubility, lipophilicity, and HIA. Colors represent compounds: new chemical compounds generated from COCONUT fragments and ChemDiv fragments with physicochemical properties like FDA-approved HIV-1 protease inhibitors and easily synthetically accessible (orange), FDA-approved drugs (blue), FDA-approved HIV-1 protease inhibitors (purple). Solubility is expressed in the percentage of Silicos-IT LogSw, and lipophilicity is expressed in the percentage of consensus LogP.

Median values for solubility, lipophilicity, and HIA are described below. FDA-approved drugs had consensus LogP = 2.36, Silicos-IT LogSw = −4.34, HIA = 90.6%. FDA-approved HIV-1 protease inhibitors had consensus LogP = 3.50, Silicos-IT LogSw = −8.49, HIA = 64.4%. Compounds derived from COCONUT and ChemDiv had consensus LogP = 4.70 and Silicos-IT LogSw = −6.45, HIA = 67.9%.

New drug candidates have poor water solubility, and it is often the result of highly lipophilic compounds. Log P < 2, the crystal lattice becomes the main determining factor for solubility. LogP values above 2, the lipophilicity is the main factor [54]. FDA-approved HIV-1 protease inhibitors were highly soluble, followed by compounds derived from COCONUT and ChemDiv fragments, both had Log P > 2; in this case, solubility is strongly influenced by lipophilicity. Contrary to FDA-approved drugs that had Log P close to 2 and were less soluble, solubility mainly depends on the crystal lattice. Compounds derived from COCONUT and ChemDiv fragments had higher HIA in comparison to FDA-approved HIV-1 protease inhibitors.

3.5.2. Distribution

The relative frequency of BBB permeability is described in Figure 11. The median value of BBB permeability was −0.38 for FDA-approved drugs; −1.21 for compounds generated from COCONUT and ChemDiv fragments, and −1.25 for FDA-approved HIV-1 protease inhibitors. Compounds generated from COCONUT and ChemDiv fragments had similar BBB permeability.

The percentage of compounds that are P-glycoprotein substrate, P-glycoprotein I inhibitor, and P-glycoprotein II inhibitor were summarized in Figure 12 and Table S13 in the supplementary material. All FDA-approved HIV-1 protease inhibitors and 96% of compounds generated from COCONUT and ChemDiv fragments were P-glycoprotein substrates. Similarly, 66.67% of HIV-1 Approved protease inhibitors and 82.9% of compounds generated from COCONUT and ChemDiv fragments were P-glycoprotein II inhibitors. Whereas no compounds generated from COCONUT and ChemDiv fragments were P-glycoprotein I inhibitors, against 100% FDA-approved HIV-1 proteases inhibitors were P-glycoprotein I inhibitors.

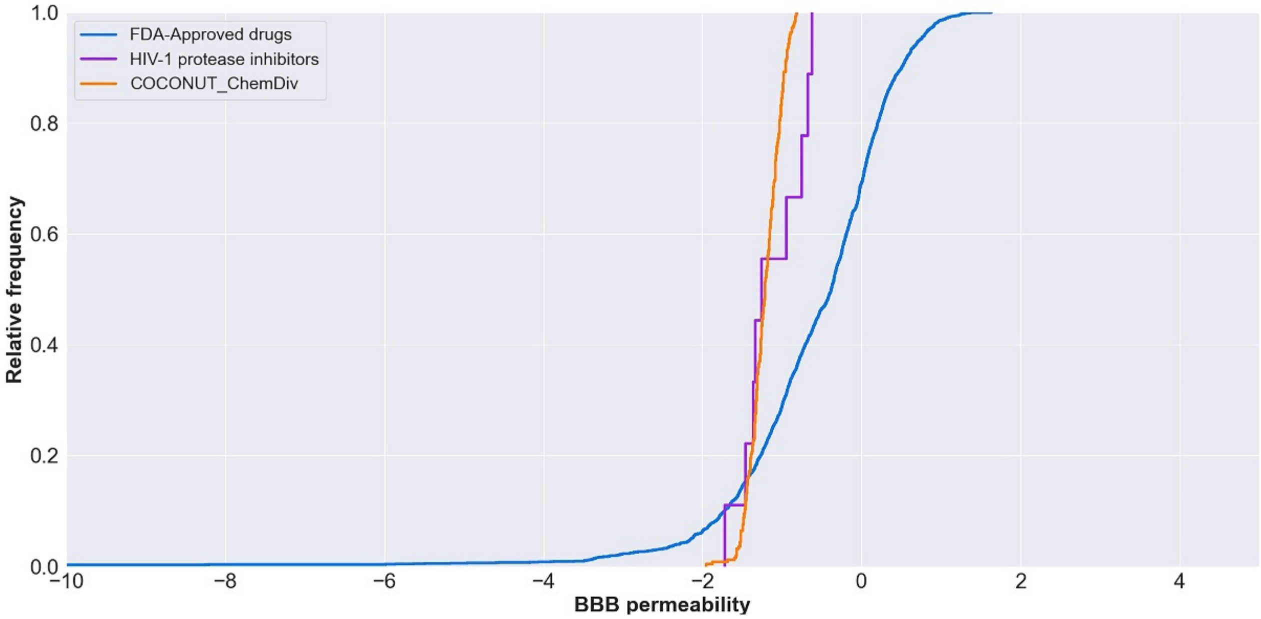


Figure 11. Distribution curve of BBB permeability. Colors represent compounds: new chemical compounds generated from COCONUT fragments and ChemDiv fragments with physicochemical properties like FDA-approved HIV-1 protease inhibitors and easily synthetically accessible (orange), FDA-approved drugs (blue), FDA-approved HIV-1 protease inhibitors (purple). The BBB permeability of FDA-approved drugs was between −34 and 2.

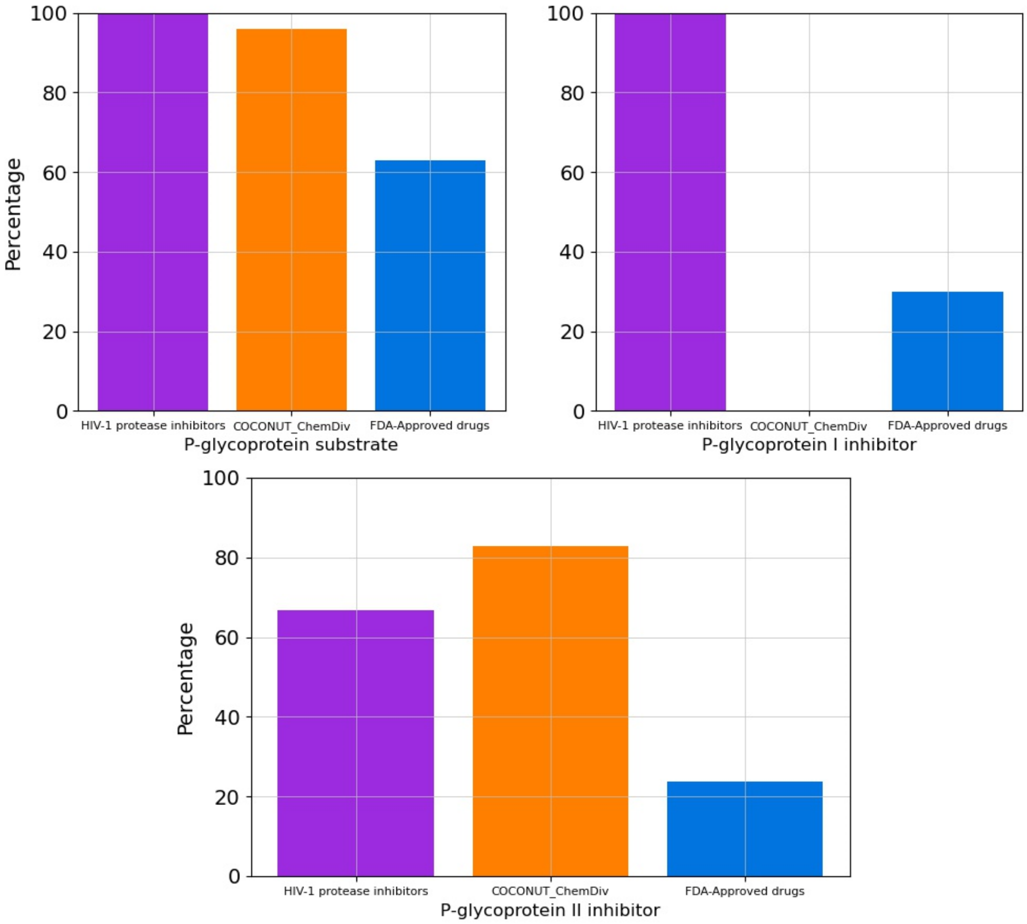


Figure 12. Percentage of compounds that are P-glycoprotein substrate, P-glycoprotein I inhibitor, and P-glycoprotein II inhibitor. Colors represent compounds: new chemical compounds generated from COCONUT fragments and ChemDiv fragments with physicochemical properties like FDA-approved HIV-1 protease inhibitors and easily synthetically accessible (orange), FDA-approved drugs (blue), FDA-approved HIV-1 protease inhibitors (purple).

3.5.3. Metabolism

The percentage of compounds CYP1A2, CYP2C19, CYP2C9, CYP2D6 and CYP3A4 inhibitors is described in Figure 13 and Table S14 in the supplementary material. No compounds generated from COCONUT and ChemDiv fragments were CYP1A2, CYP2C19, CYP2C9, CYP2D6 and CYP3A4 inhibitors. FDA-approved HIV-1 inhibitors were not CYP1A2 and CYP2D6 inhibitors similar to compounds generated from COCONUT and ChemDiv fragments. Whereas for FDA-approved HIV-1protease inhibitors, 89% were CYP3A4 inhibitors, and 33% were CYP2C19 and CYP2C9 inhibitors.

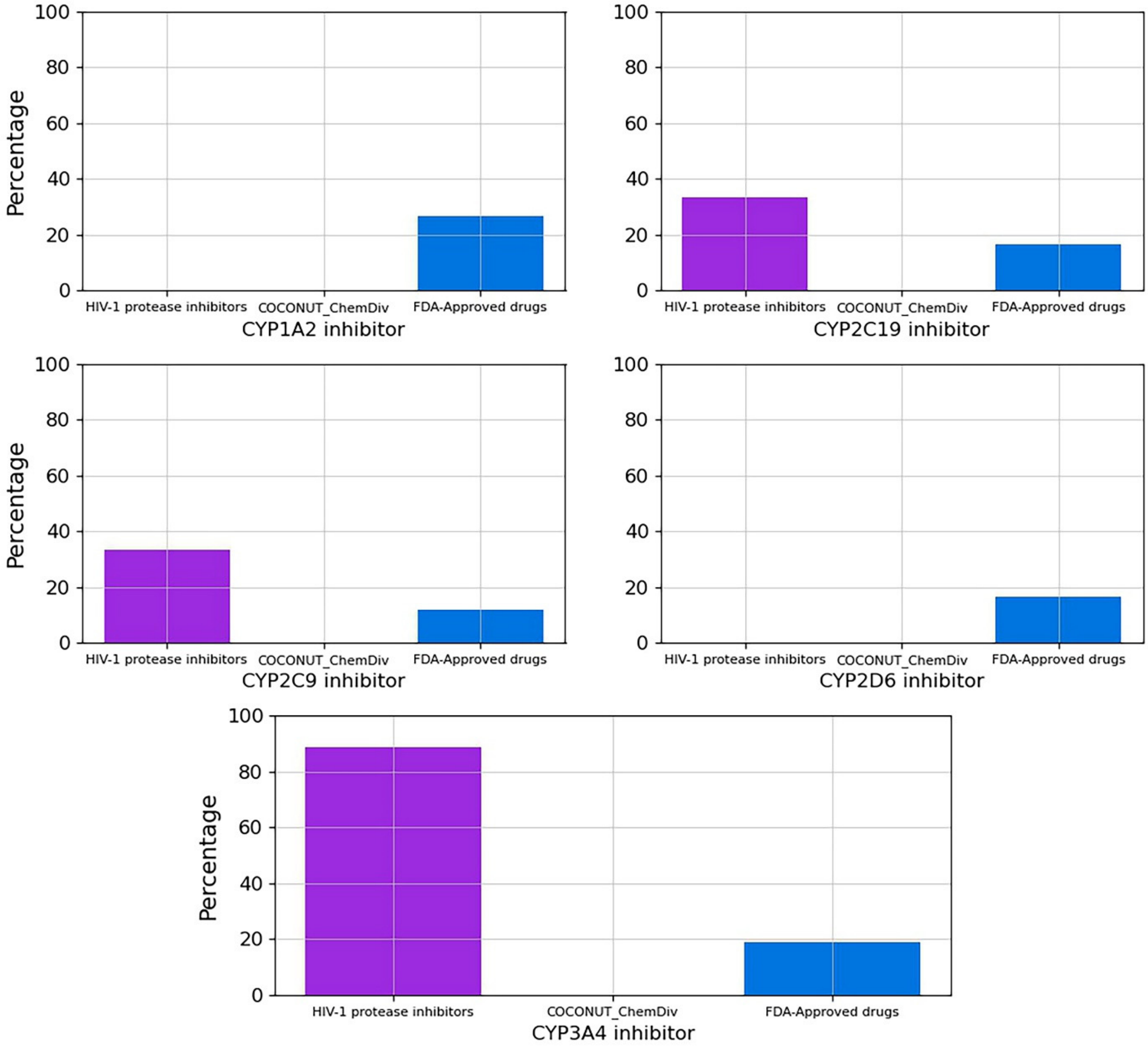


Figure 13. Percentage of compounds that inhibit the main cytochromes, CYP1A2, CYP2C19, CYP2C9, CYP2D6, CYP3A4. Colors represent compounds: new chemical compounds generated from COCONUT fragments and ChemDiv fragments with physicochemical properties like FDA-approved HIV-1 protease inhibitors and easily synthetically accessible, FDA-approved drugs (blue), FDA-approved HIV-1 protease inhibitors (purple).

3.5.4. Excretion

Clearance quantitates the irreversible removal of a drug from the measured matrix, generally, blood or plasma [55]. The total clearance logarithm expressed in units of (mL/min/Kg) is shown in Figure 14. The summary of descriptive statistics is shown in Table S15 in the Supplementary Materials. The median values of the total clearance logarithm were 0.591 for FDA-approved drugs; 0.494 for FDA-approved HIV-1 protease inhibitors, and −0.618 for compounds derived from COCONUT and ChemDiv fragments. The total clearance of FDA-approved HIV-1 protease inhibitors (0.20 $\leq$ total clearance $\leq$ 0.94) was similar to 75% FDA-approved drugs (0.27 $\leq$ total clearance $\leq$ 0.85). Whereas the total clearance of compounds generated from COCONUT and ChemDiv fragments (−1.34 $\leq$ total clearance $\leq$ 0.13) was similar to 25% FDA-approved drugs (-13.94 $\leq$ total clearance $\leq$ 0.27). The total clearance of compounds derived from COCONUT and ChemDiv fragments and FDA-approved HIV-1 inhibitors were different.

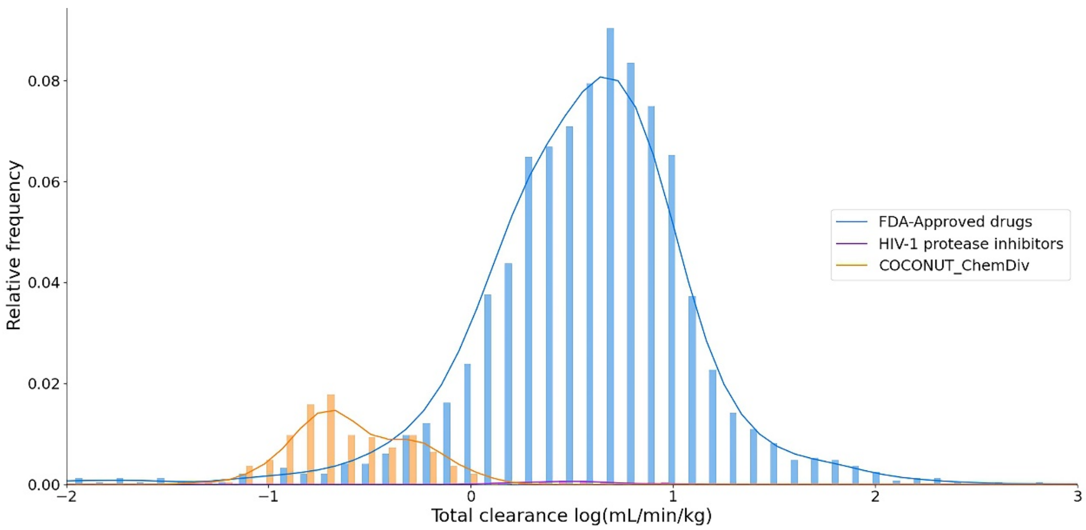


Figure 14. Distribution curve of the total clearance. Colors represent compounds: new chemical compounds generated from COCONUT fragments and ChemDiv fragments with physicochemical properties like FDA-approved HIV-1 protease inhibitors and easily synthetically accessible (orange), FDA-approved drugs (blue), FDA-approved HIV-1 protease inhibitors (purple).

3.5.5. Toxicity

Percentage of compounds from datasets that are hERG I inhibitor, hERG II inhibitor, hepatotoxicants (hepatotoxicity), and carcinogens (positive in AMES test) were described in Figure 15 and Table S16 in the supplementary material. FDA-approved HIV-1 protease inhibitors and compounds generated from COCONUT and ChemDiv fragments were not carcinogens. However, 77.22% of compounds derived from COCONUT and ChemDiv fragments were hepatotoxicants, lower than FDA-approved HIV-1 protease inhibitors (100%), and higher than FDA-approved drugs (47.42%). A total of 100% and 98.81% of compounds generated from COCONUT and ChemDiv fragments were not hERG I/II inhibitors, respectively.

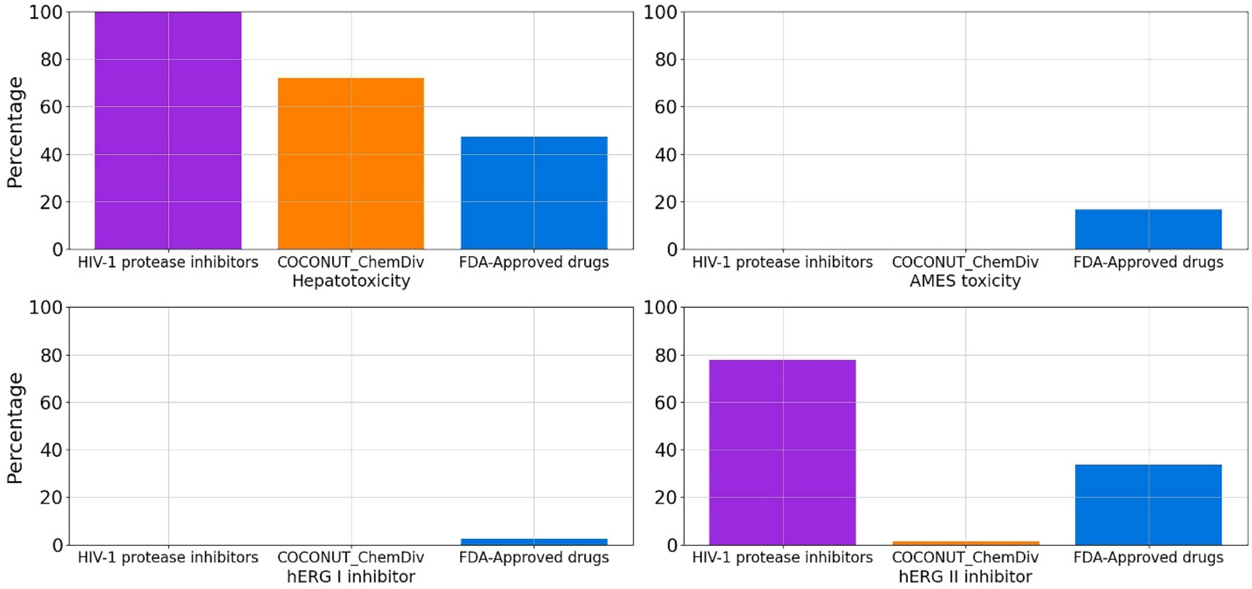


Figure 15. Percentage of compounds that are hERG I inhibitor, hERG II inhibitor, hepatotoxicity, and toxicity in AMES test in silico. Colors represent compounds: new chemical compounds generated from COCONUT fragments and ChemDiv fragments with physicochemical properties like FDA-approved HIV-1 protease inhibitors and easily synthetically accessible (orange), FDA-approved drugs (blue), FDA-approved HIV-1 protease inhibitors (purple).

4. Conclusions

We developed an HIV-1 virtual focused library using de novo design based on enumerated libraries of compounds from fragment libraries. The fragments library in-house was built from the COCONUT database, the currently largest accessible database of natural products. Using bevirimat as template, 251 out of 1534 compounds generated from COCONUT fragments, had physicochemical properties like FDA-approved HIV-1 protease inhibitors and were estimated as easy synthesizable.

Compounds generated from COCONUT fragments were more diverse than compounds generated from ChemDiv and Enamine fragments, based on chemical structure and physicochemical properties. Visual representation of the chemical space based on TMAP showed that some compounds generated from COCONUT fragments had chemical structures similar to FDA-approved drugs, such as palbociclib and pipecuronium.

ADME/Tox profiling showed that compounds generated from COCONUT fragments had adsorption (solubility and lipophilicity) and distribution (BBB permeability, P-glycoprotein substrate, and P-glycoprotein II inhibitor) similar to FDA-approved HIV-1 protease inhibitors. Concerning estimations of metabolism, no compounds generated from COCONUT fragments were CYP1A2, CYP2C19, CYP2C9, CYP2D6, and CYP3A4 inhibitors. As per excretion, the total clearance of compounds derived from COCONUT fragments and FDA-approved HIV-1 inhibitors were different, but similar to FDA-approved drugs. Compounds derived from COCONUT fragments were predicted to be no inhibitors of hERG I/II, like 97.7% and 66.4% of FDA-approved drugs, respectively. Compounds derived from COCONUT fragments were predicted to be no carcinogens.

The 251 compounds derived from COCONUT fragments with physicochemical properties like FDA-approved HIV-1 protease inhibitors, estimated as easy synthesizable, and good ADME/Tox profiling can be used in future analysis such as virtual screening to select candidates to test in biological assays. The next logical perspective of this project that this is beyond the scope of this manuscript is to conduct the chemical synthesis and experimental screening of selected compounds.

The protocol presented in this work is general and can be used to build other chemical compounds like bevirimat or other maturation inhibitors of HIV-protease. Notably, the code used for generated new chemical compounds from chemical fragments is freely available (see Data Availability statement). This can be achieved from the SMARTS and SMIRKS proposed to filter functional groups and build new chemical compounds.

Supplementary Materials: The following are available online at https://www.mdpi.com/article/10.3390/biom11121805/s1, Figure S1: Convex hull area from PCA based on physicochemical properties of new chemical compounds generated and two compound reference libraries. Table S1: Summary of fingerprint-based structural diversity of new chemical compounds generated from COCONUT, ChemDiv, and Enamine fragments, and two compound reference libraries. Figure S2: Box-whisker plots of physicochemical properties of FDA-approved drugs (blue), FDA-approved HIV-1 protease inhibitors (purple), and new chemical compounds generated from COCONUT (orange), ChemDiv (red), and Enamine (green) fragment libraries, before applying physicochemical properties filtering. Table S2: Summary of the descriptive statistics of SlogP. Table S3: Summary of the descriptive statistics of MW. Table S4: Summary of the descriptive statistics of RB. Table S5: Summary of the descriptive statistics of TPSA. Table S6: Summary of the descriptive statistics of HBA. Table S7: Summary of the descriptive statistics of HBD. Table S8: Summary of the descriptive statistics of SAscore. Table S9: Summary of the descriptive statistics of solubility (Silicos-IT LowSw). Table S10: Summary of the descriptive statistics of lipophilicity (Consensus Log P). Table S11: Summary of the descriptive statistics of HIA. Table S12: Summary of the descriptive statistics of BBB permeability. Table S13: Percentage of compounds that are P-glycoprotein substrate, P-glycoprotein I inhibitor, and P-glycoprotein II inhibitor. Table S14: Percentage of compounds that inhibit the main cytochromes, CYP1A2, CYP2C19, CYP2C9, CYP2D6, CYP3A4. Table S15: Summary of the descriptive statistics of total clearance. Table S16: Summary of the descriptive statistics of toxicity descriptors.

Author Contributions: Designed and supervised the project, J.L.M.-F.; wrote the manuscript, A.L.C.-H. and J.L.M.-F.; methodology development, J.L.M.-F. and A.L.C.-H.; data curation and

fragments generation, A.L.C.-H.; de novo design, A.L.C.-H. and F.I.S.-G.; formal analysis, A.L.C.-H. and J.L.M.-F.; data visualization, A.L.C.-H., K.E.J.-M. and J.L.M.-F. All authors have read and agreed to the published version of the manuscript.

Funding: Authors thank the Dirección General de Cómputo y de Tecnologías de Información y Comunicación (DGTIC), UNAM, for the computational resources to use Miztli supercomputer at UNAM under project LANCAD-UNAM-DGTIC-335.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All datasets used in this study are available on https://figshare.com/s/ceb58d58e8f5585ce67e (accessed on 5 November 2021). TMAP_chemical_space_visualization.html; All_fragments_COCONUT_V4_184769.csv; HIV-protease_inhibitors_from_ChemDiv.csv; HIV_protease_inhibitors_from_COCONUT.csv; HIV_protease_inhibitors_from_Enamine.csv; FDA_APPROVED_DRUGS.csv; HIV_PROTEASE_INHIBITORS.csv; ADMETOX profiling_pkCSM.csv; ADMETOX profiling_swissadme.csv. Code used for generated new chemical compounds from chemical fragments are available on https://github.com/DIFACQUIM/De-novo-desing-of-HIV-1-inhibitors (accessed on 5 November 2021).

Acknowledgments: A.L.-C.H. and F.I.S.-G. thankful to CONACyT for the granted scholarship numbers 847870 and 848061, respectively. K.E.J.-M. thanks Dirección General de Asuntos del Personal Académico for the scholarship associated with the project PAPIIT IN201321. We thank Reymond's research group for developing Faerun library for Python.

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

References

1. HIV/AIDS. Available online: https://www.who.int/news-room/fact-sheets/detail/hiv-aids (accessed on 15 July 2021).
2. Zulfiqar, H.F.; Javed, A.; Sumbal; Afroze, B.; Ali, Q.; Akbar, K.; Nadeem, T.; Rana, M.A.; Nazar, Z.A.; Nasir, I.A.; et al. HIV Diagnosis and Treatment through Advanced Technologies. *Front. Public Health* **2017**, *5*, 32. [CrossRef]
3. Lv, Z.; Chu, Y.; Wang, Y. HIV protease inhibitors: A review of molecular selectivity and toxicity. *HIV AIDS* **2015**, *7*, 95–104. [CrossRef]
4. FDA. Available online: https://www.fda.gov/consumers/free-publications-women/hiv-and-aids-medicines-help-you (accessed on 28 April 2021).
5. Schneider, G.; Clark, D.E. Automated de novo drug design: Are we nearly there yet? *Angew. Chem. Int. Ed.* **2019**, *58*, 10792–10803. [CrossRef]
6. Torjesen, I. Drug Development: The Journey of a Medicine from Lab to Shelf. Available online: https://pharmaceutical-journal.com/article/feature/drug-development-the-journey-of-a-medicine-from-lab-to-shelf (accessed on 29 May 2021).
7. Medina-Franco, J.L. Grand challenges of computer-aided drug design: The road ahead. *Front. Drug Discov.* **2021**, *1*, 728551. [CrossRef]
8. Bung, N.; Krishnan, S.R.; Bulusu, G.; Roy, A. De novo design of new chemical entities for SARS-CoV-2 using artificial intelligence. *Future Med. Chem.* **2021**, *13*, 575–585. [CrossRef] [PubMed]
9. Liu, X.; IJzerman, A.P.; van Westen, G.J.P. Computational approaches for de novo drug design: Past, present, and future. In *Artificial Neural Networks*; Cartwright, H., Ed.; Springer: New York, NY, USA, 2021; pp. 139–165, ISBN 978-1-0716-0826-5.
10. Mouchlis, V.D.; Afantitis, A.; Serra, A.; Fratello, M.; Papadiamantis, A.G.; Aidinis, V.; Lynch, I.; Greco, D.; Melagraki, G. Advances in De Novo Drug Design: From Conventional to Machine Learning Methods. *Int. J. Mol. Sci.* **2021**, *22*, 1676. [CrossRef]
11. Meyers, J.; Fabian, B.; Brown, N. De novo molecular design and generative models. *Drug Discov. Today* **2021**, *26*, 2707–2715. [CrossRef]
12. Devi, R.V.; Sathya, S.S.; Coumar, M.S. Evolutionary algorithms for de novo drug design—A survey. *Appl. Soft Comput.* **2015**, *27*, 543–552. [CrossRef]
13. Hartenfeller, M.; Schneider, G. De Novo Drug Design. In *Chemoinformatics and Computational Chemical Biology*; Bajorath, J., Ed.; Humana Press: Totowa, NJ, USA, 2011; pp. 299–323, ISBN 978-1-60761-839-3.
14. Erlanson, D.A.; Fesik, S.W.; Hubbard, R.E.; Jahnke, W.; Jhoti, H. Twenty years on: The impact of fragments on drug discovery. *Nat. Rev. Drug Discov.* **2016**, *15*, 605–619. [CrossRef]
15. Osborne, J.; Panova, S.; Rapti, M.; Urushima, T.; Jhoti, H. Fragments: Where are we now? *Biochem. Soc. Trans.* **2020**, *48*, 271–280. [CrossRef]
16. Shinde, P.B.; Bhowmick, S.; Alfantoukh, E.; Patil, P.C.; Wabaidur, S.M.; Chikhale, R.V.; Islam, M.A. De novo design based identification of potential HIV-1 integrase inhibitors: A pharmacoinformatics study. *Comput. Biol. Chem.* **2020**, *88*, 107319. [CrossRef]

17. Ghiandoni, G.M.; Bodkin, M.J.; Chen, B.; Hristozov, D.; Wallace, J.E.A.; Webster, J.; Gillet, V.J. Enhancing reaction-based de novo design using a multi-label reaction class recommender. *J. Comput. Aided Mol. Des.* **2020**, *34*, 783–803. [CrossRef]
18. Saldívar-González, F.I.; Huerta-García, C.S.; Medina-Franco, J.L. Chemoinformatics-based enumeration of chemical libraries: A tutorial. *J. Cheminform.* **2020**, *12*, 64. [CrossRef] [PubMed]
19. Prado-Romero, D.L.; Medina-Franco, J.L. Advances in the exploration of the epigenetic televant chemical space. *ACS Omega* **2021**, *6*, 22478–22486. [CrossRef]
20. Atanasov, A.G.; Zotchev, S.B.; Dirsch, V.M.; International Natural Product Sciences Taskforce; Supuran, C.T. Natural products in drug discovery: Advances and opportunities. *Nat. Rev. Drug Discov.* **2021**, *20*, 200–216. [CrossRef] [PubMed]
21. Barnes, E.C.; Kumar, R.; Davis, R.A. The use of isolated natural products as scaffolds for the generation of chemically diverse screening libraries for drug discovery. *Nat. Prod. Rep.* **2016**, *33*, 372–381. [CrossRef]
22. Chávez-Hernández, A.L.; Sánchez-Cruz, N.; Medina-Franco, J.L. A Fragment library of natural products and its comparative chemoinformatic characterization. *Mol. Inform.* **2020**, *39*, 2000050. [CrossRef] [PubMed]
23. Karageorgis, G.; Foley, D.J.; Laraia, L.; Waldmann, H. Principle and design of pseudo-natural products. *Nat. Chem.* **2020**, *12*, 227–235. [CrossRef]
24. ChemDiv. Available online: https://store.chemdiv.com/ (accessed on 19 July 2021).
25. Enamine. Available online: https://enamine.net/compound-collections/fragment-collection (accessed on 16 July 2021).
26. Sorokina, M.; Steinbeck, C. Review on natural products databases: Where to find data in 2020. *J. Cheminform.* **2020**, *12*, 20. [CrossRef]
27. Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31–36. [CrossRef]
28. toolkit RDKit. Available online: http://rdkit.org (accessed on 21 May 2021).
29. MolVS. Available online: https://molvs.readthedocs.io/en/latest/ (accessed on 21 May 2021).
30. 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 settings1PII of original article: S0169-409X(96)00423-1. The article was originally published in Advanced Drug Delivery Reviews 23 (1997) 3. *Adv. Drug Deliv. Rev.* **2001**, *46*, 3–26. [CrossRef]
31. 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. [CrossRef]
32. Lewell, X.Q.; Judd, D.B.; Watson, S.P.; Hann, M.M. RECAPRetrosynthetic combinatorial analysis procedure: A powerful new technique for identifying privileged molecular fragments with useful applications in combinatorial chemistry. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 511–522. [CrossRef] [PubMed]
33. Zhao, Y.; Chen, C.-H.; Morris-Natschke, S.L.; Lee, K.-H. Design, synthesis, and structure activity relationship analysis of new betulinic acid derivatives as potent HIV inhibitors. *Eur. J. Med. Chem.* **2021**, *215*, 113287. [CrossRef] [PubMed]
34. Martin, D.E.; Salzwedel, K.; Allaway, G.P. Bevirimat: A novel maturation inhibitor for the treatment of hiv-1 infection. *Antivir. Chem. Chemother.* **2008**, *19*, 107–113. [CrossRef] [PubMed]
35. Lazerwith, S.E.; Siegel, D.; McFadden, R.M.; Mish, M.R.; Tse, W.C. *5.19—New Antiretrovirals for HIV and Antivirals for HBV*; Chackalamannil, S., Rotella, D., Ward, S.E., Eds.; Elsevier: Oxford, UK, 2017; pp. 628–664, ISBN 978-0-12-803201-5.
36. Qian, K.; Kuo, R.-Y.; Chen, C.-H.; Huang, L.; Morris-Natschke, S.L.; Lee, K.-H. Anti-AIDS Agents 81. Design, synthesis, and structure−activity relationship study of betulinic acid and moronic acid derivatives as potent HIV maturation inhibitors. *J. Med. Chem.* **2010**, *53*, 3133–3141. [CrossRef]
37. Huang, Q.; Chen, H.; Luo, X.; Zhang, Y.; Yao, X.; Zheng, X. Structure and anti-HIV activity of betulinic acid analogues. *Curr. Med. Sci.* **2018**, *38*, 387–397. [CrossRef]
38. Rogers, D.; Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742–754. [CrossRef]
39. Durant, J.L.; Leland, B.A.; Henry, D.R.; Nourse, J.G. Reoptimization of MDL Keys for use in drug discovery. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1273–1280. [CrossRef]
40. Probst, D.; Reymond, J.-L. Visualization of very large high-dimensional data sets as minimum spanning trees. *J. Cheminform.* **2020**, *12*, 12. [CrossRef]
41. TMAP. Available online: https://tmap.gdb.tools/ (accessed on 14 September 2021).
42. Greener, J.G.; Kandathil, S.M.; Moffat, L.; Jones, D.T. A guide to machine learning for biologists. *Nat. Rev. Mol. Cell Biol.* **2021**. [CrossRef]
43. Sánchez-Cruz, N.; Pilón-Jiménez, B.A.; Medina-Franco, J.L. Functional group and diversity analysis of BIOFACQUIM: A Mexican natural product database. *F1000Research* **2020**, *8*, 2071. [CrossRef]
44. Chávez-Hernández, A.L.; Sánchez-Cruz, N.; Medina-Franco, J.L. Fragment library of natural products and compound databases for drug discovery. *Biomolecules* **2020**, *10*, 1518. [CrossRef] [PubMed]
45. Ertl, P.; Schuffenhauer, A. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *J. Cheminform.* **2009**, *1*, 8. [CrossRef]
46. Wishart, D.S.; Feunang, Y.D.; Guo, A.C.; Lo, E.J.; Marcu, A.; Grant, J.R.; 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. [CrossRef]
47. Daina, A.; Michielin, O.; Zoete, V. SwissADME: A free web tool to evaluate pharmacokinetics, drug-likeness and medicinal chemistry friendliness of small molecules. *Sci. Rep.* **2017**, *7*, 42717. [CrossRef]

48. Pires, D.E.V.; Blundell, T.L.; Ascher, D.B. pkCSM: Predicting small-molecule pharmacokinetic and toxicity properties using graph-based signatures. *J. Med. Chem.* **2015**, *58*, 4066–4072. [CrossRef] [PubMed]
49. Durán-Iturbide, N.A.; Díaz-Eufracio, B.I.; Medina-Franco, J.L. In silico ADME/Tox profiling of natural products: A focus on BIOFACQUIM. *ACS Omega* **2020**, *5*, 16076–16084. [CrossRef]
50. Saldívar-González, F.I.; Lenci, E.; Calugi, L.; Medina-Franco, J.L.; Trabocchi, A. Computational-aided design of a library of lactams through a diversity-oriented synthesis strategy. *Bioorg. Med. Chem.* **2020**, *28*, 115539. [CrossRef]
51. Laurini, R. *5—Geographic Relations*; Laurini, R.B.T.-G.K.I., Ed.; Elsevier: Amsterdam, The Netherlands, 2017; pp. 83–109, ISBN 978-1-78548-243-4.
52. Ganesan, A. The impact of natural products upon modern drug discovery. *Curr. Opin. Chem. Biol.* **2008**, *12*, 306–317. [CrossRef] [PubMed]
53. Tinworth, C.P.; Young, R.J. Facts, Patterns, and principles in drug discovery: Appraising the Rule of 5 with measured physicochemical data. *J. Med. Chem.* **2020**, *63*, 10091–10108. [CrossRef] [PubMed]
54. Bergström, C.A.S.; Yazdanian, M. Lipophilicity in drug development: Too much or not enough? *AAPS J.* **2016**, *18*, 1095–1100. [CrossRef] [PubMed]
55. Smith, D.A.; Beaumont, K.; Maurer, T.S.; Di, L. Clearance in drug design. *J. Med. Chem.* **2019**, *62*, 2245–2255. [CrossRef] [PubMed]

Contents lists available at ScienceDirect

Artificial Intelligence in the Life Sciences

journal homepage: www.elsevier.com/locate/ailsci

Short Communications

Natural products subsets: Generation and characterization

Ana L. Chávez-Hernández, José L. Medina-Franco*

DIFACQUIM Research Group, Department of Pharmacy, School of Chemistry, Universidad Nacional Autónoma de México, Avenida Universidad 3000, México City 04510, Mexico

ARTICLE INFO

Keywords:
Artificial intelligence
Chemical space
Chemical multiverse
Chirality
De novo design
Deep learning
Molecular fingerprint
Molecular representation
Natural products

ABSTRACT

Natural products are attractive for drug discovery applications because of their distinctive chemical structures, such as an overall large fraction of sp^3 carbon atoms, chiral centers (both features associated with structural complexity), large chemical scaffolds, and diversity of functional groups. Furthermore, natural products are used in *de novo* design and have inspired the development of pseudo-natural products using generative models. Public databases such as the Collection of Open NatUral ProdUcTs and the Universal Natural Product database (UNPD) are rich sources of structures to be used in generative models and other applications. In this work, we report the selection and characterization of the most diverse compounds of natural products from the UNPD using the MaxMin algorithm. The subsets generated with 14,994, 7,497, and 4,998 compounds are publicly available at https://github.com/DIFACQUIM/Natural-products-subsets-generation. We anticipate that the subsets will be particularly useful in building generative models based on natural products by research groups, particularly those with limited access to extensive supercomputer resources.

1. Introduction

Natural products and fragments derived from natural products have been attractive in drug design and development because of their distinctive chemical structures. For example, natural products have, in general, an overall larger diversity of functional groups and larger structural complexity than synthetic molecules [1–4]. However, a drawback for some natural products, particularly those with sizeable structural complexity, is that they can be challenging to synthesize. A workaround for this issue is the so-called pseudo-natural products which are synthetically feasible compounds generated through a *de novo* combination of natural product fragments [3]. Pseudo-natural products allow the exploration of uncharted areas of biologically relevant chemical space that are different from the chemical space covered by the compounds from which they are generated.

The Collection of Open NatUral ProdUcTs (COCONUT) and the Universal Natural Product Database (UNPD) are two large compound databases. COCONUT [5] is arguably the most extensive public natural product database, with 389,184 unique structures. UNPD [6], with 153,375 natural products, is the second-largest public natural product database. A distinctive feature of UNPD compared to COCONUT is that compounds in UNPD contain chirality information. In Latin America, several public natural products databases compile the compounds isolated and characterized from the country of origin. Examples are NuBBE$_{DB}$ [7,8], SistematX [9,10] from Brazil; CIFPMA [11,12] from Panama; PeruNPDB [13] from Peru; and BIOFACQUIM [14,15] from Mexico. The latter database contains the structures of 531 natural products.

De novo design generates virtual molecules from scratch. It filters structures generated using several scoring functions and assesses synthetic chemical feasibility to remove reactive and unrealistic compounds [16]. *De novo* design based on generative algorithms such as deep learning involves using neural networks [17,18] and databases with many compounds [19]. In the source compound databases, many compounds with three-dimensional information (*e.g.*, stereochemistry) are relevant to build robust generative models [20]. However, for several research groups, it is difficult to access supercomputer resources to handle many compounds to obtain appropriate subsets that impact the model prediction [21,22].

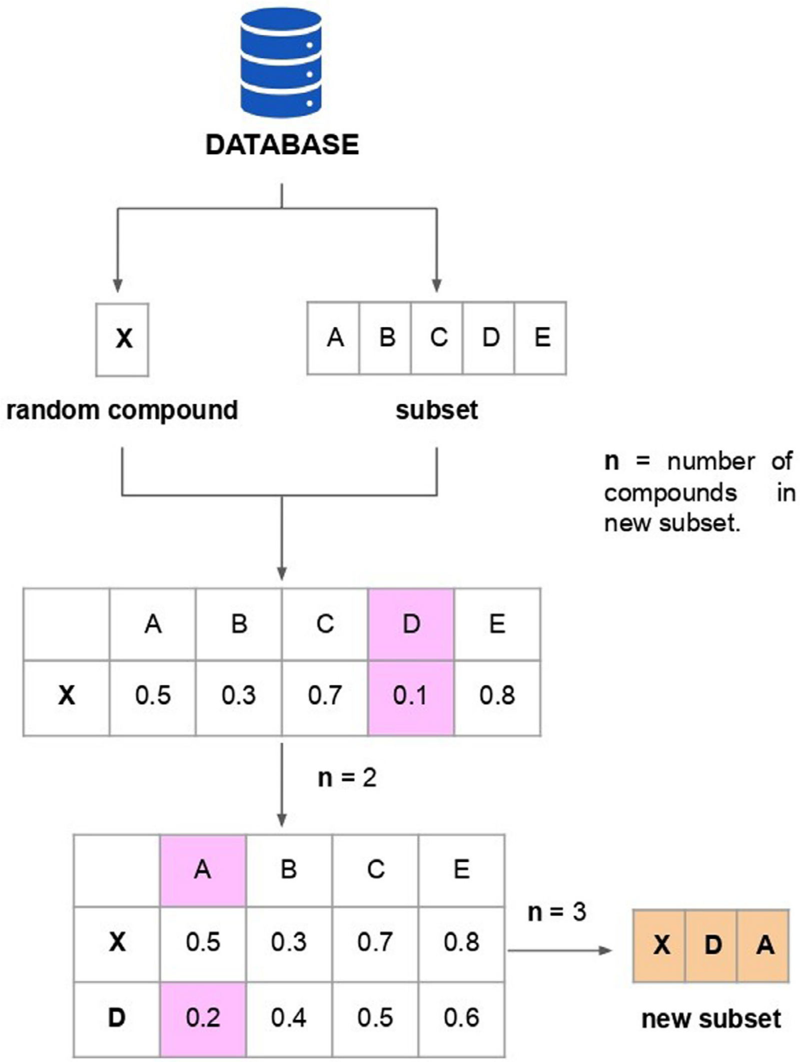
This Communication reports the selection and characterization of the most diverse compounds from UNPD. The subsets were selected using a dissimilarity-based compound selection (DBCS) method, the MaxMin algorithm [23]. The compound subsets were characterized by the Natural Product Likeness (NPL) score [24], structural diversity, and distribution in chemical space. The structural diversity was assessed with the Tanimoto coefficient and molecular fingerprints of different designs. Chemical space analysis was performed through the visual representation of the chemical multiverse [25] using T-distributed Stochastic Neighbor

Abbreviations: COCONUT, Collection of Open NatUral ProdUcTs; DNMT, DNA methyltransferase; ECFP, extended connectivity fingerprint; TMAP, Tree MAP; NPL, natural product-likeness; UNPD, Universal Natural Product database.

* Correspondence author.

E-mail address: medinajl@unam.mx (J.L. Medina-Franco).

https://doi.org/10.1016/j.ailsci.2023.100066
Received 20 December 2022; Received in revised form 5 February 2023; Accepted 21 February 2023
Available online 26 February 2023



	A	B	C	D	E
X	0.5	0.3	0.7	0.1	0.8

	A	B	C	E
X	0.5	0.3	0.7	0.8
D	0.2	0.4	0.5	0.6

Fig. 1. MaxMin algorithm implemented in this work. A database (UNPD) was split into a given number of subsets. From each subset, a random compound (X) was selected. The molecular similarity was calculated between X and each compound (A, B, C, D, and E) from the subset using the Tanimoto coefficient and the ECFP4 fingerprint. In this figure, only a subset (that contains the compounds A, B, C, D, and E) is shown for illustrative purposes. The compound with the smallest molecular similarity was chosen and deleted from the original subset. The process was repeated to generate a new subset with the number of compounds desired.

Embedding t-SNE [26] and the recently proposed Tree MAP (TMAP) [27]. We anticipate that the natural product subsets generated and thoroughly characterized in this work will be helpful to the scientific community in a variety of tasks including building generative models, in particular to those research groups with limited access to large supercomputer resources.

2. Materials and methods

2.1. Data sets

For this study, different types of databases with compounds from various origins were used including three data sets of natural products, and a data set focused on a specific molecular target of pharmaceutical relevance. Natural product databases from different sources were employed to compare the NPL score of the UNPD subsets (described in the *Natural product-likeness* section). The data set focused on a specific molecular target and the natural products from different sources were used to evaluate and compare the molecular diversity of UNPD subsets. Using a focused database gives a value of low structural diversity, which serves as a reference to compare the structural diversity of the UNPD subsets. Likewise, the chemical structures from other natural product databases served as reference compound collections to evaluate the diversity of the data sets newly generated. Each type of data set and its contents are described below.

Three data sets of natural products were used, namely, COCONUT, UNPD, and BIOFACQUIM. COCONUT [5] had 389,184 unique structures, not including chirality information. UNPD [6] with 153,372 natural products, encoding their chirality. BIOFACQUIM [14,15] is a database with 531 natural products isolated and characterized in Mexico, and the chirality of the compounds is included. A data set with 715 compounds focused on DNA methyltransferase 1 (DNMT1) inhibitors. DNMT1 is an epigenetic target relevant to drug discovery [28,29]. The data set of DNMT1 inhibitors was obtained from the ChEMBL database, release 31 [30,31].

2.2. Data set standardization

The preparation of compounds, encoded in SMILES strings [32] was performed using the open-source cheminformatics toolkit RDKit [33] and MolVS [34]. Compounds with valence errors or any chemical element other than H, B, C, N, O, F, Si, P, S, Cl, Se, Br, and I, were removed. Stereochemistry information, when available, was retained. Compounds with multiple components were split, and the largest component was retained. The remaining compounds were neutralized and reionized to generate the corresponding canonical tautomer.

2.3. Natural product-likeness

The NPL score, introduced by Ertl et al. [24], was computed for all compounds in COCONUT, UNPD, and BIOFACQUIM. The NPL score

Table 1
Descriptive statistics of NPL scores computed for COCONUT, UNPD, and BIOFACQUIM.

Dataset	COCONUT	UNPD	BIOFACQUIM
count	389,184	153,372	531
mean	0.89	1.81	1.73
[a]std	1.38	0.96	0.90
[b]min	−3.53	−2.51	−0.36
[c]Q1	−0.32	1.14	1.03
median	0.97	1.87	1.65
[d]Q3	2.01	2.56	2.45
[e]max	4.08	4.08	3.87

[a] std: standard deviation.

[b] min: minimum value.

[c] Q1: value under which 25% of data points are found in increasing order.

[d] Q3: value under which 75% of data points are found in increasing order.

[e] max: maximum value.

ranges between −5 (if the compound is more similar to a synthetic compound) and 5 (if the compound is more similar to a natural product).

2.4. Selection of sets of diverse compounds

Dissimilarity-based compound selection (DBCS) methods allow the identification of diverse compound data sets by directly calculating distances or dissimilarities between the chemical structures [23]. Among the DBCS methods, the MaxMin algorithm reduces the number of compounds choosing the molecules with the largest diversity from the original databases. Three subsets of UNPD were generated using the MaxMin algorithm as follows (Fig. 1): The initial 153,372 compounds from UNPD were split into three ways: (A) thirty random subsets of 5,000 compounds each; (B) fifteen random subsets of 10,000 compounds each; and (C) ten random subsets of 15,000 compounds each. A new diverse set with 500 compounds was selected from each one using MaxMin [23]. First, a random compound was picked from each subset. The binary similarity between the compound selected (query set) and the remaining compounds (target set) was calculated. A new compound was selected from the target set if this had the lowest similarity value and then removed from the target set. The iteration process continued until the number of compounds desired set to 500 was reached. In total, three diverse subsets from UNPD were generated: UNPD-A (15,000 compounds), UNPD-B (7,500), and UNPD-C (5,000). For the diversity selection, we used the Tanimoto coefficient and the extended connectivity fingerprint (ECFP) [35] of 1024-bits and diameter 4 (ECFP4). For the diversity set calculations, we used an E5–2670v1 processor, 16 cores, and 64 Gbytes of RAM). The maximum calculation time for each initial subset of the natural product subsets were: UNPD-A (15,000 compounds), 19,989.21 s; UNPD-B (7,500 compounds), 102,569.61 s, and UNPD-C (5,000 compounds), 209,241.25 s.

2.5. Structural diversity

The structural diversity of three UNPD subsets, UNPD, BIOFACQUIM, and DNMT1, was compared through the distribution of the pairwise similarity values generated with the Tanimoto coefficient using three molecular fingerprints Molecular ACCes System (MACCS) Keys (166-bits) [36], extended connectivity fingerprint (ECFP) [35] of 1024-bits with diameter 6 (ECFP6) and diameter 4 (ECFP4).

2.6. Chemical multiverse visualization

The chemical multiverse [25] of the three UNPD subsets was compared to the chemical multiverse of the entire UNPD collection. A chemical multiverse is a group of numerical vectors that differently describe a set of molecules depending on the molecular representation [25]. So, each chemical space is an M-dimensional cartesian space in which compounds are located by a set of M descriptors. Each type of molecular

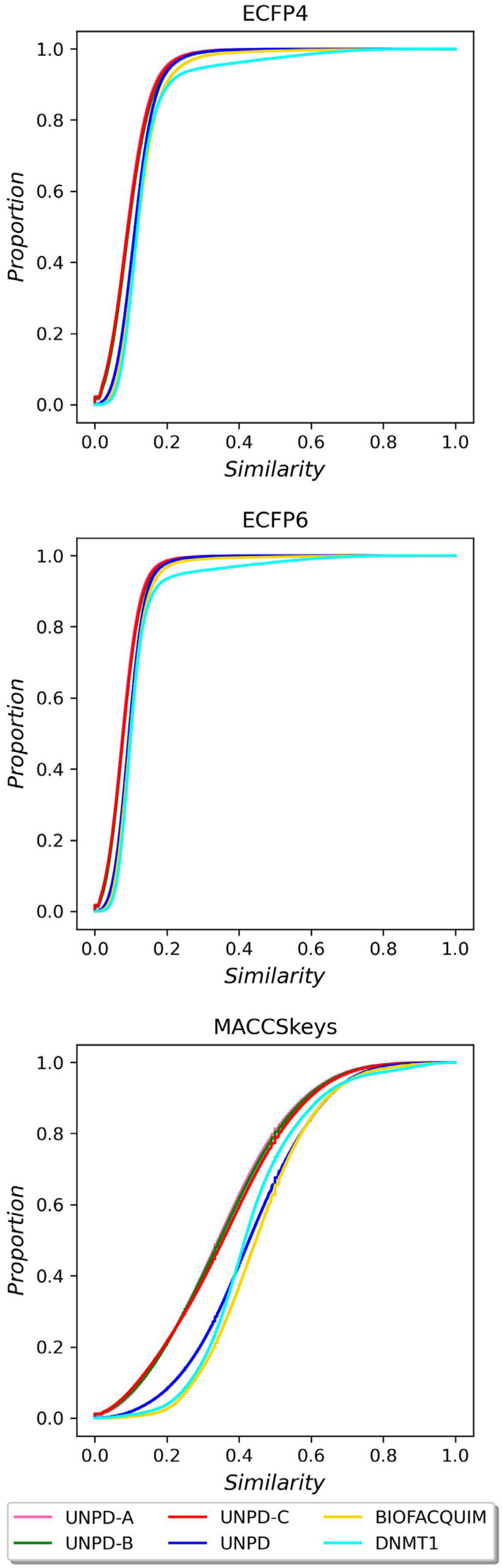


Fig. 2. Cumulative distribution functions of the pairwise Tanimoto similarity using MACCS keys (166-bits), ECFP4, and ECFP6, as molecular representations. The datasets are UNPD-A (pink line), UNPD-B (green), UNPD-C (red), and UNPD (blue).

Table 2
Summary of the structural diversity of the three UNPD subsets, BIOFACQUIM, and DNMT1 datasets. The number of initial and unique compounds from each database is indicated.

Dataset	Compounds	Unique compounds	MACCS keys (166-bits)[a]	ECFP4[a]	ECFP6[a]
UNPD-A	15,000	14,994	0.341	0.091	0.077
UNPD-B	7,500	7,497	0.346	0.094	0.08
UNPD-C	5,000	4,998	0.356	0.092	0.079
UNPD	153,372	153,372	0.43	0.111	0.094
BIOFACQUIM	531	531	0.447	0.119	0.099
DNMT1	714	714	0.417	0.119	0.1

[a] The structural diversity is reported as the median value of the distribution of the pairwise comparison using the Tanimoto coefficient and molecular fingerprints (MACCS keys, ECFP4 and ECFP6). A full diversity assessment is presented at the cumulative distribution functions of the pairwise similarity values in Fig. 2.

Table 3
Descriptive statistics of the number of hydrogen bond donors and acceptors (HBD, HBA) of the UNPD subsets and the entire UNPD.

Property	HBD				HBA			
Dataset	UNPD-A	UNPD-B	UNPD-C	UNPD	UNPD-A	UNPD-B	UNPD-C	UNPD
count	14,994	7,497	4,998	153,372	14,994	7,497	4,998	153,372
mean	2.51	2.6	2.69	3.15	5.58	5.65	5.94	7.05
[a]std	3.17	3.10	3.40	3.55	4.95	4.80	5.46	5.68
[b]min	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
[c]Q1	0.00	1.00	0.00	1.00	2.00	3.00	2.00	3.00
median	2.00	2.00	2.00	2.00	4.00	5.00	5.00	5.00
[d]Q3	3.00	3.00	4.00	4.00	7.00	7.00	8.00	9.00
[e]max	36.00	33.00	33.00	36.00	53.00	53.00	52.00	54.00

[a] std: standard deviation.
[b] min: minimum value.
[c] Q1: value under which 25% of data points are found in increasing order.
[d] Q3: value under which 75% of data points are found in increasing order.
[e] max: maximum value.

Table 4
Descriptive statistics of the number of rotatable bonds (RB) and LogP values of the UNPD subsets and the entire UNPD.

Property	RB				LogP			
Dataset	UNPD-A	UNPD-B	UNPD-C	UNPD	UNPD-A	UNPD-B	UNPD-C	UNPD
count	14,994	7,497	4,998	153,372	14,994	7,497	4,998	15,372
mean	4.74	5.34	4.81	5.97	2.94	2.98	2.94	2.94
[a]std	6.02	6.66	5.69	6.08	3.02	3.00	3.13	3.12
[b]min	0.00	0.00	0.00	0.00	−18.53	−14.10	−18.16	−20.82
[c]Q1	1.00	1.00	1.00	2.00	1.46	1.40	1.43	1.33
median	3.00	3.00	3.00	4.00	2.87	2.79	2.90	2.96
[d]Q3	6.00	7.00	6.00	8.00	4.32	4.24	4.45	4.58
[e]max	59.00	63.00	59.00	63.00	24.43	23.11	21.63	25.12

[a] std: standard deviation.
[b] min: minimum value.
[c] Q1: value under which 25% of data points are found in increasing order.
[d] Q3: value under which 75% of data points are found in increasing order.
[e] max: maximum value.

Table 5
Descriptive statistics of the topological surface area (TPSA) and molecular weight (MW) of the UNPD subsets and the entire UNPD.

Property	TPSA				MW			
Dataset	UNPD-A	UNPD-B	UNPD-C	UNPD	UNPD-A	UNPD-B	UNPD-C	UNPD
count	14,994	7,497	4,998	153,372	14,994	7,497	4,998	153,372
mean	90.78	93.58	94.78	114.63	371.94	369.57	388.43	450.55
[a]std	82.74	80.35	90.04	93.94	196.43	194.20	210.40	228.63
[b]min	0.00	0.00	0.00	0.00	16.04	16.04	17.03	16.04
[c]Q1	40.46	46.53	39.10	54.37	246.31	248.32	252.28	302.28
median	69.67	74.60	69.92	85.97	330.29	328.45	342.47	396.20
[d]Q3	112.05	116.20	119.61	145.91	445.60	444.48	466.74	534.66
[e]max	877.36	927.43	900.36	927.43	1,887.28	1,889.30	1,875.31	1,907.36

[a] std: standard deviation.
[b] min: minimum value.
[c] Q1: value under which 25% of data points are found in increasing order.
[d] Q3: value under which 75% of data points are found in increasing order.
[e] max: maximum value.

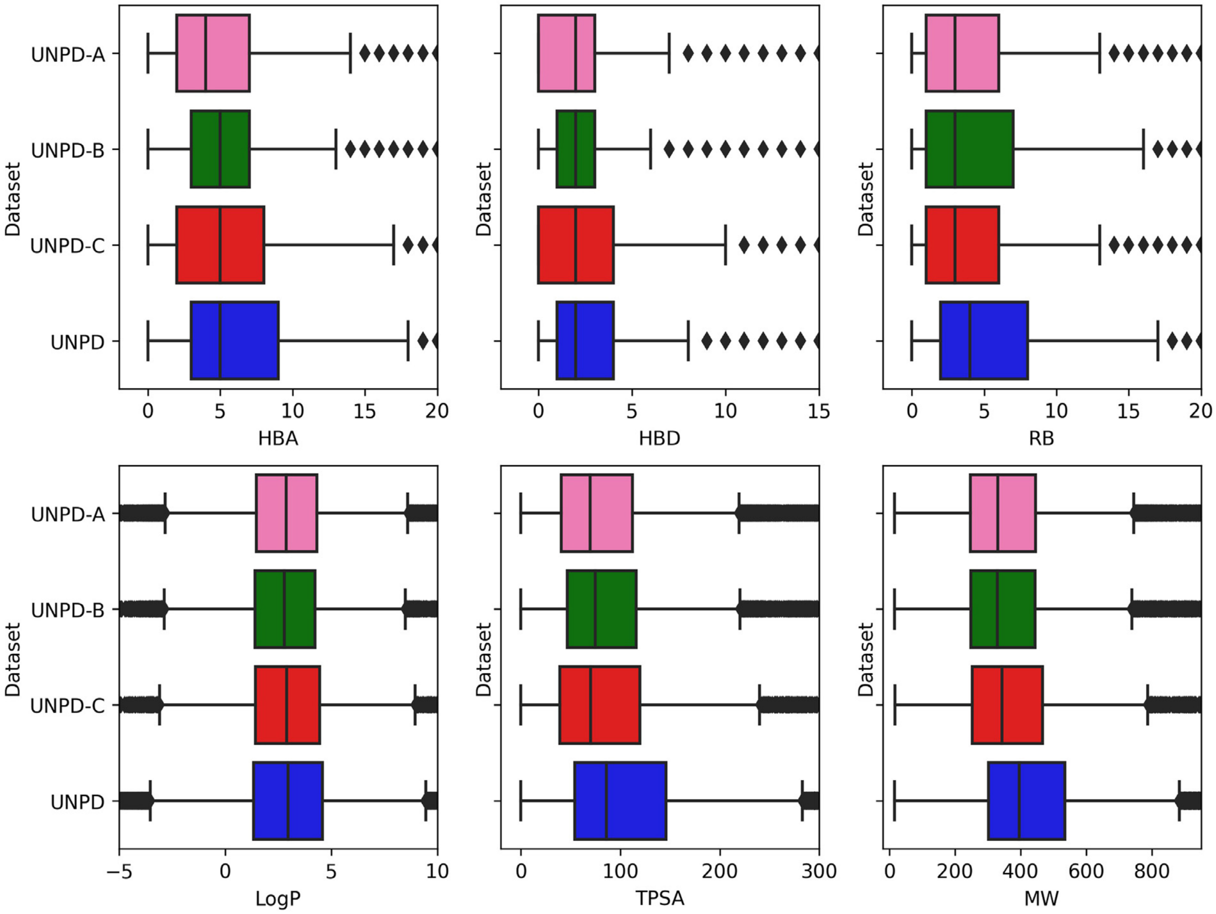


Fig. 3. Box-whisker plots of six properties of pharmaceutical relevance: hydrogen bond donors (HBD), hydrogen bond acceptors (HBA), topological polar surface area (TPSA), number of rotatable bonds (RB), molecular weight (MW), and partition coefficient octanol/water (LogP). The datasets are represented in different colors, UNPD-A (pink), UNPD-B (green), UNPD-C (red), and UNPD (blue). Black diamonds represent outliers.

fingerprint generates a distinct chemical space, and the sets of chemical spaces comprise the chemical multiverse. For this work, the visual representation of the chemical multiverse was done using t-SNE [26] and TMAPs [27] as visualization methods, and MACCS keys (166 bits), ECFP4, and ECFP6 as molecular representations.

The t-SNE generates plots that organize compounds. Similar compounds form clusters and dissimilar compounds are distant from each other. TMAP allows visualizing of many chemical compounds through the distance between clusters. Local Sensitive Hashing (LSH) allows each compound to be grouped hierarchically according to common substructures using molecular fingerprints. The number of nearest neighbors, $k = 50$, and the factor used by the augmented query algorithm, kc = 10, were used to develop the TMAP graphs. In previous studies, we used TMAP to describe the molecular diversity of natural products such as COCONUT [2], BIOFACQUIM [15], and a focused library of HIV-1 protease inhibitors using natural fragments from COCONUT [37].

3. Results and discussion

3.1. Natural-product-likeness

Table 1 summarizes the descriptive statistics for NPL scores calculated for BIOFACQUIM, UNPD, and COCONUT. The COCONUT, UNPD, and BIOFACQUIM had NPL score values in the range of [−3.53,4.08], [−2.51,4.08], and [−0.36,3.87], respectively. As anticipated, natural product compounds from COCONUT, UNPD, and BIOFACQUIM had NPL scores *ca.* 4. Only 25% of COCONUT´s compounds (min=−3.53, Q1=−0.32) had NPL scores *ca.* −3.53, and 25% of UNPD´s compounds (min=−2.51, Q1=1.14) had NPL close to −2.51 meaning that COCONUT had compounds with more chemical structures similar to synthetic compounds regarding the UNPD because of NPL scores *ca.* −5 is associated with compounds derived from the synthetic origin [24]. In general, natural products (COCONUT, UNPD, BIOFACQUIM) had NPL values closer to 5.

3.2. Molecular diversity

Three subsets of UNPD (A-C) were generated using the MaxMin algorithm, as described in Section 2.4. UNPD-A, UNPD-B, and UNPD-C had 15,000, 7,500, and 5,000 compounds, respectively. Then, the new subsets were curated, as described in Section 2.2. Table 2 shows that six, three, two, and one duplicated compounds were found in UNPD-A, UNPD-B, and UNPD-C, respectively (between 0.04% and 0.02%). Fig. 2 shows the cumulative distribution functions (CDF) of the pairwise Tanimoto similarity using MACCS keys (166-bits), ECFP4, and ECFP6, as molecular representations. The UNPD subsets (A-C) are represented in continous line as UNPD-A (pink), UNPD-B (green), and UNPD-C (red); UNPD (blue); BIOFACQUIM (yellow); and DNMT1 (cyan). The CDF with ECFP4 and ECFP6 shows that UNPD subsets (A-C) were the most diverse data sets (median=0.09); followed by the UNPD (median=0.1), BIOFACQUIM (median=0.12), and DNMT1 (median=0.12). The CDF with MACCS keys shows that UNPD subsets (A-C) were the most diverse (median=0.3) and UNPD (median=0.4), followed by DNMT1 (median=0.4) and BIOFACQUIM (median=0.43). Likewise, Table 2 shows that UNPD subsets were more diverse than UNPD, DIFACQUIM, and DNMT1. The UNPD-A was the most di-

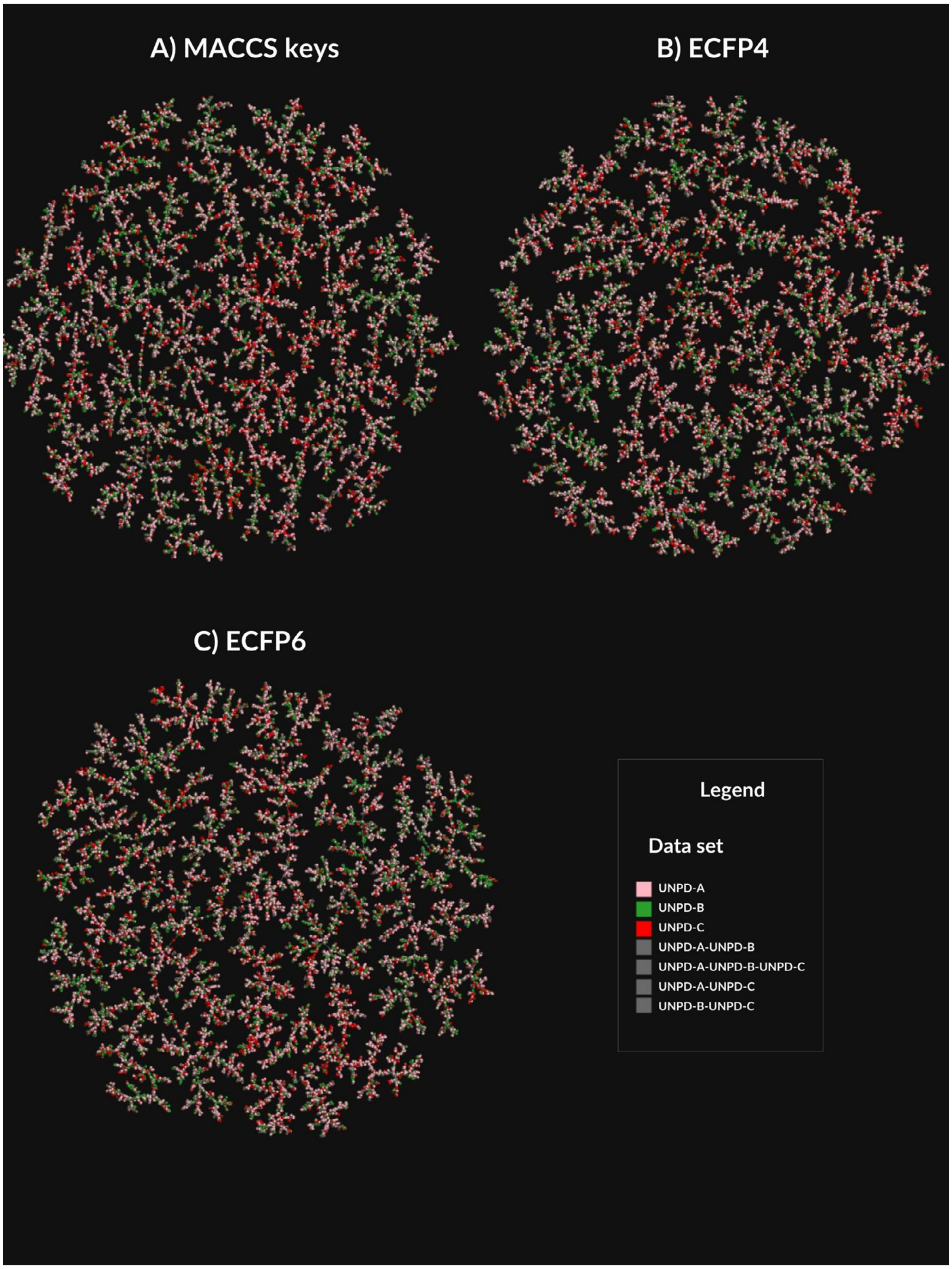


Fig. 4. Chemical multiverse visualization of UNDP subsets using TMAPs and MACCS keys (166-bits), ECFP4 and ECFP6, as molecular representations. The data sets are represented as UNPD-A (pink), UNPD-B (green), and UNPD-C (red). The overlap between natural product subsets is represented in gray.

verse dataset and had the lowest median similarity value (0.091, 0.077) calculated with ECFP4 and ECFP6, respectively, followed by the UNPD-B (0.094, 0.08), and UNPD-C (0.092, 0.079); and the reference databases: UNPD (0.111, 0.094), BIOFACQUIM (0.119, 0.099) and DNMT1 (0.119, 0.1). The structural diversity calculated with MACCS keys (Table 2 and Fig. 2) indicated that UNPD subsets had the same trend, being the UNPD-A the most diverse (median=0.341) followed by UNPD-B (median=0.346), and UNPD-C (median=0.356); followed by UNPD (median=0.43), BIOFACQUIM (median=0.447), and DNMT1 (median=0.417). For UNPD, the data set generated from a larger number of subsets (thirty *versus* fifteen or ten subsets) was the most diverse, *i.e.*, UNPD-A according to the ECFP4, ECFP6, and MACCS keys fingerprints.

3.3. *Properties of pharmaceutical relevance*

The natural products subsets were further characterized by means of the distribution of six properties of pharmaceutical relevance: hydrogen bond donors (HBD), hydrogen bond acceptors (HBA), topological polar surface area (TPSA), number of rotatable bonds (RB), molecular weight (MW), and partition coefficient octanol/water (LogP). Fig. 3 shows box-whisker plots of the distribution of the property values calculated for the three UNPD subsets and the entire UNPD (UNPD-A (pink), UNPD-B (green), UNPD-C (red), and UNPD (blue)). Tables 3-5 summarize the descriptive statistics. Fig. 3 and Tables 3-5 show that 75% of compounds of UNPD subsets and UNPD had the same range of LogP values, namely UNPD-A: LogP<=4.32, UNPD-

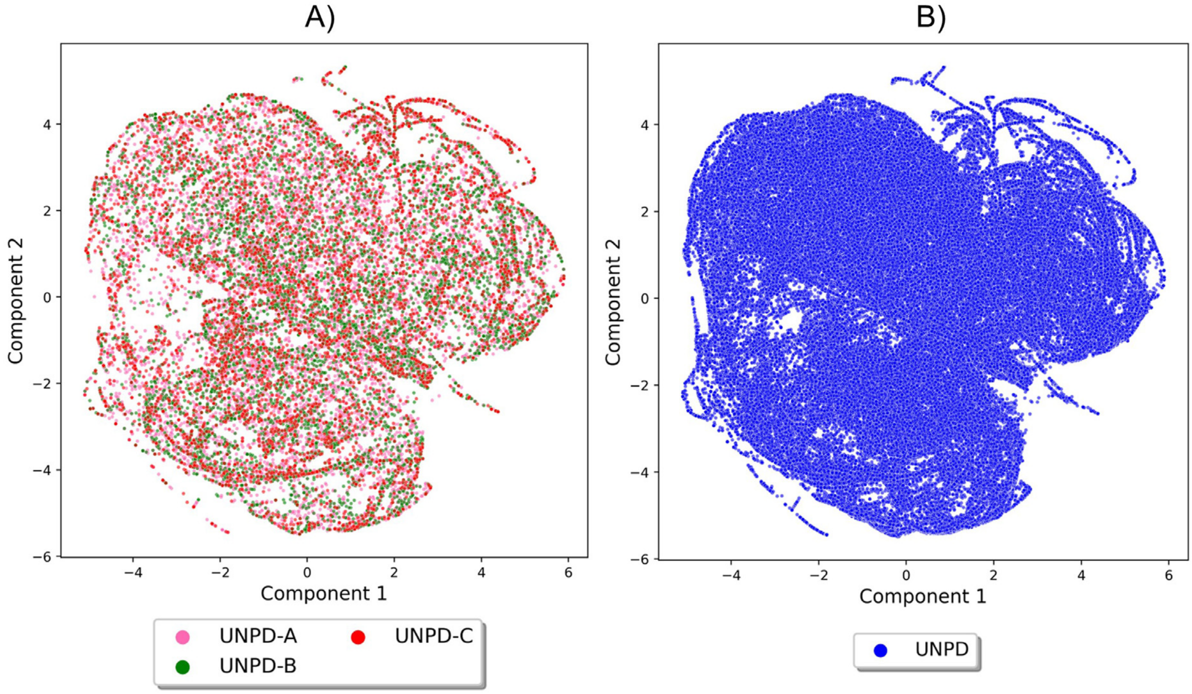


Fig. 5. Chemical space visualization of UNPD and UNPD subsets using t-SNE based on the properties: hydrogen bond donors, hydrogen bond acceptors, topological polar surface area, number of rotatable bonds, molecular weight, and partition coefficient octanol/water. The datasets are represented in scatter points as UNPD-A (pink), UNPD-B (green), UNPD-C (red), and UNPD (blue). Panel A shows the UNPD subsets (A-C), and panel B shows the entire UNPD.

B: LogP<=4.24, UNPD-C: LogP<=4.45, and UNPD: LogP<=4.58. Regarding MW and TPSA, 75% of UNPD's compounds (TPSA<=145.91 and MW<=534.66) were more diverse than 75% of UNPD subsets (A-C): UNPD-A's compounds (TPSA<=112.05, MW<=445.60); UNPD-B's compounds (TPSA<=116.20, MW<=444.48), and UNPD-C's compounds (TPSA<=116.61, MW<=466.74). Regarding RB, UNPD's compounds (RB<=8.0) and UNPD-B's compounds (RB<=7.0) were more diverse than UNPD-A's compounds (RB<=6.0) and UNPD-C's compounds (RB<=6.0). Regarding HBA and HBD, UNPD's compounds (HBD<=4.0, HBA<=9.0) and UNPD-C's compounds (HBD<=4.0, HBA<=8.0) were more diverse than UNPD-A (HBD<=3.0, HBA<=7.0) and UNPD-B (HBD<=4.0, HBA<=7.0). Overall, the UNPD-C dataset was the most diverse in terms of the properties of pharmaceutical relevance after the entire UNPD.

3.4. Visualization of the chemical space and chemical multiverse

As described in Section 2.6, a chemical multiverse is conceptualized as a group of molecular representations that each describe a compound dataset (in contrast to a chemical space that is defined by only one set of descriptors). Fig. 4 shows a visual representation of the chemical multiverse of the UNPD datasets using TMAPs and MACCS keys (166-bits), ECFP4, and ECFP6 as molecular representations. In this study, the chemical multiverse is comprised of three molecular fingerprints: one based on structural keys, MACCS keys (Fig. 4A); and the hashed molecular fingerprint, ECFP4 and ECFP6 (Figs. 4B and C). The data sets are represented in scatter points with different colors as UNPD-A (pink), UNPD-B (green), and UNPD-C (red). The overlap between different data sets is depicted in gray. The TMAP generated with MACCS keys shows fewer clusters with more compounds than ECFP4 and ECFP6. For natural products, the TMAP generated with ECFP6 shows that the compounds of the UNPD subsets are more evenly distributed with respect to the TMAPs constructed with ECFP4 and MACCS keys. The chemical multiverse represented by ECFP6 is more accurate in describing the structural diversity than ECFP4 because ECFP6 encodes molecular fragments in more detail than ECFP4 [35] and thus has an impact on quantifying structural diversity.

We also visualize the chemical space of the datasets using t-SNE based on the six physicochemical properties of pharmaceutical interest discussed in Section 3.3. Fig. 5A shows the chemical space of UNPD subsets (A-C). Each data point represents a compound: UNPD-A (pink), UNPD-B (green), and UNPD-C (red). As reference, Fig. 5B depicts a visualization of the chemical space of the entire UNPD (blue data points). t-SNE in Fig. 5A shows that the three UNPD subsets overlap with diverse regions of the chemical space of UNPD (Fig. 5B). UNPD-C was the most diverse regarding the six physicochemical properties of pharmaceutical interest, as described before in Section 3.3.

3.5. Applications of selection of subsets in drug discovery

The herein diverse subsets derived from natural products annotated with chirality information can be used in several different ways. For example, they can be used as reference sets for diversity analysis and coverage of chemical space of other compound libraries. For instance, Vivek-Ananth et al. recently reported a comparative analysis of more than 1800 secondary metabolites of medicinal fungi [38]. In that work, the authors included nine reference databases including natural products datasets. The same research group has reported and characterized an extensive database of Indian Medicinal Plants, Phytochemistry, and Therapeutics [39]. Having diverse subsets from UNDP included herein would further expand the outcome of such diversity studies. Other applications of the diverse subsets include its use in the development of generative models, such as *de novo* design. This approach generates new chemical entities with the properties desired, and recently uses algorithms of deep learning [16]. Deep learning algorithms require a large number of compounds, but it means more computer resources. The rational design of drugs involves quality data [40] to develop good models with the best predictions [21], and computer scientists advise the use of algorithms that can detect meaningful patterns in small data sets characteristics of the early stage of drug discovery can generate prospective studies [41]. For instance, a first approach to *de novo* design is to start from small data sets of compounds with diverse structures and diverse properties of pharmaceutical relevance herein generated; and add to the distinctive structural complexity and diversity of the natural products as a larger fraction of sp^3 carbon atoms and chiral centers [3,4].

4. Conclusions

We report the selection and characterization of the three subsets with the most diverse compounds from UNPD using the MaxMin algorithm. Three subsets with 14,994, 7497, and 4998 compounds selected from the UNPD contain the most structurally diverse natural products. Unlike compounds in the COCONUT database, molecules in UNPD are annotated with chirality. The structural diversity of compounds is not affected by the number of subsets derived from the original database from which a new database is generated, and an analysis of the chemical multiverse supports that UNPD subsets contain the most diverse molecules. During the study, we also concluded that the visualization of the chemical space described with ECFP6 is more accurate to describe the structural diversity of compounds compared with ECFP4 and MACCS keys (166 bits). The natural product subsets had a large diversity of chemical compounds with different structural features and properties of pharmaceutical relevance. The NPL score supports that the chemical structures of natural products are very different and diverse as defined by the threshold of the NPL score, and as expected, natural products (COCONUT, UNPD, and BIOFACQUIM) had NPL scores values close to 5.

A significant perspective of this work is that the natural product subsets derived from the UNPD can be used to develop generative models that use deep learning algorithms and require the most diverse compounds, such as *de novo* design. The natural products subsets can also be used to develop predictive models; for virtual screening; and reference databases for evaluating the structural diversity or similarity to a specific subset, among other applications. The public availability of the natural product subsets can save costly computational resources for research groups with limited accessibility to supercomputer means.

Supplementary material

The MaxMin algorithm and structural diversity implemented in Python language, the interactive TMAPs, and the three subsets generated from UNPD with 14,994, 7,497, and 4,998 compounds with stereochemical information, are publicly available at GitHub: https://github.com/DIFACQUIM/Natural-products-subsets-generation.

Funding

Authors thank the Dirección General de Cómputo y de Tecnologías de Información y Comunicación (DGTIC), UNAM, for the computational resources to use Miztli supercomputer at UNAM under project LANCAD-UNAM-DGTIC-335; and the innovation space UNAM-HUAWEI the computational resources to use their supercomputer under project No. 7 "*Desarrollo y aplicación de algoritmos de inteligencia artificial para el diseño de fármacos aplicables al tratamiento de diabetes mellitus y cáncer*". We also thank the funding of DGAPA, UNAM, Programa de Apoyo a Proyectos de Investigación e Innovación Tecnológica (PAPIIT), grant No. IN201321.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

All data is public at GitHub.

Acknowledgments

A.L.C.-H. is thankful to CONACyT, Mexico, for the Ph.D. scholarship number 847870.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.ailsci.2023.100066.

References

[1] Chávez-Hernández AL, Sánchez-Cruz N, Medina-Franco JL. A fragment library of natural products and its comparative chemoinformatic characterization. Mol Inform 2020;39:e2000050.
[2] Chávez-Hernández AL, Sánchez-Cruz N, Medina-Franco JL. Fragment library of natural products and compound databases for drug discovery. Biomolecules 2020;10:1518.
[3] Grigalunas M, Brakmann S, Waldmann H. Chemical evolution of natural product structure. J Am Chem Soc 2022;144:3314–29.
[4] Atanasov AG, Zotchev SB, Dirsch VM. International Natural product sciences taskforce, C.T. Supuran, natural products in drug discovery: advances and opportunities. Nat Rev Drug Discov 2021;20:200–16.
[5] Sorokina M, Merseburger P, Rajan K, Yirik MA, Steinbeck C. COCONUT online: collection of Open Natural Products database. J Cheminform 2021;13:2.
[6] 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.
[7] 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.
[8] Saldívar-González FI, Valli M, Andricopulo AD, da Silva Bolzani V, Medina-Franco JL. Chemical space and diversity of the NuBBE database: a chemoinformatic characterization. J Chem Inf Model 2019;59:74–85.
[9] Costa RPO, Lucena LF, Silva LMA, Zocolo GJ, Herrera-Acevedo C, Scotti L, Da–Costa FB, Ionov N, Poroikov V, Muratov EN, Scotti MT. The SistematX web portal of natural products: an update. J Chem Inf Model 2021;61:2516–22.
[10] Scotti MT, Herrera-Acevedo C, Oliveira TB, Costa RPO, de O Santos SYK, Rodrigues RP, Scotti L, Da-Costa FB. SistematX, an online web-based cheminformatics tool for data management of secondary metabolites. Molecules 2018;23:103.
[11] Olmedo DA, González-Medina M, Gupta MP, Medina-Franco JL. Cheminformatic characterization of natural products from Panama. Mol Divers 2017;21:779–89.
[12] Olmedo DA, Medina-Franco JL. Chemoinformatic approach: the case of natural products of panama. Cheminformatics and its applications. IntechOpen; 2020.
[13] H.L. Barazorda-Ccahuana, L.G. Ranilla, M.A. Candia-Puma, E.G. Cárcamo-Rodriguez, A.E. Centeno-Lopez, G.D. Del-Carpio, J.L. Medina-Franco, M.A. Chávez-Fumagalli, PeruNPDB: the Peruvian Natural Products Database for in silico drug screening, bioRxiv. (2023) 2023.01.15.524152. 10.1101/2023.01.15.524152.
[14] Pilón-Jiménez BA, Saldívar-González FI, Díaz-Eufracio BI, Medina-Franco JL. BIOFACQUIM: a Mexican compound database of natural products. Biomolecules 2019;9:31.
[15] N. Sánchez-Cruz, B.A. Pilón-Jiménez, J.L. Medina-Franco, Functional group and diversity analysis of BIOFACQUIM: a Mexican natural product database, F1000Res. 8 (2019) (Chem Inf Sci) 2071.
[16] Palazzesi F, Pozzan A. Deep learning applied to ligand-based de novo drug design. Methods Mol Biol 2022;2390:273–99.
[17] Hessler G, Baringhaus K-H. Artificial intelligence in drug design. Molecules 2018;23:2520.
[18] Sousa T, Correia J, Pereira V, Rocha M. Generative deep learning for targeted compound design. J Chem Inf Model 2021;61:5343–61.
[19] Jing Y, Bian Y, Hu Z, Wang L, Xie X-Q. Deep Learning for drug design: an artificial intelligence paradigm for drug discovery in the big data era. AAPS J 2018;20:58.
[20] Miljković F, Rodríguez-Pérez R, Bajorath J. Impact of artificial intelligence on compound discovery, design, and synthesis. ACS Omega 2021;6:33293–9.
[21] Schneider P, Walters WP, Plowright AT, Sieroka N, Listgarten J, Goodnow Jr RA, Fisher J, Jansen JM, Duca JS, Rush TS, Zentgraf M, Hill JE, Krutoholow E, Kohler M, Blaney J, Funatsu K, Luebkemann C, Schneider G. Rethinking drug design in the artificial intelligence era. Nat Rev Drug Discov 2020;19:353–64.
[22] Bajorath J, Chávez-Hernández AL, Duran-Frigola M, Fernández-de Gortari E, Gasteiger J, López-López E, Maggiora GM, Medina-Franco JL, Méndez-Lucio O, Mestres J, Miranda-Quintana RA, Oprea TI, Plisson F, Prieto-Martínez FD, Rodríguez-Pérez R, Rondón-Villarreal P, Saldívar-Gonzalez FI, Sánchez-Cruz N, Valli M. Chemoinformatics and artificial intelligence colloquium: progress and challenges in developing bioactive compounds. J Cheminform 2022;14:82.
[23] . Selecting diverse sets of compounds. In: Leach AR, Gillet VJ, editors. An introduction to chemoinformatics. Netherlands, Dordrecht: Springer; 2007. p. 119–39.
[24] Ertl P, Roggo S, Schuffenhauer A. Natural product-likeness score and its application for prioritization of compound libraries. J Chem Inf Model 2008;48:68–74.
[25] Medina-Franco JL, Chávez-Hernández AL, López-López E, Saldívar-González FI. Chemical multiverse: an expanded view of chemical space. Mol Inform 2022;41:e2200116.
[26] G. Hinton, Visualizing Data using t-SNE, (2008). https://www.jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf?fbcl (accessed February 4, 2023).

[27] Probst D, Reymond J-L. Visualization of very large high-dimensional data sets as minimum spanning trees. J Cheminform 2020;12:12.
[28] Prado-Romero DL, Medina-Franco JL. Advances in the exploration of the epigenetic relevant chemical space. ACS Omega 2021;6:22478–86.
[29] Conery AR, Rocnik JL, Trojer P. Small molecule targeting of chromatin writers in cancer. Nat Chem Biol 2022;18:124–33.
[30] Mendez D, Gaulton A, Bento AP, Chambers J, De Veij M, Félix E, Magariños MP, Mosquera JF, Mutowo P, Nowotka M, Gordillo-Marañón M, Hunter F, Junco L, Mugumbate G, Rodriguez-Lopez M, Atkinson F, Bosc N, Radoux CJ, Segura-Cabrera A, Hersey A, Leach AR. ChEMBL: towards direct deposition of bioassay data. Nucleic Acids Res 2019;47:D930–40.
[31] Davies M, Nowotka M, Papadatos G, Dedman N, Gaulton A, Atkinson F, Bellis L, Overington JP. ChEMBL web services: streamlining access to drug discovery data and utilities. Nucleic Acids Res 2015;43:W612–20.
[32] Weininger D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. J Chem Inf Comput Sci 1988;28:31–6.
[33] RDKit, (n.d.). https://www.rdkit.org (accessed 08 January 08 2022).
[34] MolVS, (n.d.). https://molvs.readthedocs.io/en/latest/(accessed 08 accessed January 2022).
[35] Rogers D, Hahn M. Extended-connectivity fingerprints. J Chem Inf Model 2010;50:742–54.
[36] Durant JL, Leland BA, Henry DR, Nourse JG. Reoptimization of MDL keys for use in drug discovery. J Chem Inf Comput Sci 2002;42:1273–80.
[37] Chávez-Hernández AL, Juárez-Mercado KE, Saldívar-González FI, Medina-Franco JL. Towards the de novo design of HIV-1 protease inhibitors based on natural products. Biomolecules 2021;11:1805.
[38] Vivek-Ananth RP, Sahoo AK, Baskaran SP, Samal A. Scaffold and structural diversity of the secondary metabolite space of medicinal fungi. ACS Omega 2023;8:3102–13.
[39] Mohanraj K, Karthikeyan BS, Vivek-Ananth RP, Chand RPB, Aparna SR, Mangalapandi P, Samal A. IMPPAT: a curated database of Indian medicinal plants, phytochemistry and therapeutics. Sci Rep 2018;8:4329.
[40] Perron Q, da Silva VBR, Atwood B, Gaston-Mathé Y. Key points to succeed in Artificial Intelligence drug discovery projects. Chem Int 2022;44:19–21.
[41] Schneider G, Clark DE. Automated de novo drug design: are we nearly there yet? Angew Chem Int Ed Engl 2019;58:10792–803.

frontiers | Frontiers in Drug Discovery

TYPE Review
PUBLISHED 21 June 2023
DOI 10.3389/fddsv.2023.1222655

Check for updates

OPEN ACCESS

EDITED BY
Ho Leung Ng,
Atomwise Inc., United States

REVIEWED BY
Rodolpho C. Braga,
InsilicAll, Brazil
Jeremy J. Yang,
University of New Mexico, United States

*CORRESPONDENCE
José L. Medina-Franco,
✉ medinajl@unam.mx

RECEIVED 15 May 2023
ACCEPTED 12 June 2023
PUBLISHED 21 June 2023

CITATION
Chávez-Hernández AL, López-López E and Medina-Franco JL (2023), Yin-yang in drug discovery: rethinking *de novo* design and development of predictive models.
Front. Drug Discov. 3:1222655.
doi: 10.3389/fddsv.2023.1222655

Yin-yang in drug discovery: rethinking *de novo* design and development of predictive models

Ana L. Chávez-Hernández[1], Edgar López-López[1,2] and José L. Medina-Franco[1]*

[1]Department of Pharmacy, DIFACQUIM Research Group, School of Chemistry, Universidad Nacional Autónoma de México, Avenida Universidad, Mexico City, Mexico, [2]Department of Chemistry and Graduate Program in Pharmacology, Center for Research and Advanced Studies of the National Polytechnic Institute, Mexico City, Mexico

Chemical and biological data are the cornerstone of modern drug discovery programs. Finding qualitative yet better quantitative relationships between chemical structures and biological activity has been long pursued in medicinal chemistry and drug discovery. With the rapid increase and deployment of the predictive machine and deep learning methods, as well as the renewed interest in the *de novo* design of compound libraries to enlarge the medicinally relevant chemical space, the balance between quantity and quality of data are becoming a central point in the discussion of the type of data sets needed. Although there is a general notion that the more data, the better, it is also true that its quality is crucial despite the size of the data itself. Furthermore, the active versus inactive compounds ratio balance is also a major consideration. This review discusses the most common public data sets currently used as benchmarks to develop predictive and classification models used in *de novo* design. We point out the need to continue disclosing inactive compounds and negative data in peer-reviewed publications and public repositories and promote the balance between the positive (Yang) and negative (Yin) bioactivity data. We emphasize the importance of reconsidering drug discovery initiatives regarding both the utilization and classification of data.

KEYWORDS

big data, chemoinformatics, chemical libraries, data quality, *de novo* design, drug discovery, machine learning, negative results

Abbreviations: 2D/3D, two-dimensional/three-dimensional; ADMET, absorption, distribution, metabolism, excretion, and toxicity; AI, artificial intelligence; CADD, computer-aided drug design; COCONUT, Collection of Open NatUral ProdUcTs; DNN, deep neural networks; HBA, hydrogen bond acceptors; HBD, hydrogen bond donors; IMPPAT, A curated database of Indian Medicinal Plants, Phytochemistry And Therapeutics; MW, molecular weight; LBDD, ligand-based drug design; log P, octanol-water partition coefficient; NCE, new chemical entities; NIH(US), National Institutes of Health; PAINS, pan-assay interference compounds; Peru NPDB, Peruvian Natural Products Database; QSAR, quantitative structure-activity relationships; REAL, Enamine's REadily AccessibLe; RNNs, recurrent neural networks; SBDD, structure-based drug design; TCM, Traditional Chinese Medicine; TPSA, topological surface area; UNPD, Universal Natural Product Database.

1 Introduction

Data and the increasing role of predictive models, including machine and deep learning (Mouchlis et al., 2021; Bajorath et al., 2022), are the cornerstone of modern drug discovery programs (Zhang et al., 2022). The increasing use of computational methods that recently included deep learning is reducing the time and financial costs of finding drug candidates (Zhang et al., 2022). For instance, computer-aided drug design (CADD) has led to the discovery of more than seventy approved drugs (Sabe et al., 2021) including remdesivir as an emergency treatment against SARS-CoV-2 in 2021 (Dos Santos Nascimento et al., 2021).

CADD methods are typically divided into two main categories, structure-based drug design (SBDD) and ligand-based drug design (LBDD) that rely on the three-dimensional (3D) structure data available for one or more molecular targets, or the structure-activity data of ligands, respectively. Examples of deep learning applications in SBDD include AlphaFold to assist in homology modeling, and DiffDock in molecular docking. AlphaFold predicts 3D protein structures according to their amino acid sequences (Jumper et al., 2021), and DiffDock predicts the binding mode between the ligand and specific protein target (Corso et al., 2022). One of the most notable approaches in LBDD are quantitative structure-activity relationships (QSAR) (Dos Santos Nascimento et al., 2021). Current QSAR methods use machine learning and deep learning (Soares et al., 2022) that can be divided into linear methods and nonlinear methods (Patel et al., 2014; Greener et al., 2022). Linear methods include linear regression, multiple linear regression, partial least squares, and principal component analysis (Patel et al., 2014). Nonlinear methods include artificial neural networks, k-nearest neighbors, and Bayesian neural nets, to name a few examples (Patel et al., 2014; Greener et al., 2022).

Advances in deep learning models have a significant progress in molecule generation, representing a big step forward in bridging the gap between chemical entities and drug-like properties (Krishnan et al., 2021). Deep learning algorithms are currently used in the renewed interest in the *de novo* design of chemical libraries. In 2020, the successful application of deep learning in drug discovery, that included the *de novo* design using deep learning, was selected by the Massachusetts Institute of Technology Technology Review as one of the top ten breakthrough technologies (Juskalian et al., 2023).

De novo design is aimed at generating new chemical entities (NCE) with desired properties (Palazzesi and Pozzan, 2022). *De novo* design based on deep learning algorithms (Palazzesi and Pozzan, 2022) requires a large number of compounds that may demand significant computational resources. However, bioactivity data for a biological endpoint is not always sufficient. The lack of data has led to the development of new methods for compound selection and applications for deep learning algorithms are being developed (Guo M et al., 2021).

Knowledge-based drug design frequently involves quality data (Perron et al., 2022b) to develop models with useful predictions (Schneider et al., 2020). To this end, rethinking the methodologies used for drug discovery and development campaigns is crucial. The quality of data sets, decoy data sets and inactive compounds used in predictive models, and *de novo* design models need to be reviewed and discussed.

The main purpose of this manuscript is discussing the importance of quality data, decoy data sets, and the balance needed between inactive (i.e., “Yin”) and active (“Yang”) compounds currently employed in *de novo* design and developing predictive models of biological activity to generate NCE. Following up on previous studies (Schneider et al., 2020; Bajorath et al., 2022; Cherkasov, 2023), we comment on the need to rethink the way to drug design and develop campaigns. The manuscript is organized into four main sections. After this Introduction, Section 2 presents an overview of *de novo* design. Section 3 discusses the main public data sources used to develop predictive models. Section 4 discusses criteria to generate quality data sets. The last section presents a summary of conclusions and perspectives.

2 *De novo* design overview

De novo design aims to generate new chemical structures from scratch with desired predicted properties, e.g., absorption, distribution, metabolism, excretion, toxicity (ADMET), other drug-likeness properties, and biological activities (Palazzesi and Pozzan, 2022). The two main strategies for *de novo* design can be classified into SBDD and LBDD (*vide supra*) (Zhang et al., 2022). A recent example of a structured-based *de novo* design is the RELATION model that learns from the desired geometric features of protein-ligand complexes to generate new molecules (Wang et al., 2022). The generation process applies a fragment-based strategy given an initial chemical scaffold embedded in the binding site of the target protein. The pre-trained model generates molecules iteratively by sequentially adding, deleting, inserting, or replacing and linking fragments (Zhang et al., 2022).

In contrast, ligand-oriented *de novo* design focuses on the ligands themselves, thereby generating compounds with new chemical structures with novel scaffolds from active compounds while optimizing the desired properties (Xie et al., 2022). A general workflow is schematically summarized in Figure 1 which has seven main steps (Krishnan et al., 2021; Zhang et al., 2022): 1) Selecting compound data sets from public or in-house sources (further discussed in Section 3); 2) Filtering molecular data sets with desired properties such as drug-likeness. In the example of Figure 1 a data set with three subsets of compounds is represented with a star, triangle, and circle, respectively. The compounds represented with a star have drug-like properties (Lipinski et al., 2001; Veber et al., 2002); those represented with triangles comply with some of the drug-likeness properties, and those represented with circles are not compliant. Other approaches to select compounds from the data sets use molecular fingerprints (Kadurin et al., 2017) or filter compounds directly via similarity-based virtual screening instead of designing NCE from scratch (Tong et al., 2021). 3) Selecting the molecular representation as a basis to learn and represent the structures and properties of molecules, e.g., SMILES (Weininger, 1988), SELFIES (Krenn et al., 2020) or molecular graphs (Simonovsky and Komodakis, 2018). 4) Developing and validating the model for molecule generation using metrics such as the operating characteristic curve. 5) Optimizing the model by combining reinforcement learning and property prediction (Olivecrona et al., 2017). 6)

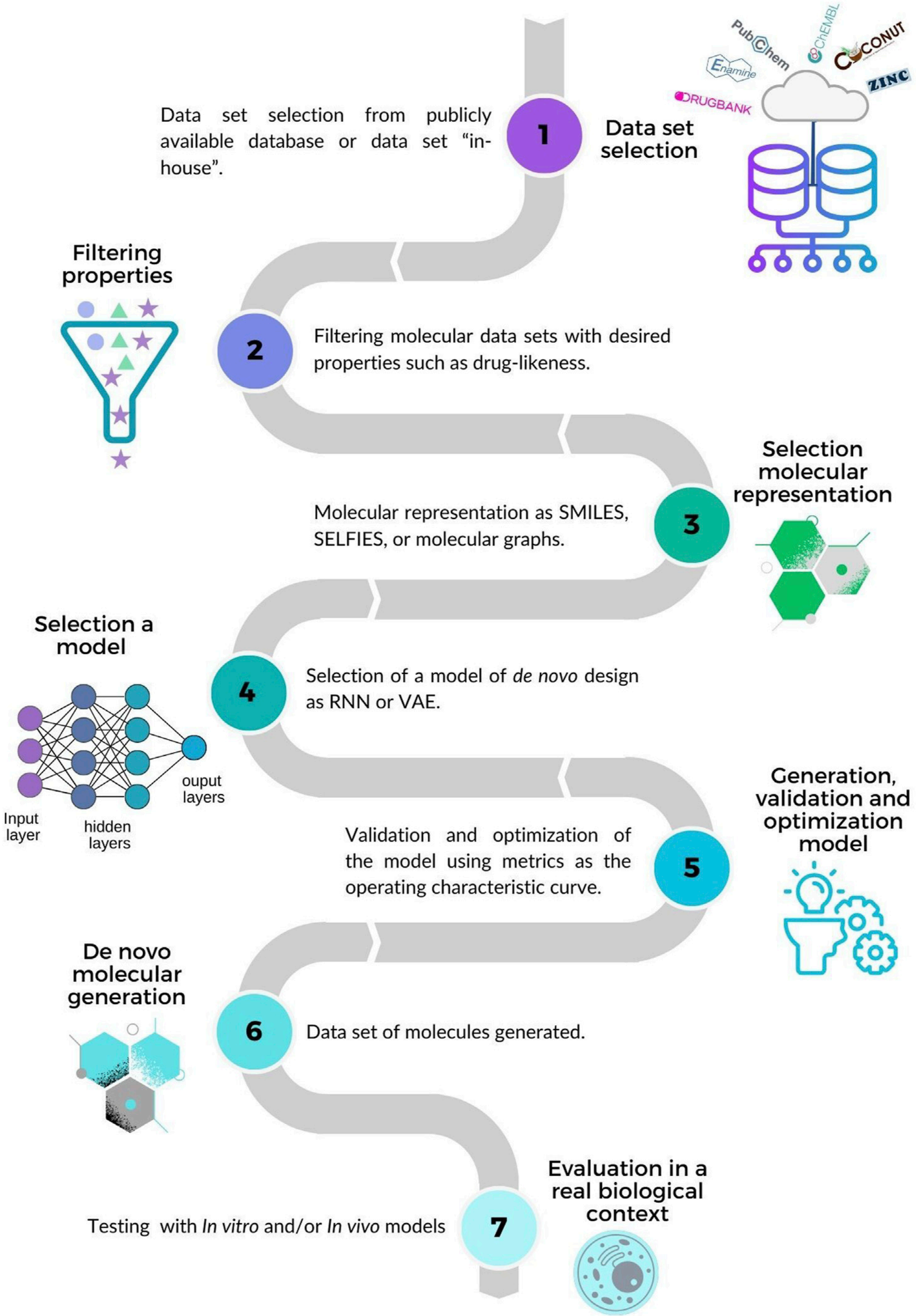

Data set selection from publicly available database or data set "in-house".
1
Data set selection
PubChem
ChEMBL
COCONUT
Enamine
DRUGBANK
ZINC
Filtering properties
2
Filtering molecular data sets with desired properties such as drug-likeness.
Selection molecular representation
Molecular representation as SMILES, SELFIES, or molecular graphs.
3
Selection a model
4
Selection of a model of de novo design as RNN or VAE.
Input layer
hidden layers
ouput layers
Generation, validation and optimization model
Validation and optimization of the model using metrics as the operating characteristic curve.
5
De novo molecular generation
6
Data set of molecules generated.
Evaluation in a real biological context
Testing with In vitro and/or In vivo models
7

FIGURE 1
Overview of ligand-based *de novo* design. 1) Selecting data sets. 2) Filtering molecular data sets with desired properties such as drug-likeness. In this example, compounds represented with stars comply with drug-likeness properties (Lipinski et al., 2001; Veber et al., 2002). 3) Choosing a molecular representation. 4) Selecting a *de novo* design model. 5) Developing, validating and optimizing the model. 6) Generating molecules *de novo*. 7) Testing the compounds in a relevant biological experiment.

Generating molecules *de novo*, 7) Assessing the biological activity of the compounds designed in relevant *in vitro* or *in vivo* models.

Deep learning, currently used in ligand-based *de novo* design, learns the probability distribution of molecular data and generates continuous or discrete latent representations for molecules with property optimization (Gómez-Bombarelli et al., 2018). The algorithms map the learned probability distribution and molecule representation into novel molecules while optimizing molecular properties (Bilodeau et al., 2022) through the tuning of hyperparameters (Perron et al., 2022a; Bender et al., 2022). Advances in deep learning are significantly advancing molecule generation, representing a big step forward in bridging the gap between chemical entities and drug-like properties (Krishnan et al., 2021).

Ligand´s properties can be optimized in two steps: 1) property-based generation, wherein models would learn the chemical space of molecules with desirable properties; and 2) novel molecules are generated within a desired property space (Bilodeau et al., 2022). Examples of ligand-based *de novo* design are deep neural networks (DNN), recurrent neural networks (RNNs) (Olivecrona et al., 2017), and variational autoencoders (VAE) (Gómez-Bombarelli et al., 2018). Olivercroma *et al.* (Olivecrona et al., 2017) proposed the REIVENT model that uses RNN for *de novo* design. They introduced a reinforcement learning method to fine-tune the pre-trained RNN so the model could generate structures with desirable properties. Recently, Blaschke et al. released REINVENT 2.0 (Blaschke et al., 2020) making the code freely accessible in Github.

Ligand-based *de novo* design using DNN (Palazzesi and Pozzan, 2022) requires a large number of compounds that demand more computational resources. The DNN architecture is prone to problems because of fitting numerous parameters. For this reason, a large training data set is needed to reduce the risk of overfitting. However, sufficient bioactivity data for a biological endpoint is not always available (Wu et al., 2018). The lack of sufficient data has led to using methods for compound selection or the development of new methods for compound selection. Altae-Tran et al. (Altae-Tran et al., 2017) demonstrated how the one-shot learning paradigm can be used to address the overfitting problem; they used DNN to transform small molecules into embedding vectors in a continuous feature space whose similarity measures are then iteratively learned. They showed that this DNN architecture offers convincing performance in many activity prediction tasks given limited amounts of training. On the other hand, computer scientists advise using algorithms that can detect meaningful patterns in small data sets, which is a typical case in the early stage of drug discovery (Schneider and Clark, 2019). For instance, an initial approach to *de novo* design is to start from small data sets of compounds with diverse structures and diverse properties of pharmaceutical relevance (Chávez-Hernández and Medina-Franco, 2023).

The availability of gold standard datasets as well as independently generated data sets are valuable in generating well-performing models (Vamathevan et al., 2019). Dissimilarity-based compound selection could be improved if one focused the selection on a structural diverse dataset (for instance derived from natural products). Some approaches proposed suggest using quality data sets using a dissimilarity-based compound selection method such as the MaxMin or MaxSum algorithms (Leach and Gilleteds, 2007). Recently, we reported the use of the MaxMin algorithm for the selection of natural product subsets (Chávez-Hernández and Medina-Franco, 2023) using the Universal Natural Product Database (UNPD) (Gu et al., 2013). In that study, the natural product subsets generated had the most diverse chemical structures with physicochemical properties of pharmaceutical interest similar to the original data set. Chemical structures in the natural product subsets were represented with SMILES encoding chirality, an important feature of natural products.

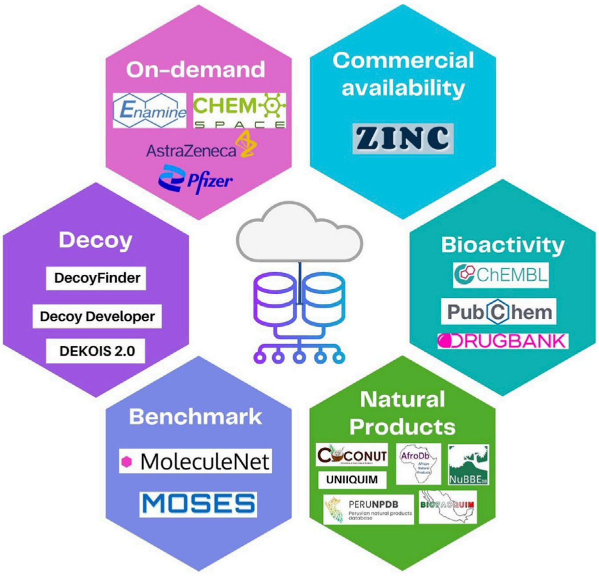


FIGURE 2
Classification of compound databases and representative examples of each one. For the discussion of this manuscript, databases are split into six main categories: on-demand, commercial availability, bioactivity, natural products, benchmark and decoys.

3 Main sources of data sets used to develop generative and predictive models

3.1 Current status of reference and benchmark datasets

The first step in *de novo* design is to select, from the vast chemical space, the appropriate subset of all possible molecules for a desired biological activity (Schneider et al., 2000). To have an idea, the size of the chemical space has been estimated at around 10^{60} small molecules and between 10^{20}–10^{24} for all molecules up to 30 atoms that comply with Lipinski’s rule-of-five (Reymond, 2015). According to Yang *et al.* compound data sets can be classified into on-demand databases, collections containing bioactivity data, compounds databases commercially available, and natural products databases (Yang et al., 2019). Herein, we include benchmark, decoy and inactive compounds data sets as others categories as illustrated in Figure 2. In this figure, on-demand databases are further divided into commercially available (e.g., Enamine-REAL, CHEMriya and Freedom Space) (Chemspace, 2023) and in-house (e.g., Pfizer and AstraZeneca). The figure shows examples of compound databases in other categories which are discussed in the remainder of this section.

Among the different types of chemical databases, *de novo* design employs libraries from different categories outlined in Figure 2. Specific examples are ChEMBL (Davies et al., 2015; Mendez et al., 2019), PubChem (Kim et al., 2023), DrugBank (Wishart et al., 2006; Wishart et al., 2008; Wishart et al., 2018), Enamine´s REadily AccessibLe (REAL) (Enamine, 2023), CHEMriya (CHEMriya, 2023), Freedom Space (Chemspace, 2023), ZINC-22 (Tingle

TABLE 1 Main sources of public molecular data sets used in *de novo* design.

Data sets	Category	Description	Ref.
ChEMBL	Bioactivity	Database with 2,354,965 bioactive drug-like small molecules with 2D structures and calculated properties.	Davies et al. (2015), Mendez et al. (2019)
PubChem	Bioactivity	Database at the US National Institutes of Health with 115 million compounds. It includes names, molecular formulas, structures, physical properties, and biological activities.	Kim et al. (2023)
DrugBank	Bioactivity	Version 5.1.10 contains 15,448 drug entries including 2,740 approved small molecule drugs.	Wishart et al. (2006)
ZINC-22	Commercial	Database with over 37 billion enumerated, searchable, commercially available compounds in 2D.	Tingle et al. (2023)
CHEMriya	On-demand	Database with 12 billion novel and synthetically feasible small molecules.	CHEMriya (2023)
Freedom Space (Chemspace)	On-demand	Database with 201 million molecules; 73% of its compounds comply with drug-likeness properties.	Chemspace (2023)
Enamine-REAL	On-demand	Database with 6 billion synthetic compounds that comply with drug-likeness properties.	Enamine (2023)
MoleculeNet	Benchmark	Compilation of 17 datasets with over 700,000 compounds in total used for comparison of different machine learning algorithms.	Wu et al. (2018)
MOSES	Benchmark	Dataset with 1,936,962 molecules from ZINC Clean Lead suitable for hit identification and ADMET optimization. It does have metrics to detect common issues in generative models such as overfitting or if the model does not limit to producing only a few typical molecules.	Polykovskiy et al. (2020)

et al., 2023), and MoleculeNet (Wu et al., 2018) which more details for each one are provided in Table 1 and further commented in the next sections.

3.2 On-demand databases

Early approaches to ligand-based *de novo* design involved fragment compounds into unique building blocks which could be recombined to make new molecules. A number of commercial suppliers of chemical samples offer large make-on-demand collections that can be reliably synthesized because the building blocks are available as well as the synthetic routes and methods (Warr et al., 2022; Korn et al., 2023). There are also large collections of fragments or building blocks commercially available. Examples of on-demand compound databases and suppliers are REAL (Enamine) (Enamine, 2023), CHEMriya (OTAVA) (CHEMriya, 2023), and Freedom Space (Chemspace) (Chemspace, 2023) (Table 1). REAL database (Enamine, 2023) comprises over 6 billion molecules that comply with the traditional drug-likeness criteria. CHEMriya (CHEMriya, 2023) contains 12 billion novel and synthetically feasible small molecules whose molecules are not explicitly listed in the public domain. Freedom Space (Chemspace, 2023) contains 201 million molecules and 73% of its compounds are drug-like (as assessed with the “rule of five”). Examples of on-demand in-house databases from the pharmaceutical industry are 10^{15} compounds of AZ Space (AstraZeneca) (Grebner, 2022), 10^{19} compounds of JFS (Johnson & Johnson) (Warr, 2021), 10^{18} compounds of PGVL (Pfizer) (Hu et al., 2012), 10^{17} compounds BICLAIM (Boehringer Ingelheim) (Korn et al., 2023), and 10^{20} compounds MASSIV (Merck/EMD) (Korn et al., 2023).

3.3 Commercially available databases

One of the largest and long-standing compendiums of commercially available compounds in ZINC. The most recent version, ZINC-22 (Tingle et al., 2023) contains over 37 billion enumerated, searchable, commercially available compounds in 2D. Over 4.5 billion have been built in biologically relevant ready-to-dock 3D formats (Tingle et al., 2023). Some examples of *de novo* design using ZINC include the design of inhibitors of DDR1 (discoidin domain receptor 1, a kinase target implicated in fibrosis and other diseases) (Zhavoronkov et al., 2019) and compounds with activity towards the dopamine receptor D2 (Liu et al., 2019; Maziarka et al., 2020).

3.4 Bioactivity databases

De novo design based on deep learning algorithms frequently use PubChem, ChEMBL, and DrugBank to select subsets of compounds focused on a biological target or biological endpoint as the design of ligands (Li et al., 2018; Li et al., 2022; Liu et al., 2019). PubChem (Kim et al., 2023) is a freely accessible database from the US National Institutes of Health (NIH) with over 115 million compounds. At the time of writing, the most recent version release of ChEMBL is 32 (Davies et al., 2015; Mendez et al., 2019) and contains 2,354,965 compounds bioactive drug-like small molecules with 2D structures and calculated properties. DrugBank (Wishart et al., 2006; Wishart et al., 2008; Wishart et al., 2018) version 5.1.10 (released 2023-01-04) contains 15,448 drug entries including 2,740 approved small molecule drugs, 1,577 approved biologics (proteins, peptides, vaccines, and allergens), 134 nutraceuticals and over 6,717 experimental (discovery-phase) drugs. Some applications include the *de novo* design of SARS-CoV-2 Mpro inhibitors (Li et al., 2022), the design of ligands against the adenosine receptor ($A_{2A}R$) (Liu et al., 2019), and the generation of compounds analogs to celecoxib (used to manage symptoms of various types of arthritis pain and reduce precancerous polyps in the colon) (Li et al., 2018; DRUGBANK, 2023).

TABLE 2 Examples of natural product databases in the public domain.

Data sets	Description	Ref.
COCONUT	Extensive database with 406,076 unique structures.	Sorokina et al. (2021)
SuperNatural 3.0	A database with 449 058 natural compounds and derivatives. It includes chemical structure, physicochemical information, information on pathways, mechanism of action, toxicity, vendor information if available, drug-like chemical space prediction for several diseases such as antiviral, antibacterial, antimalarial, anticancer, and target-specific cells.	Gallo et al. (2023)
UNPD	Second-largest database with around 229,000 natural products that contain chirality information.	Gu et al. (2013)
TCM Database@Taiwan	Database with more than 20,000 pure compounds isolated from 453 TCM ingredients.	Chen (2011)
IMPPAT	Database of 9,596 phytochemicals from 1,742 Indian medicinal plants.	Mohanraj et al. (2018)
AfroDB	Compound collection with more than 1,000 compounds from African medicinal plants.	Ntie-Kang et al. (2013)
NuBBE$_{DB}$	Brazilian database with 2,223 natural products encoding as SMILES, InChI, and InChIKey strings, Ro5 and Veber descriptors, source, therapeutic effect, and reference.	Valli et al. (2013), Pilon et al. (2017), Saldívar-González et al. (2019)
SistematX	Brazilian database with 9,514 unique secondary metabolites encoding as SMILES, InChI, and InChIKey strings, and include physicochemical drug-like descriptors, predicted biological activities, and reference.	Scotti et al. (2018), Costa et al. (2021)
CIFPMA	Database developed at the University of Panama. It contains natural products that have been tested in over 25 *in vitro* and *in vivo* bioassays, for different therapeutic targets.	Olmedo et al. (2017), Olmedo and Medina-Franco (2020)
PeruNPDB	Peru database developed at the Catholic University of Santa Maria. The current version has 280 natural products from animals and plants.	Barazorda-Ccahuana et al. (2023)
BIOFACQUIM	Mexican database with structures of 531 natural products isolated and characterized at UNAM and other Mexican institutions.	Pilón-Jiménez et al. (2019), Sánchez-Cruz et al. (2019)
UNIIQUIM	Mexican database with 1,112 plant natural products mostly isolated and characterized at the Institute of Chemistry of the UNAM.	*UNIIQUIM* (2015)

Other libraries of natural products with an emphasis on commercial availability are listed on the NIH website (NIH, 2023).

3.5 Natural product databases

Natural product databases (Gómez-García and Medina-Franco, 2022; Saldívar-González et al., 2022) are important in drug discovery. From drugs approved by 2020 about 23% are natural products or derivatives (Newman and Cragg, 2020). Natural products have a diversity of privileged scaffolds (Atanasov et al., 2021; Grigalunas et al., 2022) and molecular fragments (Chávez-Hernández et al., 2020a; Chávez-Hernández et al., 2020b) that depend on the particular source (Medina-Franco et al., 2022b); a diversity of chiral centers; and a larger fraction of sp^3 carbon atoms and functional groups (Atanasov et al., 2021; Grigalunas et al., 2022).

Privileged structures were defined by Evans et al. (Evans et al., 1988) as *chemical structures capable of providing useful ligands for more than one receptor judicious modification of such structures could be a viable alternative in the search for new receptor agonists and antagonists*. Schneider and Schneider (2017) define a privileged structure as a chemical structure that may be considered to possess geometries suitable for decoration with side chains, such that the resulting products bind to different target proteins or a ligand that potently interacts with one (selective binder) or many target receptors (promiscuous binder). To this end, natural products are used in the development of pseudo-natural products, compounds that are generated through a *de novo* combination of natural product fragments, allowing the exploration of uncharted areas of biologically relevant chemical space that are different from the chemical space covered by the compounds from which they are derived (Grigalunas et al., 2022).

Representative natural product datasets that can be used in *de novo* design are Collection of Open NatUral ProdUcTs (COCONUT) (Sorokina et al., 2021), SuperNatural 3.0 (Gallo et al., 2023), UNPD (Gu et al., 2013), NuBBE$_{DB}$ (Pilon et al., 2017; Saldívar-González et al., 2019), SistematX (Scotti et al., 2018; Costa et al., 2021), CIFPMA (Olmedo et al., 2017; Olmedo and Medina-Franco, 2020), PeruNPDB (Barazorda-Ccahuana et al., 2023), BIOFACQUIM (Pilón-Jiménez et al., 2019; Sánchez-Cruz et al., 2019), UNIIQUIM(UNIIQUIM, 2015), and are summarized in Table 2.

SuperNatural 3.0, COCONUT and UNPD are the most extensive natural product databases. SuperNatural 3.0 (Gallo et al., 2023) is arguably the most extensive natural product database with 449,058 natural compounds and derivatives; followed by COCONUT (Sorokina et al., 2021) with 406,076 unique structures (no encoding stereochemistry) and UNPD (Gu et al., 2013) with 197,201 natural products that contain chirality information.

Several public natural products databases compile the compounds isolated and characterized from a geographical region or the country of origin as China, India and Africa. For instance, Chinese Traditional Medicine (TCM) Database@Taiwan (Chen, 2011) is a non-commercial TCM database with more than 20,000 pure compounds isolated from 453 TCM ingredients; A curated database of Indian Medicinal Plants, Phytochemistry And Therapeutics (IMPPAT) (Mohanraj et al., 2018) is a manually curated database of 9,596 phytochemicals from 1,742 Indian medicinal plants; and AfroDB (Ntie-Kang et al., 2013) with more than 1,000 small and structural diversity compounds from African medicinal plants.

Representative Latin American databases (Gómez-García and Medina-Franco, 2022) are NuBBE$_{DB}$ (Pilon et al., 2017; Saldívar-González et al., 2019), SistematX (Scotti et al., 2018; Costa et al., 2021) from Brazil; CIFPMA (Olmedo et al., 2017; Olmedo and Medina-Franco, 2020) from Panama; PeruNPDB (Barazorda-Ccahuana et al., 2023) from Peru; BIOFACQUIM (Pilón-Jiménez et al., 2019; Sánchez-Cruz et al., 2019) and UNIIQUIM (*UNIIQUIM*, 2015) from Mexico. The current version of NuBBE$_{DB}$ (Pilon et al., 2017; Saldívar-González et al., 2019) contains 2,223 natural products encoding as linear notations as SMILES. SistematX (Scotti et al., 2018; Costa et al., 2021) has 9,514 unique secondary metabolites arising from 20,934 botanical occurrences across five families. Other natural product collections from Latin America are CIFPMA, the Natural Products Database from the University of Panama, Republic of Panama (Olmedo et al., 2017; Olmedo and Medina-Franco, 2020) with 354 compounds. CIFPMA molecules have the potential to show target selectivity in biochemical assays and are useful molecules to identify reference compounds for virtual screening campaigns (Olmedo et al., 2017; Olmedo and Medina-Franco, 2020). The first version of the Peruvian Natural Products Database (PeruNPDB) had 280 natural products isolated from plants and animal sources (Barazorda-Ccahuana et al., 2023). BIOFACQUIM (Pilón-Jiménez et al., 2019; Sánchez-Cruz et al., 2019) contains 531 natural products isolated and characterized at the School of Chemistry of the National Autonomous University of Mexico (UNAM) and other Mexican institutions. UNIIQUIM (UNIIQUIM, 2015) with 1,112 plant natural products mostly isolated and characterized at the Institute of Chemistry of the UNAM.

3.6 Benchmark databases

The development of reliable machine learning algorithms has been limited due to the lack of standard benchmark datasets to compare the efficacy of the methods proposed (Jain and Nicholls, 2008). Furthermore, machine learning in chemistry compared with other areas such as computer speech and vision has a main disadvantage, the data recovery (Wu et al., 2018; Guo et al., 2022), because of measuring chemical properties often requires specialized instruments; as a result, datasets with experimentally determined results are small and often not sufficiently large to cover the high-demanding needs of machine-learning tasks (Wu et al., 2018). Another challenge is data splitting (the way in which datasets are split into training data and testing data). Some are random selection and rational selection. The former is randomly extracting a compound's fraction from the data set. In contrast to rational selection, training and testing are selected from the same clusters of compounds. Random selection is common in machine learning but is often not correct for chemical data (Sheridan, 2013). In response to these challenges, standard benchmark data sets are being developed to evaluate *de novo* design protocols [(Wu et al., 2018; Brown et al., 2019; Polykovskiy et al., 2020). One example is MoleculeNet (Wu et al., 2018), a large-scale data set built upon multiple public databases. MoleculeNet is organized into regression and classification datasets and has over 700,000 compounds tested on a range of different properties subdivided into four categories (quantum mechanics, physical chemistry, biophysics, and physiology). Another example is the Molecular Sets (MOSES) (Polykovskiy et al., 2020) that contains 1,936,962 molecules (split into training, testing and scaffold datasets) and a set of metrics to evaluate the quality and diversity of generated structures. Metrics detect common issues in generative models such as overfitting or if the *de novo* design model just generates fairly common (not novel) structures (Brown et al., 2019; Polykovskiy et al., 2020). The developers of MOSES implemented and compared several molecular generation models and suggested using the results as reference points for further advancements in generative chemistry research.

3.7 Current decoy data sets and inactive compounds

Accuracy of predictive models depends on data quality and quantity. Also, the balance between active and inactive compounds is important, which remains an issue to resolve. Historically, the publication of active compounds in a given assay or with a particular endpoint has been prioritized over inactive molecules. For example, a recent comprehensive analysis of published screening bioactivity data shows that in ChEMBL V.29 (release in 2022) there is a large number of active compounds (*ca. 71%*) with respect to the inactive ones (*ca.* 31%); contrary to what it would be expected (López-López et al., 2022). These results highlight the relevance of changing the mindset about the importance and utility of inactive or negative data (keeping in mind that the definition of "inactive" is subjective as it depends on the particular biological assay and the predefined threshold to deem a compound inactive).

Decoy data sets have been developed in an attempt to reduce the gap between inactive (or negative) and active compounds. Decoy molecules are assumed non-active but have high physicochemical property similarity (but not topologically) to reference compounds (Réau et al., 2018). Decoys are useful to evaluate benchmark models that were assembled in the absence of inactive compounds experimentally measured (Irwin, 2008) and can be used to enrich *de novo* design models. Table 3 summarizes examples of large databases of experimentally tested active or inactive compounds, decoy datasets, and tools to generate decoys for specific projects.

Decoy compounds have been used to describe, explore, and expand the knowledge of active molecules. For example,

TABLE 3 Examples of potential inactive and decoy resources for enriching *de novo* design models.

Datasets with active and inactive compounds	Criteria to select inactive data	Ref.
ChEMBL	Reported activity data.	Davies et al. (2015), Mendez et al. (2019)
PubChem		Kim et al. (2023)
Binding DB	Reported ligand-receptor affinity.	Chen et al. (2002)
Decoy datasets	Common decoy selection criteria	
ZINC	Compounds that share drug-like properties with the reference (active) compounds.	Tingle et al. (2023)
DUD-E		Mysinger et al. (2012)
DUD	Database with 2950 annotated ligands and 95,316 property-matched decoys for 40 targets.	Irwin (2008)
MUV	Compounds that share structural similarity with active reported compounds.	Rohrer and Baumann (2009)
DEKOIS 2.0	Compounds that share drug-like properties and structural similarity with the reference (active) compounds.	Bauer et al. (2013)
Decoy tools	Common decoy compound selection criteria	
DecoyFinder	Allows the automatic creation of datasets of compounds with physicochemical similarity and without structural similarity respect to the reference (active) compounds.	Cereto-Massagué et al. (2012)
RADER	Allows the automatic generation of datasets of compounds with physicochemical and structural similarity with respect to the reference (active) compounds.	Wang et al. (2017)
ZINC pharmer	Enables the automatic identification of compounds with pharmacophore similarity with respect to the reference (active and inactive) compounds.	Koes and Camacho (2012)
Decoy Developer	Allows the automatic generation of peptides decoys.	Shipman et al. (2019)

TABLE 4 Examples of applications of decoys in *de novo* design.

Approach	Purpose of using decoy sets	Ref.
Ligand-based	• Validation of new protocols and scoring functions based on similarity metrics and 3D shape.	(Arús-Pous et al. (2020); Awale and Reymond. (2015); Cao et al. (2020); Medina-Franco et al. (2019); Norinder et al. (2019); Papadopoulos et al. (2021); Skalic et al. (2019b); Skalic et al. (2019a); Ullanat (2020)
	• Improvement of the accuracy of AI-based models.	
	• Improvement of the accuracy of QSAR models.	
	• Enrichment of inactive “dark regions” in chemical space.	
Structure-based	• Validation of new protocols and scoring functions based in docking, molecular dynamics, and pharmacophore modeling.	Balius et al. (2013); Beato et al. (2013); Guo J et al. (2021); Ma et al. (2021); Niitsu and Sugita (2023)
	• Peptide and protein design.	

rationalizing the physicochemical, chemical, biological, and clinical data of active compounds (López-López et al., 2021a). Recently, decoys can be employed in several *de novo* protocols based on ligand or structure as summarized in Table 4.

4 Criteria to generate compound datasets with high quality

The quality of a data set is multifaceted. Commonly, it is associated with the experimental reproducibility of each data point and the experimental similarities between the protocols used to derive such data. Another important aspect of data quality is the balance between active and inactive compound. The latter is specially a challenge in public data sets due to the overall lack of published negative data. Finding qualitative yet better quantitative relationships between chemical structures and biological activity has been long pursued in medicinal chemistry and drug discovery. With the rapid increase and deployment of the predictive machine and deep learning methods, as well as the increased interest in the *de novo* design of chemical libraries (Mouchlis et al., 2021), the quantity and quality of data are

TABLE 5 Overview of suggested general criteria to generate quality datasets useful in *de novo* design.

Criteria	Brief description	Ref.
Balance	• Quality and quantity data allow the exploration of substantial regions of chemical space.	Scannell et al. (2022); Yang et al. (2023)
Quality (confidence) data	• The reliability of the activity data (active or inactive) is crucial to develop predictive models. This is the activity data reproducibility.	Kumar et al. (2022)
Diversity	• Datasets with a high chemical and structural diversity improve the generation of novel molecules.	Saldívar-González and Medina-Franco (2022)
Preparation or curation	• Dataset curation must be focused on one or multiple drug targets. Therefore, molecular descriptors and the cut-off threshold used for the curated must be properly selected. • Dataset should be oriented to resolve specific outcomes and avoid Pan-Assay Interference Compounds (PAINS) structures or chemical structures related to side effects. • In small datasets it is very important to have as much accurate data as possible. The maximum observable accuracy of classification models also depends on the experimental uncertainty and the distribution of the measured values. For instance, datasets with large noise are not recommended for the comparison of different models.	Fourches et al. (2016); Kramer and Lewis (2012)
Complete information	• According to the main objective of each project, the dataset used must contain reliable data related to the project's objective. For example, structure containing chemical and physicochemical information, bioactivity data for the related biological endpoint, or outcomes from clinical trials, etc.	López-López et al. (2021b); López-López and Medina-Franco. (2023); Wu et al. (2023a); Wu et al. (2023b)

becoming a central point in the discussion of the type of data sets needed (Schneider et al., 2020). While the more data (Cherkasov, 2023), the better, it is also true that the quality of the data available (that might not be quite large) is also crucial. Furthermore, the balance between active and inactive compounds is also a major consideration (López-López et al., 2022). Table 5 summarizes criteria for generating quality data sets. The list is not exhaustive but covers what the authors consider key points based on experience and what has been discussed extensively in the literature. Each point is supported by the references indicated in the table and further commented in the next subsections.

4.1 Balance

As discussed previously, several current data sets in the public domain are unbalanced due to the infrequent practice of reporting inactive compounds and negative data in general. Historically, the negative and inactive data of preclinical compounds has been ignored by most journals that favor the publication of most active compounds and positive results (Medina-Franco and López-López, 2022). However, inactive and negative data are essential in drug design and development. For example, the analysis of high-quality inactive and negative data improves clinical success rate, reduces costs associated with drug development, and reduces the side effects rates (Hayes and Hunter, 2012; López-López and Medina-Franco, 2023). Moreover, data mining and AI approaches are largely benefitted from inactive compounds (Yu, 2021; López-López et al., 2022). The use of inactive and negative data allows real data augmentation to develop AI models, improve their accuracy, and reduce the rate of false-positive cases (Korkmaz, 2020; *IBM*, 2022). Also, the inactive and negative data facilitates the generation of QSPRs models that allows the rationalization of basically any property (Kramer and Lewis, 2012; Norinder et al., 2019).

4.2 Confidence of the activity data

An unwritten rule on AI and computational projects in general is "garbage in, garbage out". This perspective has direct implications in drug design (Bajorath et al., 2022). Recent studies have demonstrated that the use of quality data allows generating of AI models with higher accuracy than the AI models generated from larger datasets but with low-quality.

4.3 Chemical and structural diversity

In general, a compound dataset with a large or broad applicability domain, as captured by the diversity of the contents, can give rise to predictive models with a large coverage. This is, molecules from diverse chemical structures could be conveniently interpolated in those models. As a comparison in an experimental setting, high-throughput screening of chemical diverse libraries increases the chances to find hit compounds for targets for which no hit compounds have been previously identified.

Due to the rapid expansion of the chemical universe, recently called the 'Big Bang' of the chemical universe (Cherkasov, 2023) it is relatively easy to have access to large and diverse regions of the chemical space. However, a practical challenge is to manage such large compound data sets computationally while developing and testing new models. A similar practical problem emerged when combinatorial chemistry was at its peak: it was challenging to design rationally novel large and diverse combinatorial libraries. To tackle this problem numerous diversity selection algorithms have been developed (Leach and Gillet, 2007). We recently applied a dissimilarity-based compound selection method to obtain three diverse subsets of natural products (with 14,994, 7,497, and 4,998 compounds, respectively) from the UNP. The subsets, that are freely available, can be readily used for *the novo* design applications and as benchmarks for similarity/diversity analysis (Chávez-Hernández and Medina-Franco, 2023).

4.4 Preparation or curation

A general curation protocol used on drug discovery datasets is to eliminate duplicate structures, canonize their SMILES representation, eliminate salts, and metals. However, according to the main goal of the *de novo* design model, additional steps to prepare a dataset could be taking into account, for example: 1) eliminating compounds with structural PAINS to reduce the rate of false-positive compounds prediction; 2) deleting compounds reported with side effects and/or ADMET deficiencies, to prioritize the generation of safe and optimization compounds.; or 3) making sure to keep in the dataset compounds with high activity confidence to improve the quality of predicted outputs. This list must be adapted according to the main goal of the *de novo* design model. It is also noted the need to develop robust and consistent protocols that take into scout metal-containing compounds as they have a major role in medicinal inorganic chemistry (Medina-Franco et al., 2022a).

4.5 Completeness

Chemical structures should contain the required or relevant information for the goals of the study. For instance, compounds should be annotated with stereochemistry information if the 3D structure and conformation is critical; electronic density and quantum chemical data if the reactivity is key point to predict; the type of the biological activity data such as biochemical, cell-based or functional assays; drug-drug interaction data, pharmacogenomics, or post-marketing annotations; should be aligned with the type of outcome to be predicted and later validated experimentally.

5 Perspectives of *de novo* design

One of the major perspectives of the *de novo* design is using balanced data sets (as much as experimental data is available) to build reliable models. Similar to QSAR predictive models, it is also crucial the validation of *de novo* protocols using standard and well-curated benchmark datasets (discussed in Section 3.6). With the increasing data availability to generate and train new models, it is becoming increasingly easy to explore regions of chemical space previously uncharted and continue contributing to the so-called "big bang" expansion of the chemical space. A major perspective in this direction is to explore biologically relevant compounds but outside the traditional small molecule chemical space (Medina-Franco et al., 2014). For instance, exploring metallodrugs (Medina-Franco et al., 2022a), macrocycles (Liang et al., 2022), peptides, or the combination of commonly explored chemical spaces, e.g., pseudo-natural products (discussed in Section 3.5).

6 Conclusion

Among the main types of datasets used in *the novo* design are on-demand collections, compounds annotated with biological activity, commercially available libraries, and natural products. More recently, a large benchmark data set was developed for machine learning applications. Although there is a general agreement in machine learning that the more data, the better, it is becoming more and more evident to consider the reliability and the quality of the data sets as critical features of the data. Part of the quality is associated with the balance between inactive and active compounds (in a rough analogy with the Yin-Yang concept), tasks that are not always feasible due to the general scarcity of negative (inactive compounds). The later point further emphasizes the continued need to publish and disclose negative results. Due to the fact that the experimental data of inactive compounds are not common, the community is using decoy data sets that by themselves are subject to design and refining using rational approaches. Decoy data sets try to fill the void of experimentally determined inactive molecules. Major criteria to take into account to generate compound data sets with high quality include balanced data sets in terms of active and inactive compounds (when the experimental information is available), structural and chemical diversity, curation or preparation according to the goals of the project, and complete information. All these together contribute to the perspectives of *de novo* design that foresees a continued and rapid expansion of molecules with the potential to become drugs.

Author contributions

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

Funding

Authors are grateful to DGAPA, UNAM, Programa de Apoyo a Proyectos de Investigación e Innovación Tecnológica (PAPIIT), grant no. IN201321. We also thank the Dirección General de Cómputo y de Tecnologías de Información y Comunicación (DGTIC), UNAM, for the computational resources to use Miztli supercomputer at UNAM under project LANCAD-UNAM-DGTIC-335; and the innovation space UNAM-HUAWEI the computational resources to use their supercomputer under project-7 "*Desarrollo y aplicación de algoritmos de inteligencia artificial para el diseño de fármacos aplicables al tratamiento de diabetes mellitus y cáncer*".

Acknowledgments

AC-H and EL-L are thankful to CONACyT, Mexico, for the Ph.D. scholarships number 847870 and 894234, respectively.

Conflict of interest

The author JLM-F declared that he was an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

The remaining 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.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Altae-Tran, H., Ramsundar, B., Pappu, A. S., and Pande, V. (2017). Low data drug discovery with one-shot learning. *ACS central Sci.* 3 (4), 283–293. doi:10.1021/acscentsci.6b00367

Arús-Pous, J., Patronov, A., Bjerrum, E. J., Tyrchan, C., Reymond, J. L., Chen, H., et al. (2020). SMILES-based deep generative scaffold decorator for de-novo drug design. *J. cheminformatics* 12 (1), 38. doi:10.1186/s13321-020-00441-8

Atanasov, A. G., Zotchev, S. B., Dirsch, V. M., and Supuran, C. T.International Natural Product Sciences Taskforce (2021). Natural products in drug discovery: Advances and opportunities. *Nat. Rev. Drug Discov.* 20 (3), 200–216. doi:10.1038/s41573-020-00114-z

Awale, M., and Reymond, J-L. (2015). Similarity mapplet: Interactive visualization of the directory of useful decoys and ChEMBL in high dimensional chemical spaces. *J. Chem. Inf. Model.* 55 (8), 1509–1516. doi:10.1021/acs.jcim.5b00182

Bajorath, J., Chávez-Hernández, A. L., Duran-Frigola, M., Fernández-de Gortari, E., Gasteiger, J., López-López, E., et al. (2022). Chemoinformatics and artificial intelligence colloquium: Progress and challenges in developing bioactive compounds. *J. cheminformatics* 14 (1), 82. doi:10.1186/s13321-022-00661-0

Balius, T. E., Allen, W. J., Mukherjee, S., and Rizzo, R. C. (2013). Grid-based molecular footprint comparison method for docking and de novo design: Application to HIVgp41. *J. Comput. Chem.* 34 (14), 1226–1240. doi:10.1002/jcc.23245

Barazorda-Ccahuana, H. L., Ranilla, L. G., Candia-Puma, M. A., Cárcamo-Rodriguez, E. G., Centeno-Lopez, A. E., Davila-Del-Carpio, G., et al. (2023). PeruNPDB: The Peruvian natural products database for *in silico* drug screening. *Sci. Rep.* 13 (1), 7577. doi:10.1038/s41598-023-34729-0

Bauer, M. R., Ibrahim, T. M., Vogel, S. M., and Boeckler, F. M. (2013). Evaluation and optimization of virtual screening workflows with DEKOIS 2.0-a public library of challenging docking benchmark sets. *J. Chem. Inf. Model.* 53 (6), 1447–1462. doi:10.1021/ci400115b

Beato, C., Beccari, A. R., Cavazzoni, C., Lorenzi, S., and Costantino, G. (2013). Use of experimental design to optimize docking performance: The case of LiGenDock, the docking module of LiGen, a new de novo design program. *J. Chem. Inf. Model.* 53 (6), 1503–1517. doi:10.1021/ci400079k

Bender, A., Schneider, N., Segler, M., Patrick Walters, W., Engkvist, O., and Rodrigues, T. (2022). Evaluation guidelines for machine learning tools in the chemical sciences. *Nat. Rev. Chem.* 6 (6), 428–442. doi:10.1038/s41570-022-00391-9

Bilodeau, C., Jin, W., Jaakkola, T., Barzilay, R., and Jensen, K. F. (2022). Generative models for molecular discovery: Recent advances and challenges. *Comput. Mol. Sci.* 12 (5), e1608. doi:10.1002/wcms.1608

Blaschke, T., Arús-Pous, J., Chen, H., Margreitter, C., Tyrchan, C., Engkvist, O., et al. (2020). Reinvent 2.0: An AI tool for de novo drug design. *J. Chem. Inf. Model.* 60 (12), 5918–5922. doi:10.1021/acs.jcim.0c00915

Brown, N., Fiscato, M., Segler, M. H. S., and Vaucher, A. C. (2019). GuacaMol: Benchmarking models for de Novo molecular design. *J. Chem. Inf. Model.* 59 (3), 1096–1108. doi:10.1021/acs.jcim.8b00839

Cao, L., Goreshnik, I., Coventry, B., Case, J. B., Miller, L., Kozodoy, L., et al. (2020). De novo design of picomolar SARS-CoV-2 miniprotein inhibitors. *Science* 370 (6515), 426–431. doi:10.1126/science.abd9909

Cereto-Massagué, A., Guasch, L., Valls, C., Mulero, M., Pujadas, G., and Garcia-Vallvé, S. (2012). DecoyFinder: An easy-to-use python GUI application for building target-specific decoy sets. *Bioinformatics* 28 (12), 1661–1662. doi:10.1093/bioinformatics/bts249

Chávez-Hernández, A. L., and Medina-Franco, J. L. (2023). Natural products subsets: Generation and characterization. *Artif. Intell. Life Sci.* 3, 100066. doi:10.1016/j.ailsci.2023.100066

Chávez-Hernández, A. L., Sánchez-Cruz, N., and Medina-Franco, J. L. (2020a). A fragment library of natural products and its comparative chemoinformatic characterization. *Mol. Inf.* 39 (11), e2000050. doi:10.1002/minf.202000050

Chávez-Hernández, A. L., Sánchez-Cruz, N., and Medina-Franco, J. L. (2020b). Fragment library of natural products and compound databases for drug discovery. *Biomolecules* 10 (11), 1518. doi:10.3390/biom10111518

Chemriya (2023). CHEMriya. Available at: https://chemriya.com/ (accessed May 13, 2023).

Chemspace (2023). Freedom space. Available at: https://chem-space.com/compounds/freedom-space (accessed May 13, 2023).

Chen, C. Y-C. (2011). TCM Database@Taiwan: The world's largest traditional Chinese medicine database for drug screening *in silico*. *PloS one* 6 (1), e15939. doi:10.1371/journal.pone.0015939

Chen, X., Lin, Y., Liu, M., and Gilson, M. K. (2002). The binding database: Data management and interface design. *Bioinformatics* 18 (1), 130–139. doi:10.1093/bioinformatics/18.1.130

Cherkasov, A. (2023). The 'Big Bang' of the chemical universe. *Nat. Chem. Biol.* 19, 667–668. doi:10.1038/s41589-022-01233-x

Corso, G., Stärk, H., Jing, B., et al. (2022). *DiffDock: Diffusion steps, twists, and turns for molecular docking*. arXiv [q-bio.BM]. Available at: http://arxiv.org/abs/2210.01776.

Costa, R. P. O., Lucena, L. F., Silva, L. M. A., Zocolo, G. J., Herrera-Acevedo, C., Scotti, L., et al. (2021). The SistematX web portal of natural products: An update. *J. Chem. Inf. Model.* 61 (6), 2516–2522. doi:10.1021/acs.jcim.1c00083

Davies, M., Nowotka, M., Papadatos, G., Dedman, N., Gaulton, A., Atkinson, F., et al. (2015). ChEMBL web services: Streamlining access to drug discovery data and utilities. *Nucleic acids Res.* 43 (W1), W612–W620. doi:10.1093/nar/gkv352

Dos Santos Nascimento, I. J., de Aquino, T. M., and da Silva-Júnior, E. F. (2021). Drug repurposing: A strategy for discovering inhibitors against emerging viral infections. *Curr. Med. Chem.* 28 (15), 2887–2942. doi:10.2174/0929867327666200812215852

DRUGBANK (2023). Celecoxib. Available at: https://go.drugbank.com/drugs/DB00482 (accessed May 13, 2023).

Enamine (2023). Real database. Available at: https://enamine.net/compound-collections/real-compounds/real-database (accessed May 13, 2023).

Evans, B. E., Rittle, K. E., Bock, M. G., DiPardo, R. M., Freidinger, R. M., Whitter, W. L., et al. (1988). Methods for drug discovery: Development of potent, selective, orally effective cholecystokinin antagonists. *J. Med. Chem.* 31 (12), 2235–2246. doi:10.1021/jm00120a002

Fourches, D., Muratov, E., and Tropsha, A. (2016). Trust, but verify II: A practical guide to chemogenomics data curation. *J. Chem. Inf. Model.* 56 (7), 1243–1252. doi:10.1021/acs.jcim.6b00129

Gallo, K., Kemmler, E., Goede, A., Becker, F., Dunkel, M., Preissner, R., et al. (2023). SuperNatural 3.0-a database of natural products and natural product-based derivatives. *Nucleic acids Res.* 51 (D1), D654–D659. doi:10.1093/nar/gkac1008

Gómez-Bombarelli, R., Wei, J. N., Duvenaud, D., Hernández-Lobato, J. M., Sánchez-Lengeling, B., Sheberla, D., et al. (2018). Automatic chemical design using a data-driven continuous representation of molecules. *ACS central Sci.* 4 (2), 268–276. doi:10.1021/acscentsci.7b00572

Gómez-García, A., and Medina-Franco, J. L. (2022). Progress and impact of Latin American natural product databases. *Biomolecules* 12 (9), 1202. doi:10.3390/biom12091202

Grebner, C. (2022). Webinar: "exploration and mining of large virtual chemical spaces. Available at: https://youtu.be/fMrI11SXwpU (accessed May 13, 2023).

Greener, J. G., Kandathil, S. M., Moffat, L., and Jones, D. T. (2022). A guide to machine learning for biologists. *Nat. Rev. Mol. Cell Biol.* 23 (1), 40–55. doi:10.1038/s41580-021-00407-0

Grigalunas, M., Brakmann, S., and Waldmann, H. (2022). Chemical evolution of natural product structure. *J. Am. Chem. Soc.* 144 (8), 3314–3329. doi:10.1021/jacs.1c11270

Gu, J., Gui, Y., Chen, L., Yuan, G., Lu, H. Z., and Xu, X. (2013). Use of natural products as chemical library for drug discovery and network pharmacology. *PloS one* 8 (4), e62839. doi:10.1371/journal.pone.0062839

Guo J, J., Janet, J. P., Bauer, M. R., Nittinger, E., Giblin, K. A., Papadopoulos, K., et al. (2021). DockStream: A docking wrapper to enhance de novo molecular design. *J. cheminformatics* 13 (1), 89. doi:10.1186/s13321-021-00563-7

Guo M, M., Thost, V., Li, B., et al. (2021). "Data-efficient graph grammar learning for molecular generation," in *International conference on learning representations*, 9. February 2021. Available at: https://research.ibm.com/publications/data-efficient-graph-grammar-learning-for-molecular-generation (accessed May 13, 2023).

Guo, M., Thost, V., Li, B., et al. (2022). Data-efficient graph grammar learning for molecular generation. arXiv [cs.LG]. Available at: http://arxiv.org/abs/2203.08031.

Hayes, A., and Hunter, J. (2012). Why is publication of negative clinical trial data important? *Br. J. Pharmacol.* 167 (7), 1395–1397. doi:10.1111/j.1476-5381.2012.02215.x

Hu, Q., Peng, Z., Sutton, S. C., Na, J., Kostrowicki, J., Yang, B., et al. (2012). Pfizer global virtual library (PGVL): A chemistry design tool powered by experimentally validated parallel synthesis information. *ACS Comb. Sci.* 14 (11), 579–589. doi:10.1021/co300096q

IBM (2022). How to use AI to discover new drugs and materials with limited data. Available at: https://research.ibm.com/blog/ai-discovery-with-limited-data#fnref-1 (accessed April 16, 2023).

Irwin, J. J. (2008). Community benchmarks for virtual screening. *J. computer-aided Mol. Des.* 22 (3-4), 193–199. doi:10.1007/s10822-008-9189-4

Jain, A. N., and Nicholls, A. (2008). Recommendations for evaluation of computational methods. *J. computer-aided Mol. Des.* 22 (3-4), 133–139. doi:10.1007/s10822-008-9196-5

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., et al. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature* 596 (7873), 583–589. doi:10.1038/s41586-021-03819-2

Juskalian, R., Regalado, A., Orcutt, M., et al. (2023). *10 breakthrough technologies 2020*. Available at: https://www.technologyreview.com/10-breakthrough-technologies/2020/ (accessed February 26, 2020).

Kadurin, A., Aliper, A., Kazennov, A., Mamoshina, P., Vanhaelen, Q., Khrabrov, K., et al. (2017). The cornucopia of meaningful leads: Applying deep adversarial autoencoders for new molecule development in oncology. *Oncotarget* 8 (7), 10883–10890. doi:10.18632/oncotarget.14073

Kim, S., Chen, J., Cheng, T., Gindulyte, A., et al. (2023). PubChem 2023 update. *Nucleic acids Res.* 51 (D1), D1373–D1380. doi:10.1093/nar/gkac956

Koes, D. R., and Camacho, C. J. (2012). ZINCPharmer: Pharmacophore search of the ZINC database. *Nucleic acids Res.* 40, W409–W414. Web Server issue). doi:10.1093/nar/gks378

Korkmaz, S. (2020). Deep learning-based imbalanced data classification for drug discovery. *J. Chem. Inf. Model.* 60 (9), 4180–4190. doi:10.1021/acs.jcim.9b01162

Korn, M., Ehrt, C., Ruggiu, F., Gastreich, M., and Rarey, M. (2023). Navigating large chemical spaces in early-phase drug discovery. *Curr. Opin. Struct. Biol.* 80, 102578. doi:10.1016/j.sbi.2023.102578

Kramer, C., and Lewis, R. (2012). QSARs, data and error in the modern age of drug discovery. *Curr. Top. Med. Chem.* 12 (17), 1896–1902. doi:10.2174/156802612804547380

Krenn, M., Häse, F., Nigam, A., Friederich, P., and Aspuru-Guzik, A. (2020). Self-referencing embedded strings (SELFIES): A 100% robust molecular string representation. *Mach. Learn. Sci. Technol.* 1 (4), 045024. doi:10.1088/2632-2153/aba947

Krishnan, S. R., Bung, N., Bulusu, G., and Roy, A. (2021). Accelerating de novo drug design against novel proteins using deep learning. *J. Chem. Inf. Model.* 61 (2), 621–630. doi:10.1021/acs.jcim.0c01060

Kumar, S. A., Ananda Kumar, T. D., Beeraka, N. M., Pujar, G. V., Singh, M., Narayana Akshatha, H. S., et al. (2022). Machine learning and deep learning in data-driven decision making of drug discovery and challenges in high-quality data acquisition in the pharmaceutical industry. *Future Med. Chem.* 14 (4), 245–270. doi:10.4155/fmc-2021-0243

Leach, A. R., and Gillet, V. J. (2007). "Selecting diverse dets of compounds," in *An introduction to chemoinformatics* (Dordrecht: Springer Netherlands), 119–139. doi:10.1007/978-1-4020-6291-9_6

Li, S., Wang, L., Meng, J., Zhao, Q., Zhang, L., and Liu, H. (2022). De Novo design of potential inhibitors against SARS-CoV-2 Mpro. *Comput. Biol. Med.* 147, 105728. doi:10.1016/j.compbiomed.2022.105728

Li, Y., Zhang, L., and Liu, Z. (2018). Multi-objective de novo drug design with conditional graph generative model. *J. cheminformatics* 10 (1), 33. doi:10.1186/s13321-018-0287-6

Liang, Y., Fang, R., and Rao, Q. (2022). An insight into the medicinal chemistry perspective of macrocyclic derivatives with antitumor activity: A systematic review. *Molecules* 27 (9), 2837. doi:10.3390/molecules27092837

Lipinski, C. A., Lombardo, F., Dominy, B. W., and Feeney, P. J. (2001). Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. drug Deliv. Rev.* 46 (1-3), 3–26. doi:10.1016/s0169-409x(00)00129-0

Liu, X., Ye, K., van Vlijmen, H. W. T., Ijzerman, A. P., and van Westen, G. J. P. (2019). An exploration strategy improves the diversity of de novo ligands using deep reinforcement learning: A case for the adenosine A2A receptor. *J. cheminformatics* 11 (1), 35. doi:10.1186/s13321-019-0355-6

López-López, E., Bajorath, J., and Medina-Franco, J. L. (2021a). Informatics for chemistry, biology, and biomedical sciences. *J. Chem. Inf. Model.* 61 (1), 26–35. doi:10.1021/acs.jcim.0c01301

López-López, E., Cerda-García-Rojas, C. M., and Medina-Franco, J. L. (2021b). Tubulin inhibitors: A chemoinformatic analysis using cell-based data. *Molecules* 26 (9), 2483. doi:10.3390/molecules26092483

López-López, E., Fernández-de Gortari, E., and Medina-Franco, J. L. (2022). Yes SIR! On the structure-inactivity relationships in drug discovery. *Drug Discov. today* 27 (8), 2353–2362. doi:10.1016/j.drudis.2022.05.005

López-López, E., and Medina-Franco, J. L. (2023). Towards decoding hepatotoxicity of approved drugs through navigation of multiverse and consensus chemical spaces. *Biomolecules* 13 (1), 176. doi:10.3390/biom13010176

Ma, B., Terayama, K., Matsumoto, S., Isaka, Y., Sasakura, Y., Iwata, H., et al. (2021). Structure-based de novo molecular generator combined with artificial intelligence and docking simulations. *J. Chem. Inf. Model.* 61 (7), 3304–3313. doi:10.1021/acs.jcim.1c00679

Maziarka, L., Pocha, A., Kaczmarczyk, J., Rataj, K., Danel, T., and Warchoł, M. (2020). Mol-CycleGAN: A generative model for molecular optimization. *J. cheminformatics* 12 (1), 2. doi:10.1186/s13321-019-0404-1

Medina-Franco, J. L., Flores-Padilla, E. A., and Chávez-Hernández, A. L. (2022b). "Chapter 23 - discovery and development of lead compounds from natural sources using computational approaches," in *Evidence-based validation of herbal medicine*. Editor P. K. Mukherjee Second Edition (Elsevier), 539–560. doi:10.1016/B978-0-323-85542-6.00009-3

Medina-Franco, J. L., López-López, E., Andrade, E., Ruiz-Azuara, L., Frei, A., Guan, D., et al. (2022a). Bridging informatics and medicinal inorganic chemistry: Toward a database of metallodrugs and metallodrug candidates. *Drug Discov. today* 27 (5), 1420–1430. doi:10.1016/j.drudis.2022.02.021

Medina-Franco, J. L., and López-López, E. (2022). The essence and transcendence of scientific publishing. *Front. Res. metrics Anal.* 7, 822453. doi:10.3389/frma.2022.822453

Medina-Franco, J. L., Martinez-Mayorga, K., and Meurice, N. (2014). Balancing novelty with confined chemical space in modern drug discovery. *Expert Opin. drug Discov.* 9 (2), 151–165. doi:10.1517/17460441.2014.872624

Medina-Franco, J. L., Naveja, J. J., and López-López, E. (2019). Reaching for the bright StARs in chemical space. *Drug Discov. today* 24 (11), 2162–2169. doi:10.1016/j.drudis.2019.09.013

Mendez, D., Gaulton, A., Bento, A. P., Chambers, J., De Veij, M., Félix, E., et al. (2019). ChEMBL: Towards direct deposition of bioassay data. *Nucleic acids Res.* 47 (D1), D930–D940. doi:10.1093/nar/gky1075

Mohanraj, K., Karthikeyan, B. S., Vivek-Ananth, R. P., Chand, R. P. B., Aparna, S. R., Mangalapandi, P., et al. (2018). Imppat: A curated database of indian medicinal plants, phytochemistry and therapeutics. *Sci. Rep.* 8 (1), 4329. doi:10.1038/s41598-018-22631-z

Mouchlis, V. D., Afantitis, A., Serra, A., Fratello, M., Papadiamantis, A. G., Aidinis, V., et al. (2021). Advances in de novo drug design: From conventional to machine learning methods. *Int. J. Mol. Sci.* 22 (4), 1676. doi:10.3390/ijms22041676

Mysinger, M. M., Carchia, M., Irwin, J. J., and Shoichet, B. K. (2012). Directory of useful decoys, enhanced (DUD-E): Better ligands and decoys for better benchmarking. *J. Med. Chem.* 55 (14), 6582–6594. doi:10.1021/jm300687e

Newman, D. J., and Cragg, G. M. (2020). Natural products as sources of new drugs over the nearly four decades from 01/1981 to 09/2019. *J. Nat. Prod.* 83 (3), 770–803. doi:10.1021/acs.jnatprod.9b01285

NIH (2023). Natural product libraries. Available at: https://www.nccih.nih.gov/grants/natural-product-libraries.

Niitsu, A., and Sugita, Y. (2023). Towards de novo design of transmembrane α-helical assemblies using structural modelling and molecular dynamics simulation. *Phys. Chem. Chem. Phys. PCCP* 25 (5), 3595–3606. doi:10.1039/d2cp03972a

Norinder, U., Naveja, J. J., López-López, E., Mucs, D., and Medina-Franco, J. L. (2019). Conformal prediction of HDAC inhibitors. *SAR QSAR Environ. Res.* 30 (4), 265–277. doi:10.1080/1062936X.2019.1591503

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 (10), e78085. doi:10.1371/journal.pone.0078085

Olivecrona, M., Blaschke, T., Engkvist, O., and Chen, H. (2017). Molecular de-novo design through deep reinforcement learning. *J. cheminformatics* 9 (1), 48. doi:10.1186/s13321-017-0235-x

Olmedo, A. D., and Medina-Franco, J. L. (2020). "Chemoinformatic approach: The case of natural products of Panama," in *Cheminformatics and its applications* (IntechOpen). doi:10.5772/intechopen.87779

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 (4), 779–789. doi:10.1007/s11030-017-9781-4

Palazzesi, F., and Pozzan, A. (2022). "Deep learning applied to ligand-based de novo drug DesignDe novo drug design," in *Artificial intelligence in drug design*. Editor A. Heifetz (New York, NY: Springer US), 273–299. doi:10.1007/978-1-0716-1787-8_12

Papadopoulos, K., Giblin, K. A., Janet, J. P., Patronov, A., and Engkvist, O. (2021). De novo design with deep generative models based on 3D similarity scoring. *Bioorg. Med. Chem.* 44, 116308. doi:10.1016/j.bmc.2021.116308

Patel, H. M., Noolvi, M. N., Sharma, P., Jaiswal, V., Bansal, S., Lohan, S., et al. (2014). Quantitative structure–activity relationship (QSAR) studies as strategic approach in drug discovery. *Med. Chem. Res. Int. J. rapid Commun. Des. Mech. action Biol. Act. agents* 23 (12), 4991–5007. doi:10.1007/s00044-014-1072-3

Perron, Q., da Silva, V. B. R., Atwood, B., and Gaston-Mathé, Y. (2022b). Key points to succeed in Artificial Intelligence drug discovery projects. *Chem. Int.* 44 (1), 19–21. doi:10.1515/ci-2022-0106

Perron, Q., Mirguet, O., Tajmouati, H., Skiredj, A., Rojas, A., Gohier, A., et al. (2022a). Deep generative models for ligand-based de novo design applied to multi-parametric optimization. *J. Comput. Chem.* 43 (10), 692–703. doi:10.1002/jcc.26826

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 (1), 7215. doi:10.1038/s41598-017-07451-x

Pilón-Jiménez, B. A., Saldívar-González, F. I., Díaz-Eufracio, B. I., and Medina-Franco, J. L. (2019). Biofacquim: A Mexican compound database of natural products. *Biomolecules* 9 (1), 31. doi:10.3390/biom9010031

Polykovskiy, D., Zhebrak, A., Sanchez-Lengeling, B., Golovanov, S., Tatanov, O., Belyaev, S., et al. (2020). Molecular sets (MOSES): A benchmarking platform for molecular generation models. *Front. Pharmacol.* 11, 565644. doi:10.3389/fphar.2020.565644

Réau, M., Langenfeld, F., Zagury, J-F., Lagarde, N., and Montes, M. (2018). Decoys selection in benchmarking datasets: Overview and perspectives. *Front. Pharmacol.* 9, 11. doi:10.3389/fphar.2018.00011

Reymond, J-L. (2015). The chemical space project. *Accounts Chem. Res.* 48 (3), 722–730. doi:10.1021/ar500432k

Rohrer, S. G., and Baumann, K. (2009). Maximum unbiased validation (MUV) data sets for virtual screening based on PubChem bioactivity data. *J. Chem. Inf. Model.* 49 (2), 169–184. doi:10.1021/ci8002649

Sabe, V. T., Ntombela, T., Jhamba, L. A., Maguire, G. E. M., Govender, T., Naicker, T., et al. (2021). Current trends in computer aided drug design and a highlight of drugs discovered via computational techniques: A review. *Eur. J. Med. Chem.* 224, 113705. doi:10.1016/j.ejmech.2021.113705

Saldívar-González, F. I., Aldas-Bulos, V. D., Medina-Franco, J. L., and Plisson, F. (2022). Natural product drug discovery in the artificial intelligence era. *Chem. Sci.* 13 (6), 1526–1546. doi:10.1039/d1sc04471k

Saldívar-González, F. I., and Medina-Franco, J. L. (2022). Approaches for enhancing the analysis of chemical space for drug discovery. *Expert Opin. drug Discov.* 17 (7), 789–798. doi:10.1080/17460441.2022.2084608

Saldívar-González, F. I., Valli, M., Andricopulo, A. D., da Silva Bolzani, V., and Medina-Franco, J. L. (2019). Chemical space and diversity of the NuBBE database: A chemoinformatic characterization. *J. Chem. Inf. Model.* 59 (1), 74–85. doi:10.1021/acs.jcim.8b00619

Sánchez-Cruz, N., Pilón-Jiménez, B. A., and Medina-Franco, J. L. (2019) Functional group and diversity analysis of BIOFACQUIM: A Mexican natural product database. *F1000Research* 8, Chem Inf Sci-2071. doi:10.12688/f1000research.21540.2

Scannell, J. W., Bosley, J., Hickman, J. A., Dawson, G. R., Truebel, H., Ferreira, G. S., et al. (2022). Predictive validity in drug discovery: What it is, why it matters and how to improve it. *Nat. Rev. Drug Discov.* 21 (12), 915–931. doi:10.1038/s41573-022-00552-x

Schneider, G., and Clark, D. E. (2019). Automated de novo drug design: Are we nearly there yet? *Angew. Chem.* 58 (32), 10792–10803. doi:10.1002/anie.201814681

Schneider, G., Clément-Chomienne, O., Hilfiger, L., SchneiderKirschBöhm, et al. (2000). Virtual screening for bioactive molecules by evolutionary de novo design. *Angew. Chem.* 39 (22), 4130–4133. doi:10.1002/1521-3773(20001117)39:22<4130:aid-anie4130>3.0.co;2-e

Schneider, P., and Schneider, G. (2017). Privileged structures revisited. *Angew. Chem.* 56 (27), 7971–7974. doi:10.1002/anie.201702816

Schneider, P., Walters, W. P., Plowright, A. T., Sieroka, N., Listgarten, J., Goodnow, R. A., et al. (2020). Rethinking drug design in the artificial intelligence era. *Nat. Rev. Drug Discov.* 19 (5), 353–364. doi:10.1038/s41573-019-0050-3

Scotti, M. T., Herrera-Acevedo, C., Oliveira, T. B., Costa, R. P. O., Santos, S. Y. K. d. O., Rodrigues, R. P., et al. (2018). SistematX, an online web-based cheminformatics tool for data management of secondary metabolites. *Molecules* 23 (1), 103. doi:10.3390/molecules23010103

Sheridan, R. P. (2013). Time-split cross-validation as a method for estimating the goodness of prospective prediction. *J. Chem. Inf. Model.* 53 (4), 783–790. doi:10.1021/ci400084k

Shipman, J. T., Su, X., Hua, D., and Desaire, H. (2019). DecoyDeveloper: An on-demand, de novo decoy glycopeptide generator. *J. proteome Res.* 18 (7), 2896–2902. doi:10.1021/acs.jproteome.9b00203

Simonovsky, M., and Komodakis, N. (2018). "GraphVAE: Towards generation of small graphs using variational autoencoders," in *Artificial neural networks and machine learning – icann 2018* (Springer International Publishing), 2018, 412–422. doi:10.1007/978-3-030-01418-6_41

Skalic, M., Jiménez, J., Sabbadin, D., and De Fabritiis, G. (2019b). Shape-based generative modeling for de novo drug design. *J. Chem. Inf. Model.* 59 (3), 1205–1214. doi:10.1021/acs.jcim.8b00706

Skalic, M., Sabbadin, D., Sattarov, B., Sciabola, S., and De Fabritiis, G. (2019a). From target to drug: Generative modeling for the multimodal structure-based ligand design. *Mol. Pharm.* 16 (10), 4282–4291. doi:10.1021/acs.molpharmaceut.9b00634

Soares, T. A., Nunes-Alves, A., Mazzolari, A., Ruggiu, F., Wei, G. W., and Merz, K. (2022). The (Re)-evolution of quantitative structure-activity relationship (qsar) studies propelled by the surge of machine learning methods. *J. Chem. Inf. Model.* 62 (22), 5317–5320. doi:10.1021/acs.jcim.2c01422

Sorokina, M., Merseburger, P., Rajan, K., Yirik, M. A., and Steinbeck, C. (2021). COCONUT online: Collection of open natural products database. *J. cheminformatics* 13 (1), 2. doi:10.1186/s13321-020-00478-9

Tingle, B. I., Tang, K. G., Castanon, M., Gutierrez, J. J., Khurelbaatar, M., Dandarchuluun, C., et al. (2023). ZINC-22─A free multi-billion-scale database of tangible compounds for ligand discovery. *J. Chem. Inf. Model.* 63 (4), 1166–1176. doi:10.1021/acs.jcim.2c01253

Tong, X., Liu, X., Tan, X., Jiang, J., Xiong, Z., et al. (2021). Generative models for de novo drug design. *J. Med. Chem.* 64 (19), 14011–14027. doi:10.1021/acs.jmedchem.1c00927

Ullanat, V. (2020). "Variational autoencoder as a generative tool to produce de-novo lead compounds for biological targets," in *2020 14th international conference on innovations in information Technology (IIT)*, 102–107. doi:10.1109/IIT50501.2020.9299078

UNIIQUIM (2015). Uniiquim. Available at: https://uniiquim.iquimica.unam.mx/ (accessed May 13, 2023).

Valli, M., dos Santos, R. N., Figueira, L. D., Nakajima, C. H., Castro-Gamboa, I., Andricopulo, A. D., et al. (2013). Development of a natural products database from the biodiversity of Brazil. *J. Nat. Prod.* 76 (3), 439–444. doi:10.1021/np3006875

Vamathevan, J., Clark, D., Czodrowski, P., Dunham, I., Ferran, E., Lee, G., et al. (2019). Applications of machine learning in drug discovery and development. *Nat. Rev. Drug Discov.* 18 (6), 463–477. doi:10.1038/s41573-019-0024-5

Veber, D. F., Johnson, S. R., Cheng, H-Y., Smith, B. R., Ward, K. W., and Kopple, K. D. (2002). Molecular properties that influence the oral bioavailability of drug candidates. *J. Med. Chem.* 45 (12), 2615–2623. doi:10.1021/jm020017n

Wang, L., Pang, X., Li, Y., Zhang, Z., and Tan, W. (2017). Rader: A RApid DEcoy retriever to facilitate decoy based assessment of virtual screening. *Bioinformatics* 33 (8), 1235–1237. doi:10.1093/bioinformatics/btw783

Wang, M., Hsieh, C-Y., Wang, J., Wang, D., Weng, G., Shen, C., et al. (2022). Relation: A deep generative model for structure-based de novo drug design. *J. Med. Chem.* 65 (13), 9478–9492. doi:10.1021/acs.jmedchem.2c00732

Warr, W. A., Nicklaus, M. C., Nicolaou, C. A., and Rarey, M. (2022). Exploration of ultralarge compound collections for drug discovery. *J. Chem. Inf. Model.* 62 (9), 2021–2034. doi:10.1021/acs.jcim.2c00224

Warr, W. (2021). Report on an NIH workshop on ultralarge chemistry databases. Chemrxiv: 43. Available at: https://chemrxiv.org/engage/chemrxiv/article-details/60c75883bdbb89984ea3ada5.

Weininger, D. (1988). SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* 28 (1), 31–36. doi:10.1021/ci00057a005

Wishart, D. S., Feunang, Y. D., Guo, A. C., Lo, E. J., Marcu, A., Grant, J. R., et al. (2018). DrugBank 5.0: A major update to the DrugBank database for 2018. *Nucleic acids Res.* 46 (D1), D1074–D1082. doi:10.1093/nar/gkx1037

Wishart, D. S., Knox, C., Guo, A. C., Cheng, D., Shrivastava, S., Tzur, D., et al. (2008). DrugBank: A knowledgebase for drugs, drug actions and drug targets. *Nucleic acids Res.* 36, D901–D906. doi:10.1093/nar/gkm958

Wishart, D. S., Knox, C., Guo, A. C., Shrivastava, S., Hassanali, M., Stothard, P., et al. (2006). DrugBank: A comprehensive resource for *in silico* drug discovery and exploration. *Nucleic acids Res.* 34, D668–D672. doi:10.1093/nar/gkj067

Wu, A., Ye, Q., Zhuang, X., Chen, Q., Zhang, J., Wu, J., et al. (2023a). Elucidating structures of complex organic compounds using a machine learning model based on the 13C NMR chemical shifts. *Precis. Chem.* 1 (1), 57–68. doi:10.1021/prechem.3c00005

Wu, J., Xiao, Y., Cai, H., Zhao, D., Li, Y., et al. (2023b). DeepCancerMap: A versatile deep learning platform for target- and cell-based anticancer drug discovery. *Eur. J. Med. Chem.* 255, 115401. doi:10.1016/j.ejmech.2023.115401

Wu, Z., Ramsundar, B., Feinberg, E. N., Gomes, J., Geniesse, C., Pappu, A. S., et al. (2018). MoleculeNet: A benchmark for molecular machine learning. *Chem. Sci.* 9 (2), 513–530. doi:10.1039/c7sc02664a

Xie, W., Wang, F., Li, Y., Lai, L., and Pei, J. (2022). Advances and challenges in de novo drug design using three-dimensional deep generative models. *J. Chem. Inf. Model.* 62 (10), 2269–2279. doi:10.1021/acs.jcim.2c00042

Yang, J., Wang, D., Jia, C., Wang, M., Hao, G., and Yang, G. (2019). Freely accessible chemical database resources of compounds for *in silico* drug discovery. *Curr. Med. Chem.* 26 (42), 7581–7597. doi:10.2174/0929867325666180508100436

Yang, X., Yang, G., and Chu, J. (2023). *The balanced matrix factorization for computational drug repositioning*. arXiv [cs.CE]. Available at: http://arxiv.org/abs/2301.06448.

Yu, H. (2021). Responsible use of negative research outcomes-accelerating the discovery and development of new antibiotics. *J. antibiotics* 74 (9), 543–546. doi:10.1038/s41429-021-00439-w

Zhang, Y., Luo, M., Wu, P., Wu, S., Lee, T. Y., and Bai, C. (2022). Application of computational biology and artificial intelligence in drug design. *Int. J. Mol. Sci.* 23 (21), 13568. doi:10.3390/ijms232113568

Zhavoronkov, A., Ivanenkov, Y. A., Aliper, A., Veselov, M. S., Aladinskiy, V. A., Aladinskaya, A. V., et al. (2019). Deep learning enables rapid identification of potent DDR1 kinase inhibitors. *Nat. Biotechnol.* 37 (9), 1038–1040. doi:10.1038/s41587-019-0224-x

Capítulo de libro

CHAPTER

23

Discovery and development of lead compounds from natural sources using computational approaches

José L. Medina-Franco, E. Alexis Flores-Padilla, and Ana L. Chávez-Hernández

DIFACQUIM Research Group, Department of Pharmacy, School of Chemistry, National Autonomous University of Mexico, Mexico City, Mexico

List of abbreviations

ADME/Tox	absorption, distribution, metabolism, excretion, and toxicity
CADD	computer-aided drug design
CADS	computer-aided drug selection
COCONUT	Collection of Open Natural Products
CoVs	coronaviruses
DNMT	DNA methyltransferase
HBA	hydrogen bond acceptors
HBD	hydrogen bond donors
MD	molecular dynamics
MOE	Molecular Operating Environment
M^{pro}	membrane protein
MW	molecular weight
NP	natural products
PDB	Protein Data Bank
RB	rotatable bonds
SAH	*S*-adenosyl-*L*-homocysteine
SAM	*S*-adenosyl-L-methionine
SAR	structure-activity relationships
SARS-CoV-2	Severe Acute Respiratory Syndrome Coronavirus 2
SlogP	octanol/water partition coefficient
SMARts	structure multiple-activity relationships
TCM	Traditional Chinese Medicine
TMAP	Tree Map
TPSA	topological polar surface area
VS	virtual screening

1 Natural products in drug discovery

Natural products (NP), from either terrestrial or aquatic organisms, have a long tradition as sources of active compounds for health-related benefits. From the approved drugs between 1981 and 2019, 3.8% corresponds to unaltered NP, and 18.9% are NP derivatives [1]. An example of an NP recently approved for clinical use is migalastat (Galafold®) (Fig. 1) to treat Fabry disease. This compound that was isolated as a fermentation product of the bacterium *Streptomyces lydicus*, is approved for clinical use (as of August 2018) in Australia, Canada, Israel, Japan, South Korea, Switzerland, the United States of America, and the European Union. Other drugs recently approved for clinical use that are derivatives of NP are the antiparasitic compound moxidectin, and the antibacterial plazomicin (Fig. 1). Moxidectin is synthetically derived from nemadectin and plazomicin is synthetically derived from sisomicin. It is also well known that over millions of years, Nature has selected and optimized chemical structures to produce chemical scaffolds and compounds enriched with biological function. However, NP hurdles include challenges in the isolation and purification procedures, minimal available amounts of lead compounds, the difficulty in synthesizing NP with high structural complexity, and the associated synthesis scale-up issues. Also, for drug discovery applications, caution should be taken with compounds that have been designed by Nature for defense and are toxic. As such, one can expect that not all NP has a beneficial effect on health. However, the considerable success of using NP to produce bioactive compounds or bioactive mixtures has inspired the preparation of synthetic molecules that have become drugs approved for clinical use [1].

https://doi.org/10.1016/B978-0-323-85542-6.00009-3

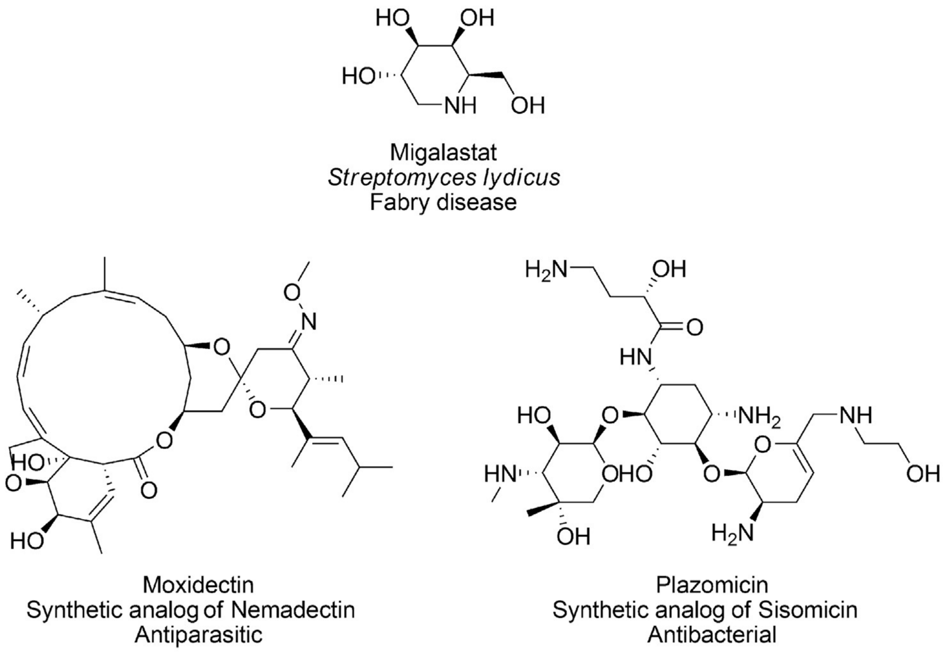


FIG. 1 Recent natural products and derivatives approved for clinical use.

Besides, the unique structural features of NP, such as structural complexity [2], represent a promising opportunity to identify active or selective compounds for emerging targets [3] or those targets that are difficult to tackle with classical synthetic molecules.

The number of applications of computational approaches to improve and accelerate NP-based drug discovery is increasing. This fact is documented in several recent book chapters and review papers [4–8] that discuss the range of molecular modeling, chemoinformatics [9], machine learning [10], and other computer-aided drug design approaches that are used to optimize drug candidates from natural origin and to understand the coverage of NP in chemical space. This chapter also discusses new molecular modeling and chemoinformatics applications to identify bioactive NP with potential therapeutic use. We also present computational techniques to optimize the biological activity and understand at the molecular level the underlying mechanism of action of bioactive compounds and anticipate their potential issues of toxicity. This chapter is an update of several topics covered 5 years ago and published in the first edition of this book [4].

The chapter is organized into six major sections: after a brief introduction of NP-based drug discovery, Section 2 presents a general overview of the drug discovery process emphasizing the different computational approaches used in the discovery and development of lead compounds. In the same section, we comment on the factors affecting drug discovery from natural products in computational approaches. Section 3 discusses recent progress on compounds databases, emphasizing collections of NP available in the public domain. The next section describes the NP's characterization and profiling in terms of chemical diversity, coverage of chemical space, toxicity, and molecular complexity. Section 5 presents the detection of compounds from natural origin with a potential therapeutic application and identifies potential molecular targets of bioactive compounds. This section emphasizes the role of NP for COVID-19 drug discovery and epigenetic drug discovery. Section 6 covers molecular modeling's role in exploring the mechanism of action of NP at the molecular level. The last section presents concluding remarks and an outlook.

2 Computer-aided drug design

Computer-aided drug design (CADD) includes a large group of theoretical and computational approaches that are part of modern drug discovery. These methods include molecular modeling, chemoinformatics, bioinformatics, and other theoretical disciplines [11]. CADD has made major contributions to help to bring compounds to the clinic. Indeed, several marketed drugs such as imatinib, zanamivir and nelfinavir, and other clinical candidates have been identified or optimized with the aid of molecular modeling techniques [12]. Fig. 2 shows a schematic representation of the main steps involved in the drug discovery process and the type of application and representative computational approaches that are used along the process. The role of CADD in the drug discovery process lies mainly in the phase of lead identification and optimization.

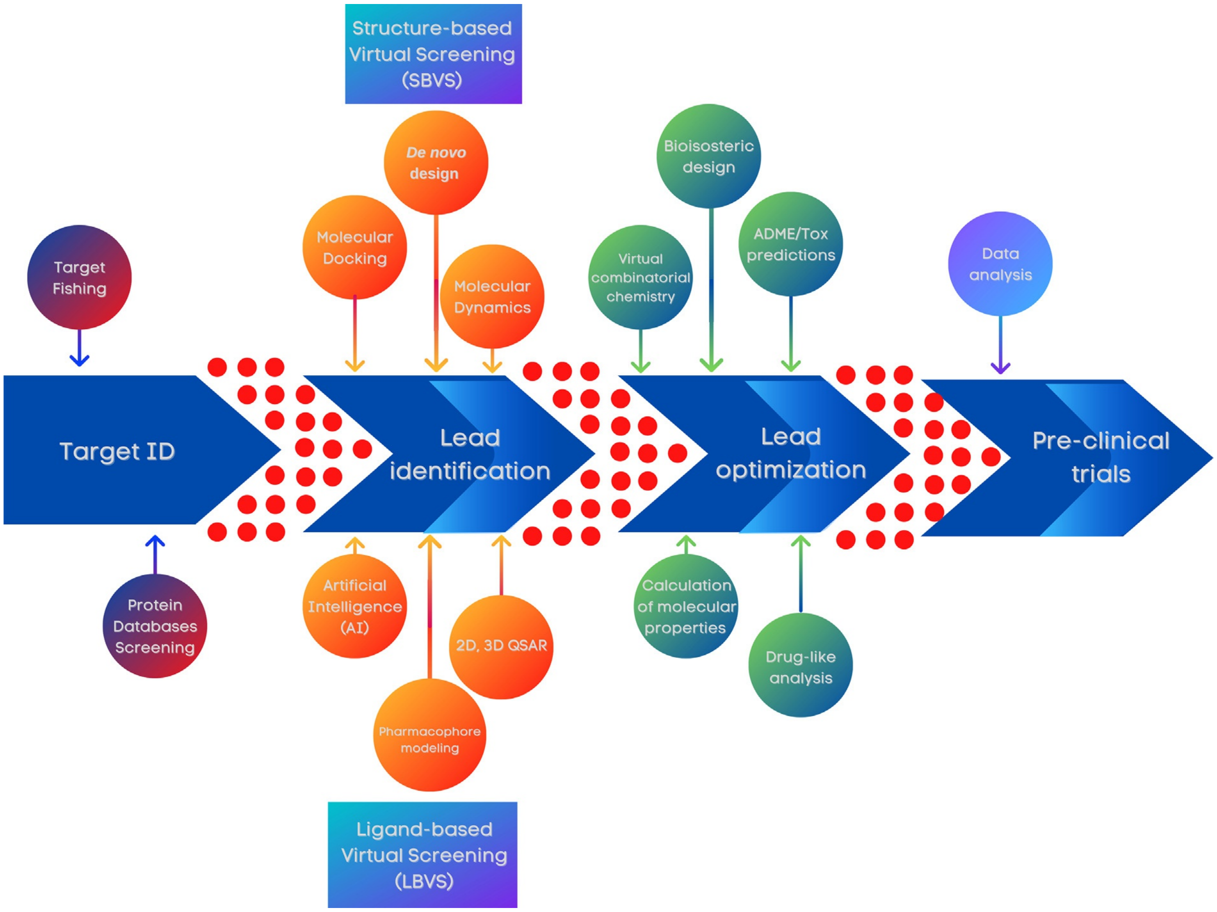


FIG. 2 Schematic representation of main steps involved in the drug discovery and the applications of computational methods at different stages.

Recent progress on CADD has been recently reviewed and the reader is referred to the literature for the details [12]. Briefly, computational calculations have played a significant role in the investigation of molecules that are currently in clinical use. For instance, *in silico* methods have made notable contributions in the treatment of acquired immunodeficiency syndrome, influenza virus infections, in the treatment of glaucoma, and in the treatment of patients with non-small cell lung cancer.

As reviewed elsewhere, CADD approaches can be classified in three main areas: structure-based, ligand-based and hybrid methods. Structure-based methods, such as molecular docking and molecular dynamics, depending on the three-dimension information of the molecular target. Applications of structure-based methods include characterization of biding sites, elucidation of the mechanism of action of active molecules at the molecular level, and evaluation of the kinetics and thermodynamics involved in the ligand-target recognition process.

Ligand-based methods depend on the information of the chemical structures of a group of molecules (aka ligands) with known biological activity. One of the main goal of this method is to identify some bioactive compounds or to improve the activity of active molecules. Typical examples of ligand-based methods are similarity searching, QSAR modeling, and machine learning models that depend on the structure of the ligands [13].

When the structure of the target is known as well as the structure of active molecules, it is feasible to apply hybrid methods, i.e., a combination of structure- and ligand-based. Examples are *in silico* approaches to predict bioactivity based on the biological profile of compounds tested vs. one or multiple targets. Other examples are pharmacophore modeling.

One of the major issues to deal with natural products using chemoinformatics approaches is the type of molecules that could be too large or complex. For instance, several molecular fingerprints typically used to represent small organic compounds are challenging to apply to natural products. Therefore, there is an effort to develop molecular representations general as possible [14].

3 Natural product databases

Compound databases have a prominent role in drug discovery. This is particularly relevant with the advent of big data. Indeed, high-throughput experimental and virtual screening (VS) of large chemical databases generate an enormous amount of data that need to be stored and made accessible to convert data into information and finally into knowledge [15]. One of the applications of chemoinformatics (also known as cheminformatics or chemical informatics) [11] in NP research is the organization, analysis, and dissemination of chemical information of NP in compound databases [16,17]. In fact, the increasing amount of informatics applications in drug discovery has led to the term “natural products informatics” [11]. There are several excellent and extensive reviews of NP databases published over the past 5 years [6,16,18,19]. One of the most recent reviews is a compilation of more than 100 public NP databases from different sources that collect more than 400,000 non-redundant molecules [20,21]. Among the numerous NP databases in the public domain, there are initiatives to compile in a single platform, NP from different geographical regions, including Africa (e.g., African Natural Products Database—ANPDB) [22], Latin America (e.g., Latin American Natural Product Database—LANaPD) [17], and Vietnam [23]. Also, there are compound databases of NP focused on a specific therapeutic indication. A contemporary example is DiaNat-DB, a compound collection of more than 330 antidiabetic compounds from medicinal plants [24].

Besides, there are efforts to make publicly available databases of fragments derived from NP for NP-based fragment-based drug discovery and the generation of “pseudo-NP” [25]. For instance, Chávez-Hernández recently reported an extensive fragment library with nearly 206,000 fragments derived from a drug-like subset of the Collection of Open Natural Products (COCONUT) database. In that work, the fragment library of NP was compared to fragment libraries of ChEMBL as representative of biologically relevant compounds and a vast on-demand database of synthetic molecules. The fragment library of NP was made freely available [26]. Of note, COCONUT is one of the largest compilations of NP available for which a website has been developed to browse the contents [21].

4 Chemoinformatic studies

In addition to the construction and maintenance of NP databases, computational methods are used to analyze compound databases’ contents and obtain a detailed profile of various features of common interest for drug discovery applications. Common examples include the systematic analysis of chemical diversity using different structural and molecular representations, a profile of physicochemical properties of pharmaceutical interest, molecular complexity, visual representation of the chemical space, and *in silico* profiling of absorption, distribution, metabolism, excretion, and toxicity (ADME/Tox). There are well-established chemoinformatic protocols to obtain a detailed profile of these characteristics [27].

4.1 Physicochemical properties

Molecular descriptors frequently used to describe chemical libraries include molecular weight (MW), the octanol/water partition coefficient (SlogP), topological polar surface area (TPSA), hydrogen bond donors (HBD), hydrogen bond acceptors (HBA), and the number of rotatable bonds (RB). These descriptors are typically used to quantify lead-like and drug-like features of compound data sets. In general, these descriptors are intended to capture three significant features of interest in drug development, namely size (MW), polarity (SlogP, TPSA, HBDs, HBAs), and flexibility (RBs) [28,29]. NP data sets have been profiled for more than 15 years in terms of such six molecular descriptors. It is also common to include in such analysis the distribution of other simple yet relevant structural features such as counts of carbon, nitrogen, oxygen atoms, and different types of rings (total number, aromatic, heteroaromatic, etc.) [30,31].

Pilón-Jimenez et al. reported a comparative analysis of BIOFACQUIM, a NP database from Mexico with drugs approved for clinical use, NP from NuBBE$_{DB}$, marine NP, cyanobacteria, and fungi metabolites [32]. The authors concluded that compounds in BIOFACQUIM are more similar to NuBBE$_{DB}$ and fungi data sets. In a separate and also recent study, Simoben et al. reported the drug-likeness of 1870 compounds from the EANPDB database [22]. It was found that about 85% of the compounds in this database have drug-like features.

Saldívar-González recently reported a diversity analysis based on physicochemical properties of 154,680 compounds from the Universal Natural Product Database [27] and compare the diversity of such database with compound data sets from the different origins such as 188 morpholine peptidomimetics from a diversity-oriented-synthesis approach, 37 analogs of indinavir from a combinatorial compound library, 27 non-nucleoside DNA-methyltransferase inhibitors from a lead optimization program representative of a target-oriented synthesis approach, and drugs approved for clinical use. The authors concluded that compounds from the extensive NP database are the most diverse, while compounds from the combinatorial library, followed by the TOS set, are the least diverse.

4.2 Molecular scaffolds

Molecular scaffolds also termed “chemotypes,” are the main or core of a molecular structure. Like physicochemical properties discussed in Section 4.1, molecular scaffolds are straightforward to interpret and facilitate communication across disciplines such as NP and medicinal chemists, and chemoinformaticians. Certainly, molecular scaffolds are firmly bound to general concepts in drug discovery, such as "privileged structures" [33] and "scaffold hopping" [34]. There are different ways to generate scaffolds of compound databases systematically and consistently that have been extensively reviewed by Langdon et al. [35].

Systematic analysis of NP databases' scaffold content has been reported revealing the most frequent and distinct scaffolds in the data sets. For instance, Saldívar-González et al. identified the most common scaffolds found in NP from Brazilian diversity [36]. Tran et al. recently discussed the unique molecular scaffolds present in compounds from honey bee and stingless bee propolis. In the same study, authors readily identified that benzene, coumarin, flavan, and flavone are the four scaffolds present in the propolis plus approved drugs and food chemicals [37]. Al Sharie et al. analyzed the scaffold diversity of metabolites from red, brown, and green algae from the Seaweed Metabolite Database, concluding that red algae metabolites are the least diverse while metabolites from green algae are the most diverse [38]. Similarly, González-Medina et al. also recently analyzed the scaffold diversity of cyanobacteria compounds from freshwater and marine sources, concluding that the former are less diverse than metabolites from marine sources. In that work, it was also revealed the most frequent scaffolds found in both data sets and the molecular scaffold common to both compound collections.

4.3 Molecular complexity

Likewise to the notion of molecular similarity [39], molecular complexity is an ambiguous and subjective concept which definition depends on the person's experience and application. For instance, the complexity of a molecule can be assessed in terms of the final structure itself (atom connectivity or three-dimensional shape) or by how difficult to synthesize. Some metrics have been proposed to quantify the complexity of a molecule structure [2]. Similarly, there are different approaches to measure synthetic accessibility [40]. Saldívar-González et al. recently reviewed the three main methods to evaluate chemical complexity and synthetic accessibility, namely graph-theoretical methods, (sub)structure-based approaches, and physicochemical and topological descriptors [27]. In that review, it was noted that the results of the quantitative metrics should coincide with the chemical intuition. Likewise, it is emphasized that simple and easy to compute metrics can provide insightful results [27].

Quantitative assessment of molecular complexity is becoming a crucial factor in drug discovery since it has been associated with increased probabilities to advance in clinical development [41], selectivity, and safety. Recently, it has been proposed that a classical metric to quantity structural complexity such as Fsp^3 is a drug-likeness criterion [42]. Several metrics are straightforward to compute with open-source and free software such as DataWarrior [43,44].

It is well known that NP can have highly complex structures. Likewise, several NP's synthetic accessibility is challenging, particularly when they have several stereocenters. Chemoinformatic methods are useful for quantifying the molecular complexity and comparing it with the complexity of compounds from other sources such as organic synthesis. Over the past few years, the fraction of sp^3 carbon atoms, the number of stereocenters, and other descriptors have been used to compare the molecular complexity of NP from different sources and geographical regions. Prieto-Martínez et al. recently reviewed several analyses [5]. More recent studies include the molecular complexity profiling of the Universal Natural Products Database, NP in $NuBBE_{DB}$, marine NP, cyanobacteria, fungi metabolites, and other data sets [36]. Also, it has been recently analyzed the complexity of compounds from the Seaweed Metabolite Database [38]. The final conclusion of the quantitative comparisons was that, overall, NPs are more complex than drugs approved for clinical use and that NP have large differences in complexity, depending on the particular source. For instance, cyanobacterial metabolites are more complex than fungi metabolites. Also, marine metabolites are more complex than NP available from commercial sources [36].

4.4 Fragments

The overall complex chemical structures of NP make them attractive sources to investigate novel areas of chemical space. Simultaneously, high structural complexity represents a challenge to further obtain them in large quantities needed in advance stages of the drug development face. For this reason, there has been a recent interest in developing synthetic plans to generate semi-synthetic compound libraries inspired by NP [45]. Also, NPs are attractive starting points for fragment-based drug design and generate "pseudo-NPs" [25]. Based on the need to generate fragment libraries based on NP, Chávez-Hernández et al. recently reported an exhaustive fragment library with 205,903 fragments obtained from a sizeable drug-like subset of COCONUT (*vide supra*) [26]. In that work, Chávez-Hernández et al. compared the NP-based fragment collection with a fragment library obtained from more than one million drug-like compounds tested for biological activity and stored in ChEMBL [46], and with a second fragment collection derived from more than 15 million synthetically accessible and novel compounds. It was concluded that there is an extensive diversity of unique fragments derived from NPs that could be used as building blocks for the *de novo* design and synthesis of unique molecules. It was also found that the entire structures and fragments derived from NP are more diverse and have larger structural complexity than the two reference compound collections [26].

4.5 Acid/base profiling

Acidic and basic functional groups of a molecule determine its charge state at different pH values. This, in turn, can affect its solubility, physicochemical properties, affinity for a molecular receptor, pharmacokinetics, and toxicity (*vide infra*). For instance, molecular basicity has been correlated with molecular promiscuity, hERG blockade, and phospholipidosis. The reader is directed to an in-depth discussion by Manallack et al. [47] on the effect of acid/base properties on ADME/Tox properties, drug-target interaction, and drug formulation.

Despite the critical importance of the acid/based properties of molecules in drug discovery, they have been analyzed on a limited basis for NP. In this direction, Santibáñez-Morán et al. discussed the acid/base profile of NP libraries from different geographic locations and sources. The calculated profile was compared to food chemicals and drugs approved for clinical use [48,49]. The NP data sets analyzed were the Universal Natural Product Database, NP from NuBBE$_{DB}$ and BIOFACQUIM databases, marine NP, fungi and cyanobacteria metabolites, and NP from commercial vendors (pure and semi-synthetic). The NP data sets were compared to food chemicals and drugs approved for clinical use. Santibáñez-Morán et al. concluded that, regardless of the different characteristics of the various NP data sets depending on the source of origin (marine, fungi, cyanobacteria) and geographical location (e.g., Brazil, Mexico), NP contain about 45% of neutral compounds. NP also have about 25% of single acids with a pK_a distribution comparable to approved drugs and less than 7% of single bases.

4.6 ADME/Tox profiling

ADME/Tox properties play a significant role in drug discovery [50]. It is estimated that a significant percentage of all drug failures are related to issues with such properties. Therefore, early measurement or at least *in silico* prediction of ADME/Tox properties has an enormous impact on drug development projects. However, accurate prediction of such properties is not a trivial endeavor and but big data and machine learning are largely contributing to improving ADME/Tox predictions [51,52]. A large variety of prediction methods have been implemented into public web servers [53]. For instance, Jia et al. reviewed public online resources to evaluate the ADME and drug-likeness properties of compound data sets [50]. The authors emphasized that quality and updated information in comprehensive databases are key factors for constructing reliable models to evaluate drug-likeness *in silico*. Jia et al. also concluded that online ADME/Tox resources provide useful guidelines to extract rational compounds that match the desirable pharmacokinetic properties or to filter compounds that are not likely to be drugs. Chen et al. have recently pointed out that, despite the fact, there are several web servers and computational models of free access to evaluate ADME/Tox properties, the user should be careful as many of such models have been trained on synthetic compounds, and the applicability domain of NP could be outside those models [7].

Since NP are excellent sources of drug candidates, NP data sets have been profiled for the past 15 years [54]. For instance, Fatima et al. recently discussed a computational ADME/Tox profiling of four phytochemical databases, analyzing different parameters. The authors concluded that 24 compounds have the ADME/Tox properties that can be considered for drug development [55].

Durán-Iturbide et al. reported a comparative *in silico* profile of compounds in BIOFACQUIM with NP from AfroDB, NuBBE$_{DB}$, molecules from the Traditional Chinese Medicine (TCM), and drugs approved for clinical use. The authors of that work found that the absorption and distribution profile of compounds in BIOFACQUIM is similar to approved drugs, while the metabolism profile is comparable to other NP databases. The excretion profile of compounds in BIOFACQUIM was different from approved drugs, but their predicted toxicity profile was comparable [56].

Recently, Simoben et al. reported the ADME/Tox profile of 1870 compounds from the EANPDB database [22]. To that end, the authors employed the free-server pkCSM-pharmacokinetics [57]. It was found that 99.7% of the molecules in EANPDB were predicted to do not interfere with the inhibition of the potassium ion (K^+) channels. It was also found that about 85% of compounds in EANPDB were estimated to do not have no-hepatotoxic or skin sensitization effects [22].

4.7 Global diversity

As commented in the previous sections, different representations of chemical structures (physicochemical properties, sub-structural features, molecular fingerprints, etc.) are used to quantitatively measure compound data sets' chemical diversity. Indeed, chemical representation is one (or perhaps the most important) feature in chemoinformatics (*vide infra*). Therefore, molecular diversity is highly attached to the particular method used to quantify diversity. To reduce molecular diversity dependence with molecular representation has been proposed to combine multiple representations into a single graph termed Consensus Diversity Plot (CDP) [58]. A CDP is a bi-dimensional graph that shows on the same plot four measures of diversity (more metrics of diversity could be added), and it is intended to analyze the "total" or "global" diversity of compound data sets. In current CDP applications, the most common representations to analyze diversity have been scaffold-based, fingerprint, drug-like molecular properties, and the number of compounds (or size) in the data set. Complexity has also been represented. There is a free webserver to generate CDPs [58].

CDPs have been used to analyze the total diversity of NP from Brazil, Mexico, and Panama [36,59,60]. Recently Al Sharie et al. employed the consensus technique to compare metabolites from red, brown, and green algae from the Seaweed Metabolite Database, concluding that, overall, metabolites from green algae are the most diverse [38]. The graphs have also been used to analyze the global diversity of compounds tested with epigenetic targets and synthetic libraries [61]. Further discussion of CDP has been published recently [27].

4.8 Visual representation of chemical space

The concept of "chemical space" has received different definitions. For example, Virshup et al. defined chemical space as "an M-dimensional Cartesian space in which compounds are located by a set of M physiochemical and/or chemoinformatic descriptors" [62]. Such definition emphasizes the dependence of chemical space with molecular representation. Although many quantitative assessments of the structural diversity of compound data sets are linked to the concept of chemical space (e.g., analysis of the profile of the six physicochemical properties of pharmaceutical interest, *vide supra*), the chemical space exploration is usually associated with a visual representation of the multi-dimensional space. To this end, different visualization techniques have been implemented. Among the most common are principal component analysis, self-organizing maps, t-distributed stochastic neighbor embedding (t-SNE), ChemMaps [63], and others extensively reviewed in [64,65]. For example, Olmedo et al. used PCA to generate a comparative chemical space visualization of NP from Panama with compounds in TCM, synthetic molecules and drugs approved for clinical use [59,66].

Recently, Probst et al. proposed the technique Tree Map (TMAP) tuned to visualize high-dimensional chemical spaces [67]. This technique has been used to visualize the chemical space of the NP database BIOFACQUIM with the reference databases ChEMBL and NP assembled from the Universal Natural Products Database, the Natural Products Atlas, and Natural Products in PubChem Substance Database [60]. Visual representation of the chemical space if often used to explore the structure-activity relationships (SAR) or structure multiple-activity relationships (SMARts) [68] of compound data sets systematically and identify valuable "StARs" in chemical space [69].

To illustrate TMAPs, three Natural Product datasets were used: Traditional Chinese Medicine (TCM), with 17,986 compounds [70], BIOFACQUIM with 531 NPs that were isolated and characterized from Mexico [60], and NuBBE$_{DB}$ with 2215 NP from Brazil [71]. Data sets were prepared with the open-source cheminformatics toolkit RDKit (http://www.rdkit.org), version 2020.03.2.0. supporting us from the Standardizer, LargestFragmentChoser, Uncharger, Reionizer, and TautomerCanonicalizer functions implemented in the molecule validation and standardization tool MolVS [72]. All calculations were performed using Python programming language. Curated data sets have standardized chemical structures. Compounds were removed if they had any element other than H, B, C, N, O, F, Si, P, S, Cl, Se, Br, and I. Structures with valence errors were removed from the dataset. Those had multiple compounds (e.g., ionic salts) were split, and the largest component was retained (LargestFragmentChoser function). The remaining compounds were neutralized (Uncharger function) and reionized (Reionizer function) to generate a canonical tautomer (TautomerCanonicalizer function) subsequently. Also, duplicated structures within each database were removed. After curation and preparation, data sets had the following sizes: 17,905 TCM compounds, 2015 NuBBE$_{DB}$ compounds, and 503 BIOFACQUIM compounds.

After data curation, fragments were generated with Retrosynthetic Combinatorial Analysis Procedure (RECAP) that is based on 11 cleavage rules derived from chemical reactions. A molecule is cleaved if this had any of the following bonds: ether, olefin, amide, ester, amine, urea, quaternary nitrogen, aromatic nitrogen-aliphatic carbon, lactam nitrogen-aliphatic carbon, aromatics carbon-aromatic carbon, and sulfonamide [73]. After that, to carry out the visual representation of chemical space using TMAP, Morgan fingerprint with radius 2 (Morgan2, 1024 bits) [74] were generated for all compounds and fragments from data sets of TCM, BIOFACQUIM, and NuBBE$_{DB}$. It was generated a binary vector (0,1) with the possible substructures that each molecule in each database can have. Where 1 is that a molecule has a substructure and 0 is that the molecule does not have. It should be noted that the reference substructures are unique to each database and depend on the chemical environment of the molecule. Later, these binary values (fingerprints) were coded to 1024 HASH functions and 128 prefix trees. After the c-approximate k-nearest neighbors' graph (c-k-NNG) was constructed taking two arguments: *k*, the number of nearest-neighbors to be searched for, and kc, the factor used by the augmented query algorithm. The *k* numbers of compounds closest to a reference compound are grouped. Closeness is measured from the Euclidean distance between each compound, and the closest compounds form groups that were joined through branches, as if it were a tree, hence the name of the graph.

Figs. 3 and 4 show a visual representation of three NP datasets using TMAP for both compounds and fragments, respectively. Both figures have six panels where the three upper panels show the three data sets of NP: TCM in cyan (light gray in print version), BIOFACQUIM in magenta (gray in print version), and NuBBE_{DB} in olive (dark gray in print version). The three lower panels show the overlaps between TCM-BIOFACQUIM in black, TCM-NuBBE_{DB} in navy color (dark gray color in print version), and BIOFACQUIM-NuBBE_{DB} in blue (light gray in print version). Each point in the TMAP graph represents a chemical structure, either compounds or fragments. In structural terms, if several points are separated means that the compounds are different. The compounds are like each other if the points are closer together. For example, Fig. 3 shows that compounds BIOFACQUIM are the most diverse, followed by NuBBE_{DB} and TCM. In addition, the chemical space of the compound data sets is mostly represented by TCM. The largest percentage of overlapping compounds was TCM-NuBBE_{DB} (3.02% overlap), followed by BIOFACQUIM-NuBBE_{DB} (2.81% overlap) and TCM-BIOFACQUIM (0.84% overlap). Whereas, Fig. 4 shows the chemical space of fragments data sets. BIOFACQUIM-NuBBE_{DB} fragments have the largest overlapping (9.09% overlap), followed by TCM-NuBBE_{DB} (5.33% overlap), and TCM-BIOFACQUIM (2.15% overlap). Because the percentage of structures in common between NuBBE_{DB} and BIOFACQUIM fragments (9.09%) is greater

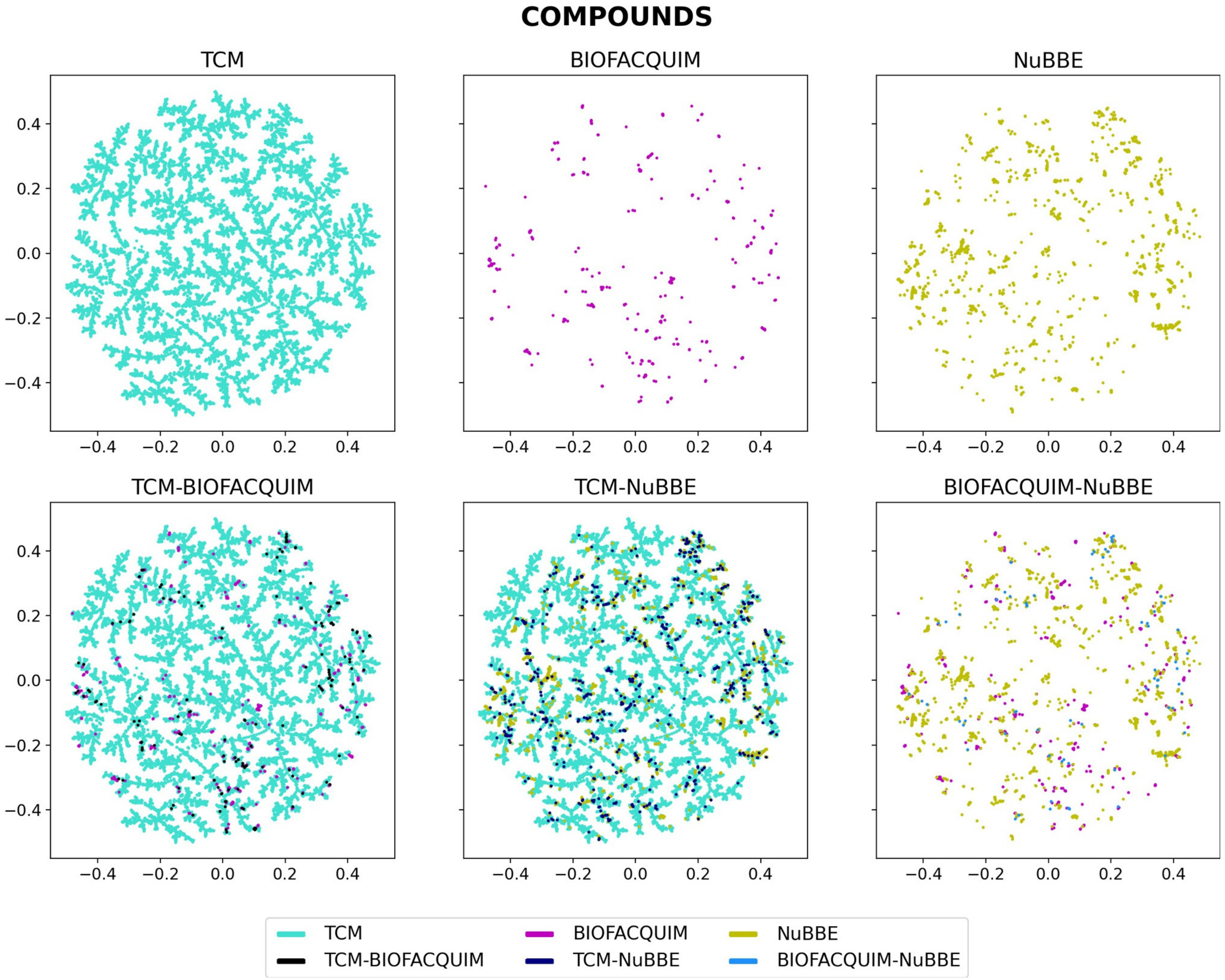


FIG. 3 Visual representation of the chemical space of the compounds datasets using Tree Map (TMAP). The figure shows compounds from each dataset represented in colors: TCM (cyan (light gray in print version)); BIOFACQUIM (magenta (gray in print version)), and NuBBE_{DB} (olive (dark gray in print version)). Overlapping compounds were indicated for TCM-BIOFACQUIM (black); TCM-NuBBE_{DB} (navy (dark gray in print version)), and BIOFACQUIM-NuBBE_{DB} (blue (light gray in print version)).

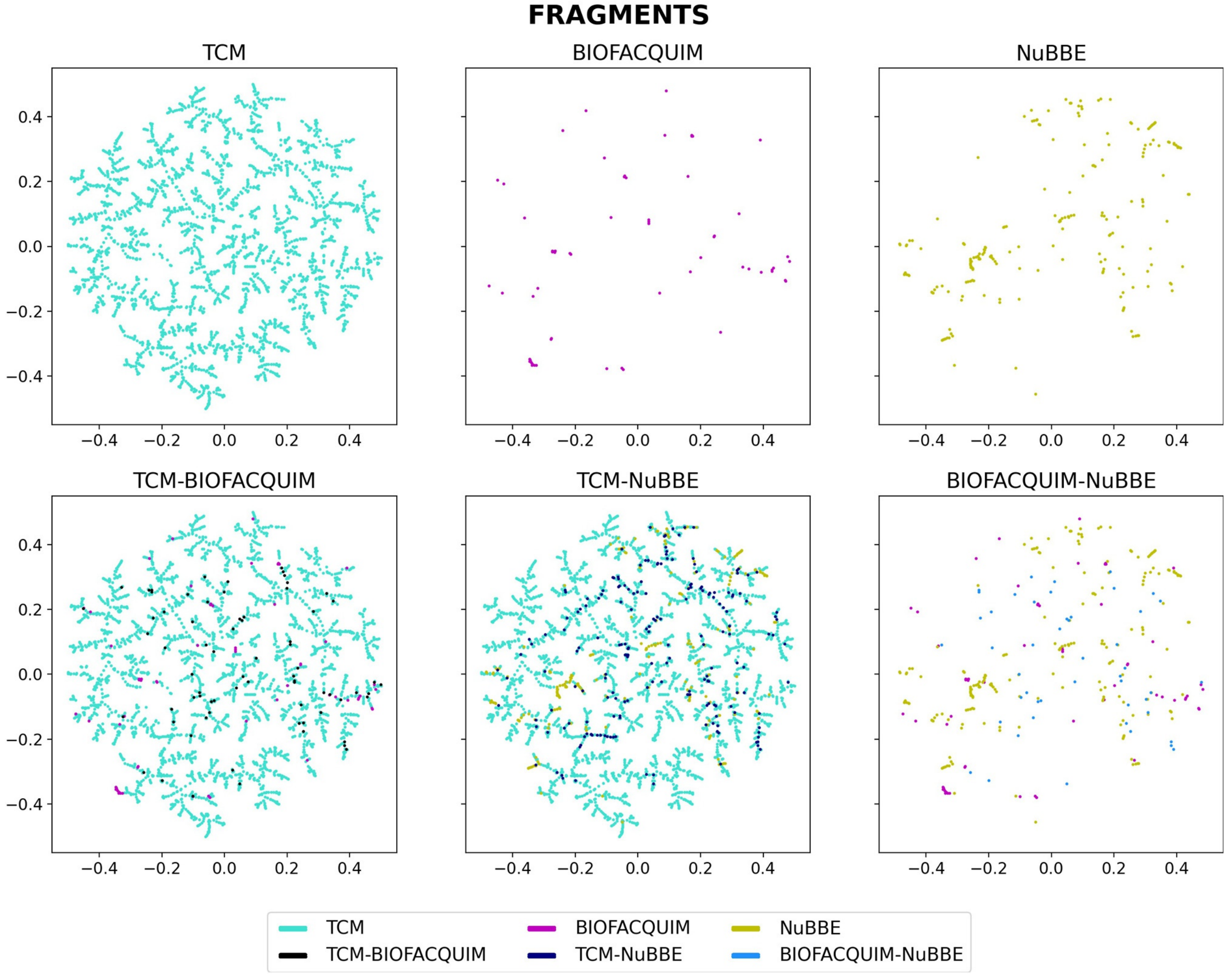


FIG. 4 Visual representation of the chemical space of the fragments datasets using Tree Map (TMAP). The figure shows fragments from each dataset represented in colors: TCM (cyan (light gray in print version)); BIOFACQUIM (magenta (gray in print version)), and NuBBE$_{DB}$ (olive (dark gray in print version)). Overlapping fragments were indicated for TCM-BIOFACQUIM (black); TCM-NuBBE$_{DB}$ (navy (dark gray in print version)), and BIOFACQUIM-NuBBE$_{DB}$ (blue (light gray in print version)).

than their respective compounds (2.81%). It would be advisable to merge both databases into one, but only fragments. Besides, relatively few fragments are obtained, 392 fragments from NuBBE$_{DB}$ and 136 fragments from BIOFACQUIM.

The chemical space of NP from plant, marine, fungi, and other sources has been extensively revised by Saldívar-González et al. [75]. In that review, the authors highlight the variety of properties calculated and different visualization methods of the chemical space. The molecular representations used more frequently to visualize the chemical space are physicochemical properties associated with drug-like features and molecular fingerprints. One of the most frequently used visualization techniques is PCA. In that work, it was concluded that the space of naturally occurring molecules is diverse and vast and that the consistent exploration of the space may have crucial implications not only in drug discovery but also in biodiversity analysis.

In a recent and novel approach, Santibáñez-Morán et al. reported a PCA representation of chemical of seven NP data set from different origins (e.g., marine, fungi, and cyanobacteria metabolites) and geographical regions (Brazil and Mexico) using nine descriptors associated with the acid/based profile [48]. The NP data sets were compared to food chemicals and drugs approved for clinical use. The first two principal components captured 76% of the variance. The visualization of the chemical space and hierarchical clustering of the same nine descriptors revealed that cyanobacteria metabolites are different from the other NP data sets due to mainly the different pK_a distribution of single acids that, in turn, is associated with the low proportion of carboxylic acids. The analysis also showed that semi-synthetic compounds from a commercial vendor are more similar to drugs approved for clinical use [48].

Sánchez-Cruz et al. used the TMAP method to visualize the chemical space of 503 compounds in BIOFACQUIM, 168,030 NP assembled from three large data sets (namely, the Natural Product Atlas, Natural Products in PubChem Substance Database, and Universal Natural Product Database), and 1,667,509 compounds from ChEMBL 25 [46]. TMAP was particularly useful in this case since, as stated above, this approach is suitable to represent visually large data sets as a two-dimensional tree. It was found that compounds in ChEMBL practically defined the biologically relevant chemical space, but this is not evenly populated. In such reference space defined by ChEMBL, NPs cover the same space but more sparsely. In contrast, BIOFACQUIM populates less dense chemical space regions but have compounds similar to the reference data sets [60].

5 Identify active compounds and potential targets

5.1 Computer-aided drug selection

VS or computer-aided drug selection (CADS) can be understood as the computational filtering of molecules to select a reduced number of molecular entities with increased probability to have the desired property (e.g., biological activity). Generally, there are two main types of filtering. One of them is screening a typically large collection of small molecules (e.g., MW below 1000 Da) to select a reduced number of molecules for experimental validation and find hit compounds. The experimental validation is done using biochemical or cell-based assays. Depending on the experimental information available, the filtering can be done using structure- or ligand-based methods. Common approaches are molecular docking for the first and similarity searching.

The second main type of filtering is referred to in the literature target fishing. This is the screening of molecular targets that could potentially interact with a given (or a handful number) of small molecules. This approach is broadly used in NP research. This is because it is frequently desired to uncover potential molecular targets of a given natural compound. This strategy is also quite relevant if the natural compounds under study are part of a mixture of molecules that are used in traditional or folk medicine.

VS has a large impact on NP-based drug discovery, as reviewed in previous publications [7,76,77]. For instance, in a recent view, de Sousa Luis et al. concluded that for 15 years (2003–2018), 230 peer-reviewed articles are reporting VS of NP collections. Most of the applications were focused on the therapeutic areas anticancer, antibacterial, and anti-inflammatory.

VS or CADS can also be facilitated by the increased availability of software and online resources for this purpose [78]. For instance, there are several free online servers that have been developed to identify potential targets of small molecules. Examples are MolTarPred [79], HitPick [80], SwissTargetPrediction [81], and the polypharmacology browser [82], to name a few. It was recently developed Epigenetic Target Profiler that is focused on predicting the potential activity of a given small molecule with a set of 55 epigenetic targets [83]. Epigenetic Target Profiler is part of D-Tools: a set of free online resources to support drug discovery [84].

Hereunder, we elaborate on a recent case study related to the CADS of compounds with potential activity vs. COVID-19. We selected this case study because of its relevance in modern drug discovery.

5.1.1 COVID-19

Coronaviruses (CoVs) are a family of enveloped RNA viruses and have a characteristic crown-shaped appearance caused by the surface glycoproteins that decorate the virus [85]. CoVs are widespread in nature, they can infect several different species, including mammals and birds [86]. Also, it has been suggested that some types of bats are the natural reservoirs of these types of CoVs [87]. CoVs can cause mainly respiratory, gastrointestinal, and hepatic diseases, with neurotropic and neuroinvasive properties in various hosts [88].

A novel CoV provisionally named 2019-nCoV [89], and then renamed as Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) by the International Committee on Taxonomy of Viruses [90], was associated with a cluster of respiratory tract infections named COVID-19 in Wuhan, Hubei Province, China on December 2019 [91,92]. It has rapidly spread across the world, causing that the World Health Organization recognized it as a pandemic in March 2020 [93,94].

SARS-CoV-2 was found to be a positive sense, single-stranded RNA virus belonging to the genus *Betacoronavirus* [85,86]. Its genome contains 14 open reading frames, encoding for 27 proteins: the ORF1 and ORF2 at the 5′-terminal region of the genome encode for 15 non-structural proteins, Fig. 5. The four non-structural proteins that are key enzymes in the viral life cycle are the 3-chymotrypsin-like protease, papain-like protease, helicase, and RNA-dependent RNA polymerase, which are important for virus replication [95]. Initial analyses of SARS-CoV-2 genomic sequences indicate that these four enzymes' catalytic sites could represent antiviral targets due to their highly conserved and high level of sequence similarity with the corresponding SARS and MERS enzymes [96]. The 3′-terminal region of the genome encodes for structural proteins, namely spike, envelope protein, membrane protein, and nucleocapsid, plus eight accessory proteins [97]. The proteins are important because the spike protein governs binding to host cell

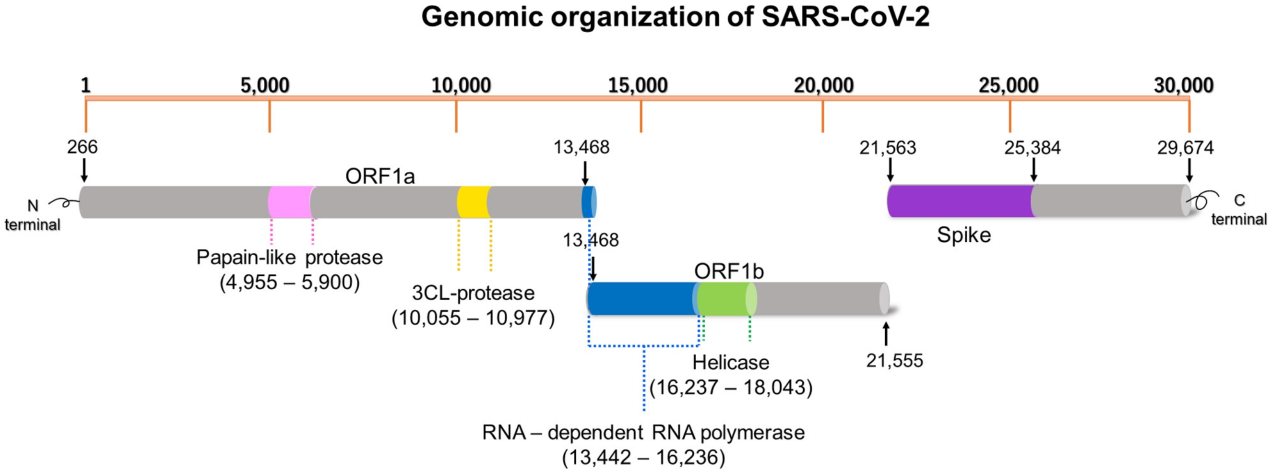

Genomic organization of SARS-CoV-2
1
5,000
10,000
15,000
20,000
25,000
30,000
266
13,468
21,563
25,384
29,674
ORF1a
N terminal
C terminal
Papain-like protease (4,955 – 5,900)
3CL-protease (10,055 – 10,977)
13,468
ORF1b
Spike
Helicase (16,237 – 18,043)
21,555
RNA – dependent RNA polymerase (13,442 – 16,236)

FIG. 5 Genomic organization of SARS-CoV-2 indicating the coding regions for proteins that are key enzymes in the viral life cycle and could be potential drug targets [95].

receptors and virus entry into cells; the membrane protein (M^{pro}) and the envelope protein, facilitates viral assembly; and the N protein, which together with genomic RNA constitutes the nucleocapsid, which is the major antigen in severe acute respiratory syndrome [86,98,99].

Currently, there are no SARS-CoV-2 specific antiviral agents. However, researchers are looking for any possible treatment for this disease. Some journal articles published since December 2019 to date. There are journal articles on potential antiviral drug candidates such as remdesivir, baricitinib, and chloroquine to treat this disease [100,101].

Besides various commercialized antiviral drugs, there are also small molecule compounds currently in research and development that have shown significant inhibitory effects on many key proteins from similar coronaviruses such as SARS-CoV and MERS-CoV. These drug candidates mostly inhibit viral enzymes, including proteases and components for RdRp. Since 3CLpro protease has a high level of sequence homology between SARS-CoV and SARS-CoV-2, inhibitors against 3CLpro of SARS-CoV may also apply to SARS-CoV-2. Compounds, including **benzopurpurin B**, **C-467929**, **C-473872**, **NSC-306711** and **N-65828**, which may inhibit the activity of viral NSP15, poly(U)-specific endoribonuclease, were tested for reduced SARS-CoV infectivity in cultured cells with IC_{50} of 0.2–40 μM.

Because of the emergency of the pandemic, one of the most attractive strategies to identify the potential treatment of COVID-19 is to screen compound data sets to identify potential drug candidates hitting one or more of the main molecular targets or directly inhibiting the SARS-CoV-2. Other rational drug discovery approaches such as *de novo* design can be explored and reported in the near future. To rapidly screen the vast chemical space [64], chemoinformatic analysis [27] and computational filtering of compound databases or CADS is a logical step to shorten the times and cost of massive experimental screening. Molecular docking, followed by molecular dynamics, is one of the most common techniques used thus far in VS. Other general approaches are similarity searching and QSAR modeling [102].

Thus far, perhaps the most common databases used in CADS campaigns are drugs approved for clinical use, under clinical investigation, or withdrawn from the market. This strategy is to pursue a drug repurposing strategy. Arguably the second most explored regions of chemical space are commercial compounds from ZINC 15 [103]. Other regions of the chemical space explored in VS are represented by small sets of NP and food chemicals. Our research group recently conducted a first large screening of food chemicals and molecules in the dark chemical matter using a combination of ligand- and structure-based screening [104]. Some chemical space regions have been explored on a limited basis such as on-demand libraries [105] and large databases of NP.

5.1.2 Natural products with potential activity with SARS-CoV-2

The broad availability of NP data sets in the public domain and the rapid structure information of potential drug targets of SARS-CoV-2 prompted the quick search in NP datasets (in-house, commercial or public collections) to identify potential drug candidates or starting points for optimization [106]. In some cases, food chemical databases have been evaluated, of which a large number are from natural sources [104]. Consequently, several reports emerged in the literature, most of them published as preprints, and several published in peer-reviewed journals. Table 1 summarizes representative studies published in peer-reviewed journals.

From the compilation of 16 works related to possible hits against SARS-CoV-2 (Table 1) of NP, it was observed that the most frequently used computational approaches are molecular docking, molecular dynamics, ADME/Tox predictions, and pharmacophore modeling. It should be

TABLE 1 Examples of computational screening on natural products against SARS-CoV-2.

Target	Natural products	Computational approaches	Hits[a]	Reference
M^{Pro}	52 compounds. Alkaloids, terpenoids, polyphenolic compounds, peptides.	Molecular docking, ADME/Tox, molecular dynamics.	One	[107]
M^{Pro}	88 compounds from the literature.	QSAR, molecular docking.	13	[108]
ACE2, $3CL^{pro}$, RdRp, Spike protein	38 compounds contained in Rhizoma Polygonatum from TCMSP (Traditional Chinese Medicine Systems Pharmacology Database).	Network pharmacology, molecular docking.	Two	[109]
$3CL^{pro}$ (6w63)	Narcissin flavonoid that belongs to monomethoxyflavone derivative.	Molecular docking.	One	[110]
Multi-target: including RdRp, $3CL^{pro}$, exoribonuclease, endoribonuclease.	4570 natural compounds from NPASS Database.	Molecular docking, molecular dynamics, ADME.	Three	[111]
Nsp10/nsp16, endoribonuclease, ADP ribose phosphate,$3CL^{pro}$.	2755 bioactive molecules of coumarin derivatives from PubChem	Virtual screening, molecular docking, ADME/Tox, molecular dynamics.	25 (five compounds for each receptor)	[112]
$3CL^{pro}$	3000 compounds from the compound library of Natural Products Research Laboratories (NPRL).	Molecular docking.	9	[113]
M^{pro}	14,064 compounds from Marine Natural Product (MNP) library.	Pharmacophore model, molecular docking, molecular dynamics.	17	[114]
Transmembrane protease serine 2 (TMPRSS2)	30,927 compounds from the natural compounds library Natural Product Activity and Species Source (NPASS).	Pharmacophore model, molecular docking, ADME/Tox.	12	[115]
M^{pro}	100+ natural compounds and synthetic analogs from in-house library.	Virtual screening, molecular docking, molecular dynamics.	Four	[116]
M^{pro}	Natural product databases: ZINC (120,720 molecules), SNP (274,363 molecules), and MNP (1464 molecules).	Pharmacophore model, molecular docking, molecular dynamics, ADME.	Six	[117]
M^{pro}	1611 natural compounds from Selleck database.	Virtual screening, molecular docking, molecular dynamics, ADME/Tox.	Four	[118]
M^{pro}	Natural products from Sigma-Aldrich plant profiler library.	Molecular docking, molecular dynamics	Six	[119]
Nsp1	2300 from DRUGBANK and 300,000 small molecules from Supernatural II database.	Virtual screening, docking, molecular dynamics, ADME.	Four NP, remdesivir, edoxudine.	[120]
$3CL^{pro}$, Nsp9, Spike receptor (ecto-domain and HR2 domain)	27 plant metabolites belonging to different classes from the PubChem database.	Molecular docking, ADME.	Three	[121]
$3CL^{pro}$, PL^{pro}	Natural products from ZINC natural product database and FDA-approved drug from ZINC drug database	Pharmacophore model based virtual screening, molecular dynamics, docking, ADME/Tox.	12 natural products, nelfinavir and tipranavir.	[122]

[a] *Number of computational hits.*

noted that pharmacophore modeling and/or molecular docking is usually used in combination with molecular dynamics and predictions of ADME/Tox properties. As it is well-known, molecular docking enables predicting the binding mode and conformation of a molecule with a molecular target and aids the identification of compounds with biological activity. ADME/Tox profiling aims to consider in advance the ADME/Tox properties for the proposed molecule's active substances and subsequently perform the in vitro assays. In the context of the studies performed, molecular dynamics simulations aimed to assess the stability of the possible inhibitor within the binding pocket and account for the flexibility of the receptor.

Table 2 summarizes the number of studies focused on each molecular target of SARS-CoV-2 and the number of proposed or computational hits. Most of the studies published at the time of writing this manuscript (December 2020) have been directed to M^{pro}.

From the studies surveyed, it was possible to find that the protein with the highest number of studies was M^{pro} and, not surprisingly, the largest number of computational hits. As elaborated above, M^{pro} is the main protease of SARS-CoV-2, and it is a promising target because there are no homologous proteins in humans in addition to having a direct impact on the natural life cycle of SARS-CoV-2 [104]. One of the computational hits to be highlighted from this search was dieckol that showed an experimental inhibitory activity of M^{pro} with IC_{50} of 2.7 μM [123].

TMPRSS2, or serine protease 2, is a transmembrane host enzyme that facilitates the entry of viral particles into host cells, therefore, inhibition of this protein blocks virus's interaction with ACE2 [115]. TMPRSS2 was the second most pursued target in the survey of VS. One of the computational hits against this target was geniposide that showed a better-predicted affinity with TMPRSS2 as compared to the reference compound, camostat.

RdRp and the spike protein are important molecular targets associated with the invasion and replication of SARS-CoV-2 (*vide supra*). The spike protein is the first protein to interact with the host cell through ACE2 (angiotensin-converting enzyme-2), propitiating invasion. In turn, RdRp is an RNA-dependent RNA polymerase, therefore, a molecule that has inhibitory activity against RdRp can satisfactorily interfere in the life cycle of the coronavirus. In the VS, the asiatic acid was identified as a potential inhibitor of the S protein. However, there are no *in vitro* studies that demonstrate its efficacy against SARS-CoV-2.

TABLE 2 Number of studies and hits found per target protein.

Target	Number of studies	Number of computational hits
M^{pro}	14	63
RdRp	2	2
Spike	2	1
TMPRSS2	1	10
Nsp 10/16	1	5
Nsp 1	1	3
Nsp 9	1	4

Non-structural proteins (Nsp1–16) have functions of viral replication and interaction with the host cell. Nsp1 has the function of cellular degradation of cellular mRNA and inhibition of IFN signaling; Nsp9 functions by dimerization and binding of RNA; Nsp10/16 is associated with the negative regulation of the innate immune response. In the computational screening, glycyrrhizic acid was identified as a potential inhibitor of Nsp1. This compound is approved by the FDA for its use as a sweetener, and antiviral activity has been reported, inhibiting replication and regulation of the immune response. Similar to the other computational hits surveyed at the time of writing this chapter (December 2020), there are not yet reports of experimental studies demonstrating its efficacy against SARS-CoV-2.

Some of the VS studies surveyed also seek to propose a multi-target design. The objective of a multi-objective design is for a molecule to interact efficiently with different target proteins of SARS-CoV-2. An example computational hit of these studies was 2,3,4-trihydroxybenzoic acid which showed potential inhibitory activity against five target proteins (RdRp, 3CLpro, exoribonuclease, endoribonuclease, methyltransferase). However, it still lacks *in vitro* experimental assays [111].

Table 3 shows the chemical structures of representative NP hits that have been proposed as potential drug candidates for the treatment of COVID-19.

One of the significant challenges that remain to be solved is the high mutation rate of viruses. The drugs are designed for a unique binding model to the protein of interest, a high mutation rate can greatly affect possible pharmacological therapies and cause setbacks to the investigation.

To carry out the *in silico* studies, the chemical structures of the NP were retrieved from different databases summarized in Table 4. Most of the compounds studied come mainly from plants and herbs, followed by marine species microorganisms' products.

6 Elucidate mechanism of action: Molecular modeling

Molecular modeling has been critical to explore the SAR of NP at molecular level [5]. Hereunder we discuss a recent case study focused on the characterization of a dietary component as an inhibitor of a major epigenetic target that has been pursued by our and other research groups [124,125].

TABLE 3 Representative NP hits that have been proposed as potential drug candidates for the treatment of COVID-19.

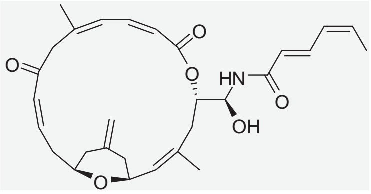
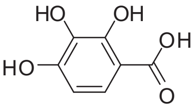
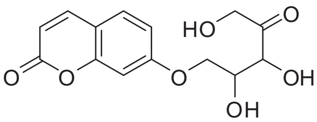
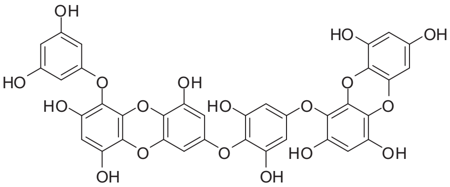
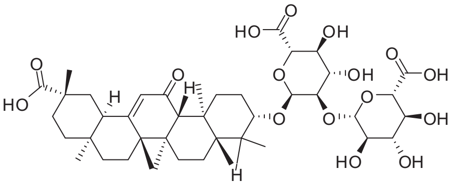
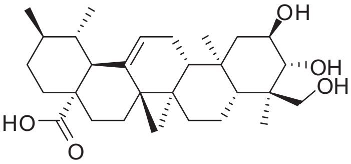
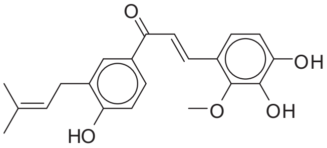
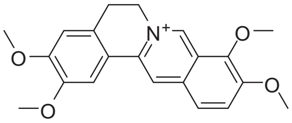
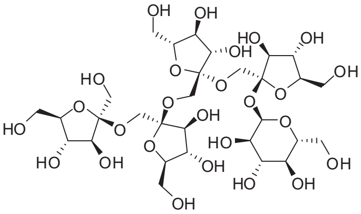
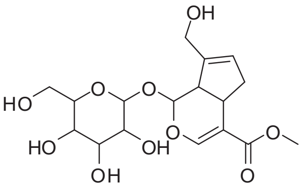
Molecular structure	ID	Target	Sources	Ref.
	10255241 PubChem 70951814 ChemSpider	M^{pro}	Terpenoid from marine sponge *Cacospongia mycofijiensis*.	[107]
	11874 PubChem 11381 ChemSpider	Multi-target: RdRp, $3CL^{pro}$, exoribonuclease, endoribonuclease, methyltransferase	Phenol compound isolated from *Pachysandra terminalis*.	[111]
	101223868 PubChem	Nsp10/nsp16, Main Protease and endoribonuclease.	Coumarin derivates known as natural coumarins from *Ruta graveolens*, *Polygala fruticose*, *Seseli sessiliflorum*.	[112]
	3008868 PubChem 2278311 ChemSpider	M^{pro}	Belonging to the family of phlorotannins, isolated in the brown algae *Ecklonia cava.*	[114]
	14982 PubChem 14263 ChemSpider	Nsp1	Extracted from the root of the licorice plant; *Glycyrrhiza glabra*.	[120]
	119034 PubChem 106361 ChemSpider	$3CL^{pro}$, Nsp9, Spike receptor (ecto-domain)	Extracted from *Centella asiatica*	[121]
	10473311 PubChem	$3CL^{pro}$, PL^{pro}	Present in extracts of *Glycyrrhiza glabra*, also known as licorice	[122]
	S3769 Selleck database 19009 PubChem	M^{pro}	Palmatine is a naturally occurring isoquinoline alkaloids found in traditional Chinese medicine.	[118]
	SN00293542 Supernatural Product (SNP) 3085157 PubChem 2342114 ChemSpider	M^{pro}	Onions, asparagus, and edible burdock, *Morinda officinalis*.	[117]

TABLE 3 Representative NP hits that have been proposed as potential drug candidates for the treatment of COVID-19—cont'd

Molecular structure	ID	Target	Sources	Ref.
	NPC306344 Natural Product Activity and Species Source 12004581 PubChem	Transmembrane protease serine 2 (TMPRSS2)	One of the major iridoid glycosides of gardenia fruit, It is present in nearly 40 species belonging to various families.	[115]

6.1 DNA methyltransferases (DNMT): Identification of a natural product

The word "epigenetics" is rooted in Waddington and Nanney's work, where it was initially defined to denote a cellular memory, persistent homeostasis in the absence of an original perturbation, or an effect on cell fate not attributable to changes in DNA [126]. However, "epigenetics" is now used with multiple meanings, for instance, to describe the heritable phenotype (cellular memory) without modification of DNA sequences [127], or the mechanism in which the environment conveys its influence to the cell, tissue or organism [128]. Regardless of the multiple definitions, the interest in epigenetic drug discovery increases, as revealed by the multiple approved epigenetic drugs or compounds in clinical development for epigenetic targets [129].

DNMTs are one of the main epigenetic modifiers. This enzyme family is responsible for promoting the covalent addition of a methyl group from *S*-adenosyl-L-methionine (SAM) to the 5-carbon of cytosine, mainly within CpG dinucleotides, yielding *S*-adenosyl-*L*-homocysteine (SAH) [130]. DNMT1, DNMT3A, and DNMT3B participate in DNA methylation in mammals to regulate embryo development, cell differentiation, gene transcription, and other normal biological functions. Abnormal functions of DNMTs are associated with tumorigenesis and other diseases [130,131]. Despite the fact there are two DNMT inhibitors approved for clinical use, both azacitidine and decitabine have poor bioavailability, low specificity, and instability in physiological conditions plus toxicity. Thus, it has been the interest of our [124,132–138] and other research groups [139–146] to identify DNMT inhibitors with novel chemical scaffolds. Inhibition of DNMTs is a major topic of research not only because of its potential therapeutic benefits but also to understand the essential mechanisms of epigenetic events in cells.

Fig. 6 shows the chemical structures of representative DNMT inhibitors or compounds with DNA demethylation activity from the natural origin [3,147–151]. Evidence suggests that environmental factors and nutrients play a major role in establishing epigenetic mechanisms, including irregular DNA methylation patterns. Thus, a regular uptake of DNA demethylating agents is hypothesized to have a chemopreventive effect [152].

Due to the known activity of polyphenols and previous evidence of theaflavin activity with DNMT3A [153], we hypothesized that theaflavin, a dietary component (Fig. 6) found in back tea, is an inhibitor of DNMT1 and DNMT3B. To test this hypothesis, we purchased theaflavin

TABLE 4 Databases of natural products used in the virtual screening toward SARS-CoV-2 molecular targets.

Database	Number of molecules	Website
TCMSP (Traditional Chinese Medicine Systems Pharmacology Database and Analysis Platform)	29,384	https://tcmspw.com/load_intro.php?id=42
NPASS (Natural Products Activity & Species Source Database)	35,032	http://bidd.group/NPASS/index.php
MNP (Marine Natural Products)	14,064	http://docking.umh.es/chemlib/mnplib
SNP (Super Natural Products)	27,363	http://bioinf-applied.charite.de/supernatural_new/index.php?site=compound_input
ZINC natural database	120,720	http://zinc15.docking.org/substances/subsets/natural-products/
Selleck Natural Product Library	2370	https://www.selleckchem.com/screening/natural-product-library.html
DrugBank	11,483	https://go.drugbank.com/
PubChem	109 M	https://pubchem-ncbi-nlm-nih-gov.pbidi.unam.mx:2443/

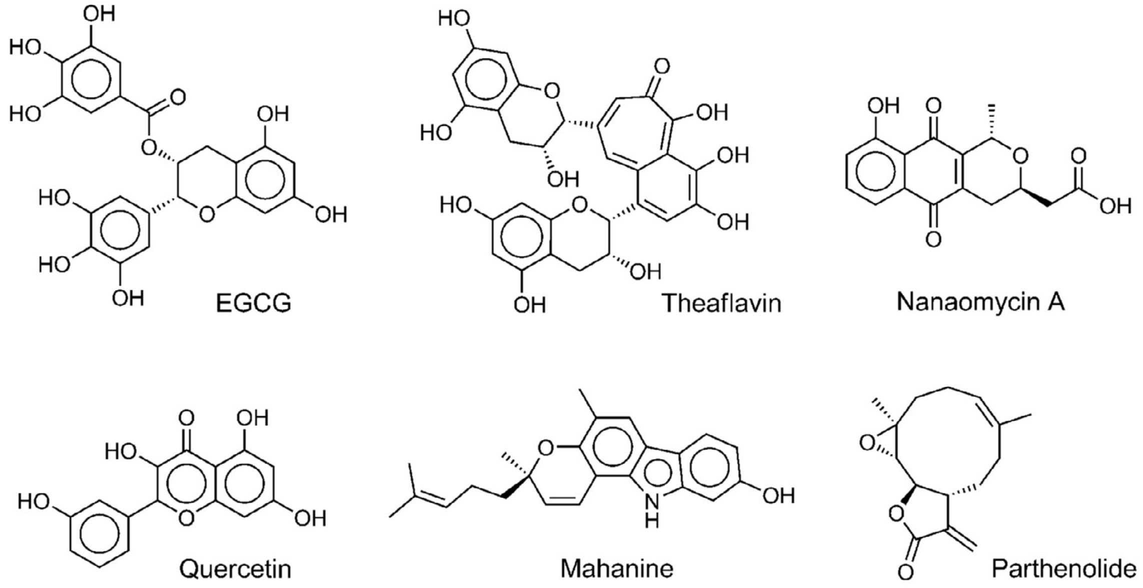


FIG. 6 Representative natural products as inhibitors of DNA methyltransferases.

from a commercial vendor (TargetMol) and performed biochemical assays as summarized in the next subsection and described in detail elsewhere [154].

6.1.1 Biochemical DNMT assays

The inhibition of the enzymatic activity of DNMT1, DNMT3B, and DNMT3B/3L was tested using the HotSpot^{SM} platform for methyltransferase assays from Reaction Biology Corporation [155]. HotSpot^{SM} is a low volume radioisotope based assay that uses tritium-labeled AdoMet (^{3}H-SAM) as a methyl donor. Theaflavin diluted in dimethyl sulfoxide was added using acoustic technology into enzyme/substrate mixture in the nano-liter range. The reactions were initiated by the addition of ^{3}H-SAM, and incubated at 30 °C. Total final methylations on the substrate (Poly dI-dC in DNMT1 assay, and Lambda DNA in DNMT3B; DNMT3B/3L assay) were identified by a filter binding method implemented in Reaction Biology. Data analysis was conducted with Graphed Prism software (La Jolla, CA) for curve fits. Reactions were conducted at 1 μM of SAM. SAH was used as a standard positive control. Theaflavin was tested first with DNMT1 and DNMT3B at one 100 μM concentration in duplicate. Then, it was tested in 10-dose IC_{50} (effective concentration to inhibit DNMT1, DNMT3B and DNMT3B/L activity by 50%) with a three-fold serial dilution starting at 100 μM.

At single-dose concentration, theaflavin showed detectable inhibition of DNMT1 and DNMT3B (65% and 33% inhibition, respectively). In the dose-response evaluation, the dietary component had an IC_{50} value of 85.33 μM with DNMT1 and had IC_{50} values >100 μM when tested with DNMT3B and DNMT3B/3L. SAH (the positive control) had an IC_{50} value of 0.26 μM with DNMT1.

Theaflavin is a natural product polyphenol found in green and black tea and coffee (*vide supra*) with previously measured enzymatic inhibitory activity of DNMT3A [153]. However, there were no previous reports on its inhibitory potential of DNMT1 and DNMT3B. In that work, it was evaluated the activity of the dietary component with both enzymes. Based on the results and previous publications, it was proposed that theaflavin could be a selective inhibitor of DNMT1. Despite the fact its IC_{50} is high (85.3 μM, under the assay conditions of that work), this dietary component could contribute to the modulation of DNMT1. Interestingly, it has been proposed that the modulation of normal levels of DNMT could be conveniently achieved through the dietary uptake of food chemicals (or other "safe" NP). A prominent example of this hypothesis has been suggested for the polyphenol compound from green tea, EGCG (Fig. 6), which has been proposed to inhibit DNMT1 and reactivate methylation-silenced genes in cancer [152].

6.1.2 Molecular docking

Of the different mechanisms described to inactivate DNMT activity [156] we hypothesized that theaflavin is a SAM competitor. Theaflavin was docked with DNMT1 (*vide infra*) using the program Molecular Operating Environment (MOE), version 2018.08 [157]. Its chemical structure was built with MOE. The docking was carried out with the crystal structure of the catalytic domain of DNMT1 obtained from the Protein Data Bank [158] PDB ID: 4WXX [159]. This crystal structure is in complex with SAH and has a resolution of 2.62 Å. The structure of the protein was prepared with the "QuickPrep" tool of MOE using the parameters established by default, which help to remove the molecules of structural water and add

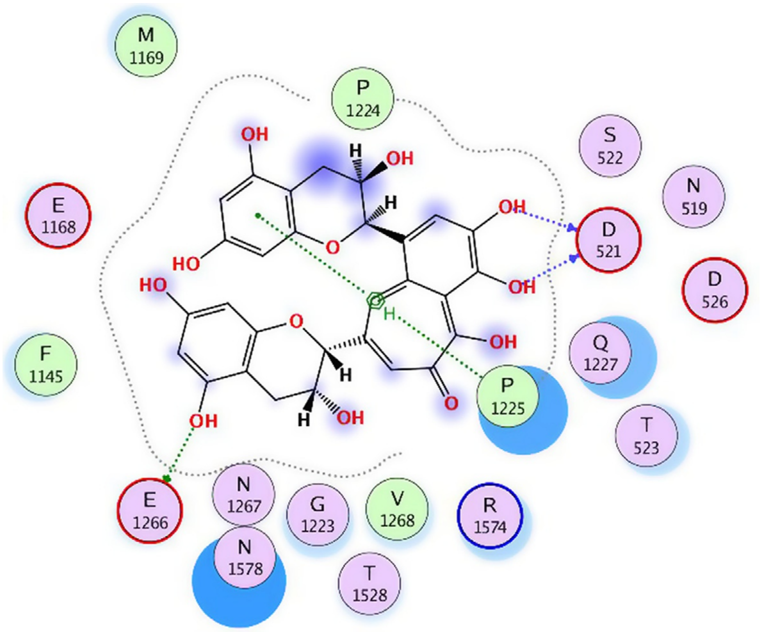


FIG. 7 Two dimensional map showing the predicted binding mode of theaflavin with DNA methyltransferase 1 (DNMT1).

hydrogens atoms to the protein. In this process, the co-crystallized SAH was removed for the binding site to realize a direct docking. The docking was done using default parameters in MOE. Before docking theaflavin, the docking protocol was validated by re-docking the SAH obtaining a root-mean-square deviation of 1.3 Angstroms and a docking score of −8.96 kcal/mol.

Fig. 7 shows the 2D interaction map and the 3D binding mode of the predicted binding mode between the human catalytic domain of DNMT1 and theaflavin. Docking simulations help to rationalize at the molecular level the experimental results of theaflavin.

7 Concluding remarks

NPs have a critical role in drug discovery and development either providing bioactive compounds that have become drugs approved for clinical use or inspiring the synthesis of small molecules. Chemoinformatics has been crucial in the investigation of NP and gives rise to the emerging sub-discipline "natural products informatics." Informatic methods have been key for the profiling and diversity analysis of NP data sets in terms of physicochemical properties, structural and sub-structures (functional groups and fragments), molecular fingerprints, complexity, acid/base properties and ADME/Tox profile. These studies have quantified, for instance, the structural complexity of NP and revealed that to all NP are alike; some have distinct structural features depending on the source. For instance, it is remarkable the unique characteristics of cyanobacteria metabolites compared to the structures of NPs from other sources. Many visualization methods have been developed and used to analyze the diversity and distribution of NP in chemical space. It is important to point out the "chemical space" is not static and unique, but relative to the descriptors used to generate the chemical space. Thus far, NPs have been represented with physicochemical properties, structural features and fingerprints, ADME/Tox, acid/base profile. Computational approaches also have an outstanding contribution to uncover bioactive NP. It is concluded that VS of NP compound libraries is an example of computer-aided drug selection. VS-CADS of NP is an excellent example of the benefits of combining to main sources of hit compounds (computational approaches and NP-based drug discovery). Given the emergency imposed by the current COVID-19 pandemic, many research groups are doing CADS-VS of NP compound libraries. Several NPs have been proposed at the time of writing this chapter (October 2020) and are awaiting experimental confirmation. Although potential molecules with inhibitory activity against some SARS-CoV-2 proteins of interest have been sought and found, it is necessary to perform *in vitro* and *in vivo* tests to corroborate the activity of computational hits selected from computational filtering of compound databases. Computational approaches have also largely contributed to the identification of compounds with activity for epigenetic targets. Also, molecular modeling methodologies play a significant role in understanding at the molecular level, the biological activity of small molecules. This manuscript illustrates that molecular docking has helped understand the in vitro activity of theaflavin that is a dietary component that inhibits DNMT1 and does not show significant inhibition of DNMT3B.

Acknowledgments

This work was supported by DGAPA, UNAM, *Programa de Apoyo a Proyectos de Investigación e Innovación Tecnológica* (PAPIIT), grant IN201321. A.L.-C.H. is thankful to the National Council of Science and Technology (CONACyT, Mexico) for the granted scholarship number 847870. Insightful contributions and discussions with Fernanda I. Saldívar-González, K. Eurídice Juárez-Mercado, B. Angélica Pilón-Jiménez, Mariel Rico-Hidalgo, Norberto Sánchez-Cruz, and other members of the DFACQUIM research group are greatly acknowledged.

References

[1] Newman DJ, Cragg GM. Natural products as sources of new drugs over the nearly four decades from 01/1981 to 09/2019. J Nat Prod 2020;83:770–803. https://doi.org/10.1021/acs.jnatprod.9b01285.

[2] Méndez-Lucio O, Medina-Franco JL. The many roles of molecular complexity in drug discovery. Drug Discov Today 2017;22:120–6. https://doi.org/10.1016/j.drudis.2016.08.009.

[3] Saldívar-González FI, Gómez-García A, Chávez-Ponce de León DE, Sánchez-Cruz N, Ruiz-Rios J, et al. Inhibitors of DNA methyltransferases from natural sources: a computational perspective. Front Pharmacol 2018;9:1144. https://doi.org/10.3389/fphar.2018.01144.

[4] Medina-Franco JL. Discovery and development of lead compounds from natural sources using computational approaches. In: Mukherjee P, editor. Evidence-based validation of herbal medicine. Elsevier; 2015. p. 455–75.

[5] Prieto-Martínez FD, Norinder U, Medina-Franco JL. Cheminformatics explorations of natural products. In: Kinghorn A, Falk H, Gibbons S, Kobayashi J, Asakawa Y, et al., editors. Progress in the chemistry of organic natural products. Cham: Springer; 2019.

[6] Koulouridi E, Valli M, Ntie-Kang F, Bolzani VDS. A primer on natural product-based virtual screening. Phys Sci Rev 2019;4:20180105. https://doi.org/10.1515/psr-2018-0105.

[7] Chen Y, Kirchmair J. Cheminformatics in natural product-based drug discovery. Mol Inf 2020;39:2000171. https://doi.org/10.1002/minf.202000171.

[8] Medina-Franco JL, Saldívar-González FI. Cheminformatics to characterize pharmacologically active natural products. Biomolecules 2020;10. https://doi.org/10.3390/biom10111566.

[9] Martinez-Mayorga K, Madariaga-Mazon A, Medina-Franco JL, Maggiora G. The impact of chemoinformatics on drug discovery in the pharmaceutical industry. Exp Opin Drug Discov 2020;15:293–306. https://doi.org/10.1080/17460441.2020.1696307.

[10] Saldívar-González FI, Aldas-Bulos VD, Medina-Franco JL, Plisson F. Natural product drug discovery in the artificial intelligence era. Chem Sci 2022. https://doi.org/10.1039/d1sc04471k [in press].

[11] López-López E, Bajorath J, Medina-Franco JL. Informatics for chemistry, biology, and biomedical sciences. J Chem Inf Model 2021;61:26–35. https://doi.org/10.1021/acs.jcim.0c01301.

[12] Prieto-Martínez FD, López-López E, Eurídice Juárez-Mercado K, Medina-Franco JL. Computational drug design methods—current and future perspectives. In: Roy K, editor. In silico drug design. Academic Press; 2019. p. 19–44 [chapter 2].

[13] Sánchez-Cruz N, Medina-Franco JL. Epigenetic target fishing with accurate machine learning models. J Med Chem 2021;64:8208–20. https://doi.org/10.1021/acs.jmedchem.1c00020.

[14] Capecchi A, Probst D, Reymond J-L. One molecular fingerprint to rule them all: drugs, biomolecules, and the metabolome. J Cheminform 2020;12:43. https://doi.org/10.1186/s13321-020-00445-4.

[15] Gasteiger J. Chemistry in times of artificial intelligence. ChemPhysChem 2020;21:2233–42. https://doi.org/10.1002/cphc.202000518.

[16] Chen Y, de Bruyn Kops C, Kirchmair J. Data resources for the computer-guided discovery of bioactive natural products. J Chem Inf Model 2017;57:2099–111. https://doi.org/10.1021/acs.jcim.7b00341.

[17] Medina-Franco JL. Towards a unified Latin American natural products database: LANaPD. Future Sci OA 2020;6:FSO468. https://doi.org/10.2144/fsoa-2020-0068.

[18] Yongye AB, Waddell J, Medina-Franco JL. Molecular scaffold analysis of natural products databases in the public domain. Chem Biol Drug Des 2012;80:717–24. https://doi.org/10.1111/cbdd.12011.

[19] Fullbeck M, Michalsky E, Dunkel M, Preissner R. Natural products: sources and databases. Nat Prod Rep 2006;23:347–56. https://doi.org/10.1039/b513504b.

[20] Sorokina M, Steinbeck C. Review on natural products databases: where to find data in 2020. J Chem 2020;12:20. https://doi.org/10.1186/s13321-020-00424-9.

[21] Sorokina M, Merseburger P, Rajan K, Yirik MA, Steinbeck C. Coconut online: collection of open natural products database. J Chem 2021;13:2. https://doi.org/10.1186/s13321-020-00478-9.

[22] Simoben CV, Qaseem A, Moumbock AFA, Telukunta KK, Günther S, et al. Pharmacoinformatic investigation of medicinal plants from East Africa. Mol Inf 2020;39:2000163. https://doi.org/10.1002/minf.202000163.

[23] Nguyen-Vo TH, Le T, Pham D, Nguyen T, Le P, et al. Vietherb: a database for Vietnamese herbal species. J Chem Inf Model 2019;59:1–9. https://doi.org/10.1021/acs.jcim.8b00399.

[24] Madariaga-Mazón A, Naveja JJ, Medina-Franco JL, Noriega-Colima KO, Martinez-Mayorga K. Dianat-Db: a molecular database of antidiabetic compounds from medicinal plants. RSC Adv 2021;11:5172. https://doi.org/10.1039/d0ra10453a.

[25] Christoforow A, Wilke J, Binici A, Pahl A, Ostermann C, et al. Design, synthesis, and phenotypic profiling of pyrano-furo-pyridone pseudo natural products. Angew Chem Int Ed 2019;58:14715–23. https://doi.org/10.1002/anie.201907853.

[26] Chávez-Hernández AL, Sánchez-Cruz N, Medina-Franco JL. A fragment library of natural products and its comparative chemoinformatic characterization. Mol Inf 2020;39:2000050. https://doi.org/10.1002/minf.202000050.

[27] Saldívar-González FI, Medina-Franco JL. Chemoinformatics approaches to assess chemical diversity and complexity of small molecules. In: Trabocchi A, Lenci E, editors. Small molecule drug discovery. Elsevier; 2020. p. 83–102 [chapter 3].

[28] Lipinski CA. Lead- and drug-like compounds: the rule-of-five revolution. Drug Discov Today Technol 2004;1:337–41. https://doi.org/10.1016/j.ddtec.2004.11.007.

[29] Veber DF, Johnson SR, Cheng HY, Smith BR, Ward KW, et al. Molecular properties that influence the oral bioavailability of drug candidates. J Med Chem 2002;45:2615–23. https://doi.org/10.1021/jm020017n.

[30] Feher M, Schmidt JM. Property distributions: differences between drugs, natural products, and molecules from combinatorial chemistry. J Chem Inf Comput Sci 2003;43:218–27. https://doi.org/10.1021/ci0200467.

[31] Singh N, Guha R, Giulianotti MA, Pinilla C, Houghten RA, et al. Chemoinformatic analysis of combinatorial libraries, drugs, natural products, and molecular libraries small molecule repository. J Chem Inf Model 2009;49:1010–24. https://doi.org/10.1021/ci800426u.

[32] Pilon-Jimenez BA, Saldivar-Gonzalez FI, Diaz-Eufracio BI, Medina-Franco JL. Biofacquim: a Mexican compound database of natural products. Biomolecules 2019;9:31. https://doi.org/10.3390/biom9010031.

[33] Evans BE, Rittle KE, Bock MG, DiPardo RM, Freidinger RM, et al. Methods for drug discovery: development of potent, selective, orally effective cholecystokinin antagonists. J Med Chem 1988;31:2235–46. https://doi.org/10.1021/jm00120a002.

[34] Schneider G, Neidhart W, Giller T, Schmid G. Scaffold-hopping by topological pharmacophore search: a contribution to virtual screening. Angew Chem Int Ed 1999;38:2894–6. https://doi.org/10.1002/(SICI)1521-3773(19991004)38:19<2894::AID-ANIE2894>3.0.CO;2-F.

[35] Langdon SR, Brown N, Blagg J. Scaffold diversity of exemplified medicinal chemistry space. J Chem Inf Model 2011;51:2174–85. https://doi.org/10.1021/ci2001428.

[36] Saldívar-González FI, Valli M, Andricopulo AD, da Silva Bolzani V, Medina-Franco JL. Chemical space and diversity of the Nubbe database: a chemoinformatic characterization. J Chem Inf Model 2019;59:74–85. https://doi.org/10.1021/acs.jcim.8b00619.

[37] Tran TD, Ogbourne SM, Brooks PR, Sánchez-Cruz N, Medina-Franco JL, et al. Lessons from exploring chemical space and chemical diversity of propolis components. Int J Mol Sci 2020;21:4988. https://doi.org/10.3390/ijms21144988.

[38] Al Sharie AH, El-Elimat T, Al Zu'bi YO, Aleshawi AJ, Medina-Franco JL. Chemical space and diversity of seaweed metabolite database (SWMD): a cheminformatics study. J Mol Graph Model 2020;100. https://doi.org/10.1016/j.jmgm.2020.107702, 107702.

[39] Medina-Franco JL, Maggiora GM. Molecular similarity analysis. Chemoinformatics for drug discovery. John Wiley & Sons, Inc.; 2013. p. 343–99.

[40] Saldívar-González FI, Huerta-García CS, Medina-Franco JL. Chemoinformatics-based enumeration of chemical libraries: a tutorial. J Chem 2020;12:64. https://doi.org/10.1186/s13321-020-00466-z.

[41] Lovering F. Escape from flatland 2: complexity and promiscuity. MedChemComm 2013;4:515–9. https://doi.org/10.1039/C2MD20347B.

[42] Wei W, Cherukupalli S, Jing L, Liu X, Zhan P. Fsp3: a new parameter for drug-likeness. Drug Discov Today 2020;25:1839–45. https://doi.org/10.1016/j.drudis.2020.07.017.

[43] 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;55:460–73. https://doi.org/10.1021/ci500588j.

[44] López-López E, Naveja JJ, Medina-Franco JL. Datawarrior: an evaluation of the open-source drug discovery tool. Expert Opin Drug Discov 2019;14:335–41. https://doi.org/10.1080/17460441.2019.1581170.

[45] Ganesan A. Natural products as a hunting ground for combinatorial chemistry. Curr Opin Biotechnol 2004;15:584–90. https://doi.org/10.1016/j.copbio.2004.09.002.

[46] Mendez D, Gaulton A, Bento AP, Chambers J, De Veij M, et al. ChEMBL: towards direct deposition of bioassay data. Nucl Acids Res 2019;47:D930–40. https://doi.org/10.1093/nar/gky1075.

[47] Manallack DT, Prankerd RJ, Yuriev E, Oprea TI, Chalmers DK. The significance of acid/base properties in drug discovery. Chem Soc Rev 2013;42:485–96. https://doi.org/10.1039/c2cs35348b.

[48] Santibáñez-Morán MG, Medina-Franco JL. Analysis of the acid/base profile of natural products from different sources. Mol Inform 2020;39. https://doi.org/10.1002/minf.201900099, e1900099.

[49] Santibáñez-Morán MG, Rico-Hidalgo MP, Manallack DT, Medina-Franco JL. The acid/base profile of a large food chemical database. Mol Inform 2019;38. https://doi.org/10.1002/minf.201800171, e1800171.

[50] Jia CY, Li JY, Hao GF, Yang GF. A drug-likeness toolbox facilitates Admet study in drug discovery. Drug Discov Today 2020;25:248–58. https://doi.org/10.1016/j.drudis.2019.10.014.

[51] Schneckener S, Grimbs S, Hey J, Menz S, Osmers M, et al. Prediction of oral bioavailability in rats: transferring insights from in vitro correlations to (deep) machine learning models using in silico model outputs and chemical structure parameters. J Chem Inf Model 2019;59:4893–905. https://doi.org/10.1021/acs.jcim.9b00460.

[52] Vo AH, Van Vleet TR, Gupta RR, Liguori MJ, Rao MS. An overview of machine learning and big data for drug toxicity evaluation. Chem Res Toxicol 2020;33:20–37. https://doi.org/10.1021/acs.chemrestox.9b00227.

[53] Gonzalez-Medina M, Naveja JJ, Sanchez-Cruz N, Medina-Franco JL. Open chemoinformatic resources to explore the structure, properties and chemical space of molecules. RSC Adv 2017;7:54153–63. https://doi.org/10.1039/C7RA11831G.

[54] Ntie-Kang F. An in silico evaluation of the Admet profile of the Streptomedb database. SpringerPlus 2013;2:353. https://doi.org/10.1186/2193-1801-2-353.

[55] Fatima S, Gupta P, Sharma S, Sharma A, Agarwal SM. Admet profiling of geographically diverse phytochemical using chemoinformatic tools. Fut Med Chem 2020;12:69–87. https://doi.org/10.4155/fmc-2019-0206.

[56] Durán-Iturbide NA, Díaz-Eufracio BI, Medina-Franco JL. In Silico Adme/Tox profiling of natural products: a focus on Biofacquim. ACS Omega 2020;5:16076–84. https://doi.org/10.1021/acsomega.0c01581.

[57] Pires DEV, Blundell TL, Ascher DB. Pkcsm: predicting small-molecule pharmacokinetic and toxicity properties using graph-based signatures. J Med Chem 2015;58:4066–72. https://doi.org/10.1021/acs.jmedchem.5b00104.

[58] González-Medina M, Prieto-Martínez FD, Medina-Franco JL. Consensus diversity plots: a global diversity analysis of chemical libraries. J Chem 2016;8:63. https://doi.org/10.1186/s13321-016-0176-9.

[59] Olmedo DA, González-Medina M, Gupta MP, Medina-Franco JL. Cheminformatic characterization of natural products from Panama. Mol Divers 2017;21:779–89. https://doi.org/10.1007/s11030-017-9781-4.

[60] Sánchez-Cruz N, Pilón-Jiménez B, Medina-Franco J. Functional group and diversity analysis of Biofacquim: a Mexican natural product database [version 2; peer review: 3 approved]. F1000Research 2020;8:2071. https://doi.org/10.12688/f1000research.21540.2.

[61] Saldívar-González FI, Lenci E, Calugi L, Medina-Franco JL, Trabocchi A. Computational-aided design of a library of lactams through a diversity-oriented synthesis strategy. Bioorg Med Chem 2020;28. https://doi.org/10.1016/j.bmc.2020.115539, 115539.

[62] Virshup AM, Contreras-García J, Wipf P, Yang W, Beratan DN. Stochastic voyages into uncharted chemical space produce a representative library of all possible drug-like compounds. J Am Chem Soc 2013;135:7296–303. https://doi.org/10.1021/ja401184g.

[63] Naveja J, Medina-Franco J. Chemmaps: towards an approach for visualizing the chemical space based on adaptive satellite compounds [version 2; peer review: 3 approved with reservations]. F1000Research 2017;6. https://doi.org/10.12688/f1000research.12095.2.

[64] Medina-Franco JL, Martínez-Mayorga K, Giulianotti MA, Houghten RA, Pinilla C. Visualization of the chemical space in drug discovery. Curr Comput Aided Drug Des 2008;4:322–33. https://doi.org/10.2174/157340908786786010.

[65] Osolodkin DI, Radchenko EV, Orlov AA, Voronkov AE, Palyulin VA, et al. Progress in visual representations of chemical space. Exp

Opin Drug Discov 2015;10:959–73. https://doi.org/10.1517/17460441.2015.1060216.

[66] Olmedo DA, Medina-Franco JL. Chemoinformatic approach: The case of natural products of Panama. In: Stefaniu A, Rasul A, Hussain G, editors. Cheminformatics and its applications. IntechOpen; 2019. https://doi.org/10.5772/intechopen.87779. Available from: https://www.intechopen.com/online-first/chemoinformatic-approach-the-case-of-natural-products-of-panama.

[67] Probst D, Reymond J-L. Visualization of very large high-dimensional data sets as minimum spanning trees. J Chem 2020;12:12. https://doi.org/10.1186/s13321-020-0416-x.

[68] Saldívar-González FI, Naveja JJ, Palomino-Hernández O, Medina-Franco JL. Getting smart in drug discovery: chemoinformatics approaches for mining structure–multiple activity relationships. RSC Adv 2017;7:632–41. https://doi.org/10.1039/C6RA26230A.

[69] Medina-Franco JL, Naveja JJ, López-López E. Reaching for the bright stars in chemical space. Drug Discov Today 2019;24:2162–9. https://doi.org/10.1016/j.drudis.2019.09.013.

[70] Chen CY-C. Tcm database@Taiwan: the world's largest traditional Chinese medicine database for drug screening in silico. PLoS One 2011;6. https://doi.org/10.1371/journal.pone.0015939, e15939.

[71] Pilon AC, Valli M, Dametto AC, Pinto MEF, Freire RT, et al. Nubbedb: an updated database to uncover chemical and biological information from Brazilian biodiversity. Sci Rep 2017;7:7215. https://doi.org/10.1038/s41598-017-07451-x.

[72] Molvs M, https://molvs.readthedocs.io/en/latest/.

[73] Lewell XQ, Judd DB, Watson SP, Hann MM. Recap retrosynthetic combinatorial analysis procedure: a powerful new technique for identifying privileged molecular fragments with useful applications in combinatorial chemistry. J Chem Inf Comput Sci 1998;38:511–22. https://doi.org/10.1021/ci970429i.

[74] Rogers D, Hahn M. Extended-connectivity fingerprints. J Chem Inf Model 2010;50:742–54. https://doi.org/10.1021/ci100050t.

[75] Saldívar-González FI, Pilón-Jiménez BA, Medina-Franco JL. Chemical space of naturally occurring compounds. Phys Sci Rev 2018;4:20180103. https://doi.org/10.1515/psr-2018-0103.

[76] Do QT, Medina-Franco JL, Scior T, Bernard P. How to valorize biodiversity? Let's go hashing, extracting, filtering, mining, fishing. Planta Med 2015;81:436–49. https://doi.org/10.1055/s-0034-1396314.

[77] de Sousa Luis JA, Barros RPC, de Sousab NF, Muratov E, Scotti L, et al. Virtual screening of natural products database. Mini-Rev Med Chem 2021;21:2657–730. https://doi.org/10.2174/1389557520666200730161549.

[78] Singh N, Chaput L, Villoutreix BO. Virtual screening web servers: designing chemical probes and drug candidates in the cyberspace. Brief Bioinform 2020;22:1790–818. https://doi.org/10.1093/bib/bbaa034.

[79] Peón A, Li H, Ghislat G, Leung KS, Wong MH, et al. Moltarpred: a web tool for comprehensive target prediction with reliability estimation. Chem Biol Drug Des 2019;94:1390–401. https://doi.org/10.1111/cbdd.13516.

[80] Hamad S, Adornetto G, Naveja JJ, Chavan Ravindranath A, Raffler J, et al. Hitpickv2: a web server to predict targets of chemical compounds. Bioinformatics 2019;35:1239–40. https://doi.org/10.1093/bioinformatics/bty759.

[81] Daina A, Michielin O, Zoete V. SwissTargetPrediction: updated data and new features for efficient prediction of protein targets of small molecules. Nucleic Acids Res 2019;47:W357–64. https://doi.org/10.1093/nar/gkz382.

[82] Awale M, Reymond JL. Polypharmacology browser Ppb2: target prediction combining nearest neighbors with machine learning. J Chem Inf Model 2019;59:10–7. https://doi.org/10.1021/acs.jcim.8b00524.

[83] Sánchez-Cruz N, Medina-Franco JL. Epigenetic target profiler: a web server to predict epigenetic targets of small molecules. J Chem Inf Model 2021;61:1550–4. https://doi.org/10.1021/acs.jcim.1c00045.

[84] Naveja JJ, Oviedo-Osornio CI, Trujillo-Minero NN, Medina-Franco JL. Chemoinformatics: a perspective from an academic setting in Latin America. Mol Divers 2018;22:247–58. https://doi.org/10.1007/s11030-017-9802-3.

[85] Mousavizadeh L, Ghasemi S. Genotype and phenotype of COVID-19: their roles in pathogenesis. J Microbiol Immunol Infect 2021;54:159–63. https://doi.org/10.1016/j.jmii.2020.03.022.

[86] Masters PS. Coronavirus genomic RNA packaging. Virology 2019;537:198–207. https://doi.org/10.1016/j.virol.2019.08.031.

[87] Mann DL. SARS-CoV-2 and bats: from flight to fighting COVID-19. JACC Basic Transl Sci 2020;5:545–6. https://doi.org/10.1016/j.jacbts.2020.04.012.

[88] Chen Y, Guo D. Molecular mechanisms of coronavirus RNA capping and methylation. Virol Sin 2016;31:3–11. https://doi.org/10.1007/s12250-016-3726-4.

[89] Lupia T, Scabini S, Mornese Pinna S, Di Perri G, De Rosa FG, et al. 2019 novel coronavirus (2019-Ncov) outbreak: a new challenge. J Glob Antimicrob Resist 2020;21:22–7. https://doi.org/10.1016/j.jgar.2020.02.021.

[90] Gorbalenya AE, Baker SC, Baric RS, de Groot RJ, Drosten C, et al. The species severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2. Nat Microbiol 2020;5:536–44. https://doi.org/10.1038/s41564-020-0695-z.

[91] Huang C, Wang Y, Li X, Ren L, Zhao J, et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. Lancet 2020;395:497–506. https://doi.org/10.1016/s0140-6736(20)30183-5.

[92] Hui DS, Azhar EI, Madani TA, Ntoumi F, Kock R, et al. The continuing 2019-Ncov epidemic threat of novel coronaviruses to global health—the latest 2019 novel coronavirus outbreak in Wuhan, China. Int J Infect Dis 2020;91:264–6. https://doi.org/10.1016/j.ijid.2020.01.009.

[93] Cucinotta D, Vanelli M. Who declares COVID-19 a pandemic. Acta Biomed 2020;91:157–60. https://doi.org/10.23750/abm.v91i1.9397.

[94] Hemmati F, Saedi S, Hemmati-Dinarvand M, Zarei M, Seghatoleslam A. Mysterious virus: a review on behavior and treatment approaches of the novel coronavirus, 2019-Ncov. Arch Med Res 2020;51:375–83. https://doi.org/10.1016/j.arcmed.2020.04.022.

[95] Li G, De Clercq E. Therapeutic options for the 2019 novel coronavirus (2019-Ncov). Nat Rev Drug Discov 2020;19:149–50. https://doi.org/10.1038/d41573-020-00016-0.

[96] Morse JS, Lalonde T, Xu S, Liu WR. Learning from the past: possible urgent prevention and treatment options for severe acute respiratory infections caused by 2019-Ncov. ChemBioChem 2020;21:730–8. https://doi.org/10.1002/cbic.202000047.

[97] Desforges M, Le Coupanec A, Stodola JK, Meessen-Pinard M, Talbot PJ. Human coronaviruses: viral and cellular factors involved in neuroinvasiveness and neuropathogenesis. Virus Res 2014;194:145–58. https://doi.org/10.1016/j.virusres.2014.09.011.

[98] Neuman BW, Buchmeier MJ. Supramolecular architecture of the coronavirus particle. In: Ziebuhr J, editor. Advances in virus research. Academic Press; 2016. p. 1–27 [chapter 1].

[99] Chang C-k, Sue S-C, Yu T-h, Hsieh C-M, Tsai C-K, et al. Modular organization of Sars coronavirus nucleocapsid protein. J Biomed Sci 2006;13:59–72. https://doi.org/10.1007/s11373-005-9035-9.

[100] Liu C, Zhou Q, Li Y, Garner LV, Watkins SP, et al. Research and development on therapeutic agents and vaccines for COVID-19 and related human coronavirus diseases. ACS Cent Sci 2020;6:315–31. https://doi.org/10.1021/acscentsci.0c00272.

[101] Bocci G, Bradfute SB, Ye C, Garcia MJ, Parvathareddy J, et al. Virtual and in vitro antiviral screening revive therapeutic drugs for COVID-19. ACS Pharmacol Transl Sci 2020. https://doi.org/10.1021/acsptsci.0c00131.

[102] Muratov EN, Bajorath J, Sheridan RP, Tetko IV, Filimonov D, et al. QSAR without borders. Chem Soc Rev 2020;49:3525–64. https://doi.org/10.1039/D0CS00098A.

[103] Sterling T, Irwin JJ. Zinc 15 – ligand discovery for everyone. J Chem Inf Model 2015;55:2324–37. https://doi.org/10.1021/acs.jcim.5b00559.

[104] Santibáñez-Morán MG, López-López E, Prieto-Martínez FD, Sánchez-Cruz N, Medina-Franco JL. Consensus virtual screening of dark chemical matter and food chemicals uncover potential inhibitors of SARS-CoV-2 main protease. RSC Adv 2020;10:25089–99. https://doi.org/10.1039/D0RA04922K.

[105] Walters WP. Virtual chemical libraries. J Med Chem 2019;62:1116–24. https://doi.org/10.1021/acs.jmedchem.8b01048.

[106] Vougogiannopoulou K, Corona A, Tramontano E, Alexis MN, Skaltsounis AL. Natural and nature-derived products targeting human coronaviruses. Molecules 2021;26:448. https://doi.org/10.3390/molecules26020448.

[107] Sepay N, Sekar A, Halder UC, Alarifi A, Afzal M. Anti-COVID-19 terpenoid from marine sources: a docking, Admet and Molecular Dynamics Study. J Mol Struct 2020;129433. https://doi.org/10.1016/j.molstruc.2020.129433.

[108] Ghosh K, Amin SA, Gayen S, Jha T. Chemical-informatics approach to COVID-19 drug discovery: exploration of important fragments and data mining based prediction of some hits from natural origins as main protease (Mpro) inhibitors. J Mol Struct 2021;1224. https://doi.org/10.1016/j.molstruc.2020.129026, 129026.

[109] Mu C, Sheng Y, Wang Q, Amin A, Li X, et al. Potential compound from herbal food of Rhizoma Polygonati for treatment of COVID-19 analyzed by network pharmacology and molecular docking technology. J Funct Foods 2020;104149. https://doi.org/10.1016/j.jff.2020.104149.

[110] Dubey K, Dubey R. Computation screening of narcissoside a glycosyloxyflavone for potential novel coronavirus 2019 (COVID-19) inhibitor. Biomed J 2020;43:363–7. https://doi.org/10.1016/j.bj.2020.05.002.

[111] Naik B, Gupta N, Ojha R, Singh S, Prajapati VK, et al. High throughput virtual screening reveals SARS-CoV-2 multi-target binding natural compounds to lead instant therapy for COVID-19 treatment. Int J Biol Macromol 2020;160:1–17. https://doi.org/10.1016/j.ijbiomac.2020.05.184.

[112] Maurya AK, Mishra N. In silico validation of coumarin derivatives as potential inhibitors against main protease, Nsp10/Nsp16-methyltransferase, phosphatase and endoribonuclease of Sars Cov-2. J Biomol Struct Dyn 2020;1–16. https://doi.org/10.1080/07391102.2020.1808075.

[113] Yang J-S, Chiang J-H, Tsai SC, Hsu Y-M, Bau D-T, et al. In silico de novo curcuminoid derivatives from the compound library of natural products research laboratories inhibit COVID-19 3CLpro activity. Nat Prod Commun 2020;15. https://doi.org/10.1177/1934578X20953262. 1934578X20953262.

[114] Gentile D, Patamia V, Scala A, Sciortino MT, Piperno A, et al. Putative inhibitors of SARS-CoV-2 main protease from a library of marine natural products: a virtual screening and molecular modeling study. Mar Drugs 2020;18:225. https://doi.org/10.3390/md18040225.

[115] Rahman N, Basharat Z, Yousuf M, Castaldo G, Rastrelli L, et al. Virtual screening of natural products against type II transmembrane serine protease (TMPRSS2), the priming agent of coronavirus 2 (SARS-CoV-2). Molecules 2020;25:2271. https://doi.org/10.3390/molecules25102271.

[116] Mazzini S, Musso L, Dallavalle S, Artali R. Putative SARS-CoV-2 M(Pro) inhibitors from an in-house library of natural and nature-inspired products: a virtual screening and molecular docking study. Molecules 2020;25. https://doi.org/10.3390/molecules25163745.

[117] Kumar BK, Faheem SK, Ojha R, Prajapati VK, et al. Pharmacophore based virtual screening, molecular docking, molecular dynamics and MM-GBSA approach for identification of prospective SARS-CoV-2 inhibitor from natural product databases. J Biomol Struct Dyn 2020;1-24. https://doi.org/10.1080/07391102.2020.1824814.

[118] Joshi T, Joshi T, Pundir H, Sharma P, Mathpal S, et al. Predictive modeling by deep learning, virtual screening and molecular dynamics study of natural compounds against SARS-CoV-2 main protease. J Biomol Struct Dyn 2020;1–19. https://doi.org/10.1080/07391102.2020.1802341.

[119] Majumder R, Mandal M. Screening of plant-based natural compounds as a potential COVID-19 main protease inhibitor: an in silico docking and molecular dynamics simulation approach. J Biomol Struct Dyn 2020;1–16. https://doi.org/10.1080/07391102.2020.1817787.

[120] Sharma A, Tiwari V, Sowdhamini R. Computational search for potential COVID-19 drugs from FDA approved drugs and small molecules of natural origin identifies several anti-virals and plant products. J Biosci 2020;45. https://doi.org/10.1007/s12038-020-00069-8.

[121] Azim KF, Ahmed SR, Banik A, Khan MMR, Deb A, et al. Screening and druggability analysis of some plant metabolites against SARS-CoV-2: an integrative computational approach. Inform Med Unlocked 2020;20. https://doi.org/10.1016/j.imu.2020.100367, 100367.

[122] Mitra K, Ghanta P, Acharya S, Chakrapani G, Ramaiah B, et al. Dual inhibitors of SARS-CoV-2 proteases: pharmacophore and molecular dynamics based drug repositioning and phytochemical leads. J Biomol Struct Dyn 2020;1-14. https://doi.org/10.1080/07391102.2020.1796802.

[123] Park JY, Kim JH, Kwon JM, Kwon HJ, Jeong HJ, et al. Dieckol, a SARS-CoV 3cl(Pro) inhibitor, isolated from the edible brown algae *Ecklonia cava*. Bioorg Med Chem 2013;21:3730–7. https://doi.org/10.1016/j.bmc.2013.04.026.

[124] Yoo J, Medina-Franco JL. Trimethylaurintricarboxylic acid inhibits human DNA methyltransferase 1: insights from enzymatic and molecular modeling studies. J Mol Model 2012;18:1583–9. https://doi.org/10.1007/s00894-011-1191-4.

[125] Juárez-Mercado KE, Prieto-Martínez FD, Sánchez-Cruz N, Peña-Castillo A, Prada-Gracia D, et al. Expanding the structural diversity of DNA methyltransferase inhibitors. Pharmaceuticals 2020;14. https://doi.org/10.3390/ph14010017.

[126] Waddington CH. The epigenotype, Endeavor, 1942, Vol. 1 (Pg. 18–20). Reprinted in Int J Epidemiol 2012;41:10–3. https://doi.org/10.1093/ije/dyr184.

[127] Wu C, Morris JR. Genes, genetics, and epigenetics: a correspondence. Science 2001;293:1103–5. https://doi.org/10.1126/science.293.5532.1103.

[128] Cavalli G, Heard E. Advances in epigenetics link genetics to the environment and disease. Nature 2019;571:489–99. https://doi.org/10.1038/s41586-019-1411-0.

[129] Ganesan A, Arimondo PB, Rots MG, Jeronimo C, Berdasco M. The timeline of epigenetic drug discovery: from reality to dreams. Clin Epigenetics 2019;11:174. https://doi.org/10.1186/s13148-019-0776-0.

[130] Lyko F. The DNA methyltransferase family: a versatile toolkit for epigenetic regulation. Nat Rev Genet 2017;19:81. https://doi.org/10.1038/nrg.2017.80.

[131] Zhang J, Yang C, Wu C, Cui W, Wang L. DNA methyltransferases in cancer: biology, paradox, aberrations, and targeted therapy. Cancers 2020;12:2123.

[132] Kuck D, Singh N, Lyko F, Medina-Franco JL. Novel and selective DNA methyltransferase inhibitors: docking-based virtual screening and experimental evaluation. Bioorg Med Chem 2010;18:822–9. https://doi.org/10.1016/j.bmc.2009.11.050.

[133] Kuck D, Caulfield T, Lyko F, Medina-Franco JL. Nanaomycin a selectively inhibits DNMT3B and reactivates silenced tumor suppressor genes in human cancer cells. Mol Cancer Ther 2010;9:3015–23. https://doi.org/10.1158/1535-7163.MCT-10-0609.

[134] Yoo J, Kim JH, Robertson KD, Medina-Franco JL. Molecular modeling of inhibitors of human DNA methyltransferase with a crystal structure: discovery of a novel DNMT1 inhibitor. Adv Protein Chem Struct Biol 2012;87:219–47. https://doi.org/10.1016/B978-0-12-398312-1.00008-1.

[135] Méndez-Lucio O, Tran J, Medina-Franco JL, Meurice N, Muller M. Towards drug repurposing in epigenetics: olsalazine as a novel hypomethylating compound active in a cellular context. ChemMedChem 2014;9:560–5. https://doi.org/10.1002/cmdc.201300555.

[136] Aldawsari FS, Aguayo-Ortiz R, Kapilashrami K, Yoo J, Luo M, et al. Resveratrol-salicylate derivatives as selective DNMT3 inhibitors and anticancer agents. J Enzyme Inhib Med Chem 2016;31:695–703. https://doi.org/10.3109/14756366.2015.1058256.

[137] Davide G, Sandra A, Emily B, Mattia C, Marta G, et al. Design and synthesis of N-benzoyl amino acid derivatives as DNA methylation inhibitors. Chem Biol Drug Des 2016;88:664–76. https://doi.org/10.1111/cbdd.12794.

[138] Palomino-Hernandez O, Jardinez-Vera A, Medina-Franco J. Progress on the computational development of epigenetic modulators of DNA methyltransferases 3a and 3b. J Mex Chem Soc 2017;61:266–72.

[139] Pechalrieu D, Dauzonne D, Arimondo PB, Lopez M. Synthesis of novel 3-halo-3-nitroflavanones and their activities as DNA methyltransferase inhibitors in cancer cells. Eur J Med Chem 2020;186. https://doi.org/10.1016/j.ejmech.2019.111829, 111829.

[140] Shao Z, Xu P, Xu W, Li L, Liu S, et al. Discovery of novel DNA methyltransferase 3a inhibitors via structure-based virtual screening and biological assays. Bioorg Med Chem Lett 2017;27:342–6. https://doi.org/10.1016/j.bmcl.2016.11.023.

[141] Krishna S, Shukla S, Lakra AD, Meeran SM, Siddiqi MI. Identification of potent inhibitors of DNA methyltransferase 1 (DNMT1) through a pharmacophore-based virtual screening approach. J Mol Graph Model 2017;75:174–88. https://doi.org/10.1016/j.jmgm.2017.05.014.

[142] Erdmann A, Arimondo PB, Guianvarc'h D. Structure-guided optimization of DNA methyltransferase inhibitors. In: Medina-Franco JL, editor. Epi-informatics. London, UK: Academic Press; 2016. p. 53–74.

[143] Joshi M, Rajpathak SN, Narwade SC, Deobagkar D. Ensemble-based virtual screening and experimental validation of inhibitors targeting a novel site of human DNMT1. Chem Biol Drug Des 2016;88:5–16. https://doi.org/10.1111/cbdd.12741.

[144] Kabro A, Lachance H, Marcoux-Archambault I, Perrier V, Dore V, et al. Preparation of phenylethylbenzamide derivatives as modulators of DNMT3 activity. MedChemComm 2013;4:1562–70. https://doi.org/10.1039/c3md00214d.

[145] Castellano S, Kuck D, Viviano M, Yoo J, López-Vallejo F, et al. Synthesis and biochemical evaluation of Δ2-Isoxazoline derivatives as DNA methyltransferase 1 inhibitors. J Med Chem 2011;54:7663–77. https://doi.org/10.1021/jm2010404.

[146] Newton AS, Faver JC, Micevic G, Muthusamy V, Kudalkar SN, et al. Structure-guided identification of DNMT3B inhibitors. ACS Med Chem Lett 2020;11:971–6. https://doi.org/10.1021/acsmedchemlett.0c00011.

[147] Medina-Franco JL, López-Vallejo F, Kuck D, Lyko F. Natural products as DNA methyltransferase inhibitors: a computer-aided discovery approach. Mol Divers 2011;15:293–304. https://doi.org/10.1007/s11030-010-9262-5.

[148] Akone SH, Ntie-Kang F, Stuhldreier F, Ewonkem MB, Noah AM, et al. Natural products impacting DNA methyltransferases and histone deacetylases. Front Pharmacol 2020;11:992. https://doi.org/10.3389/fphar.2020.00992.

[149] Lee WJ, Shim JY, Zhu BT. Mechanisms for the inhibition of DNA methyltransferases by tea catechins and bioflavonoids. Mol Pharmacol 2005;68:1018–30. https://doi.org/10.1124/mol.104.008367.

[150] Lee WJ, Zhu BT. Inhibition of DNA methylation by caffeic acid and chlorogenic acid, two common catechol-containing coffee polyphenols. Carcinogenesis 2006;27:269–77. https://doi.org/10.1093/carcin/bgi206.

[151] Martinez-Mayorga K, Montes CP. The role of nutrition in epigenetics and recent advances of in silico studies. In: Medina-Franco JL, editor. Epi-informatics. London, UK: Academic Press; 2016. p. 385–98.

[152] Prieto-Martínez F, Peña-Castillo A, Méndez-Lucio O, Fernández-de Gortari E, Medina-Franco JL. Molecular modeling and chemoinformatics to advance the development of modulators of epigenetic targets: a focus on DNA methyltransferases. Adv Protein Chem Struct Biol 2016;105:1–26. https://doi.org/10.1016/bs.apcsb.2016.05.001.

[153] Rajavelu A, Tulyasheva Z, Jaiswal R, Jeltsch A, Kuhnert N. The inhibition of the mammalian DNA methyltransferase 3a (DNMT3a) by dietary black tea and coffee polyphenols. BMC Biochem 2011;12. https://doi.org/10.1186/1471-2091-12-16.

[154] Juarez-Mercado KE, Prieto-Martinez FD, Sanchez-Cruz N, Pena-Castillo A, Prada-Gracia D, et al. DNA methyltransferase inhibitors with novel chemical scaffolds. bioRxiv 2020. https://doi.org/10.1101/2020.10.13.337709. 2020.2010.2013.337709.

[155] Reaction Biology Corporation. http://www.reactionbiology.com [Accessed May 2021].

[156] Castillo-Aguilera O, Depreux P, Halby L, Arimondo P, Goossens L. DNA methylation targeting: the DNMT/HMT crosstalk challenge. Biomolecules 2017;7:3. https://doi.org/10.3390/biom7010003.

[157] Anon. Molecular Operating Environment (MOE), version 2019.2008. Montreal, QC, Canada: Chemical Computing Group Inc; 2019. Available at: http://www.chemcomp.com. [Accessed January 2022].

[158] Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, et al. The Protein Data Bank. Nucl Acids Res 2000;28:235–42. https://doi.org/10.1093/nar/28.1.235.

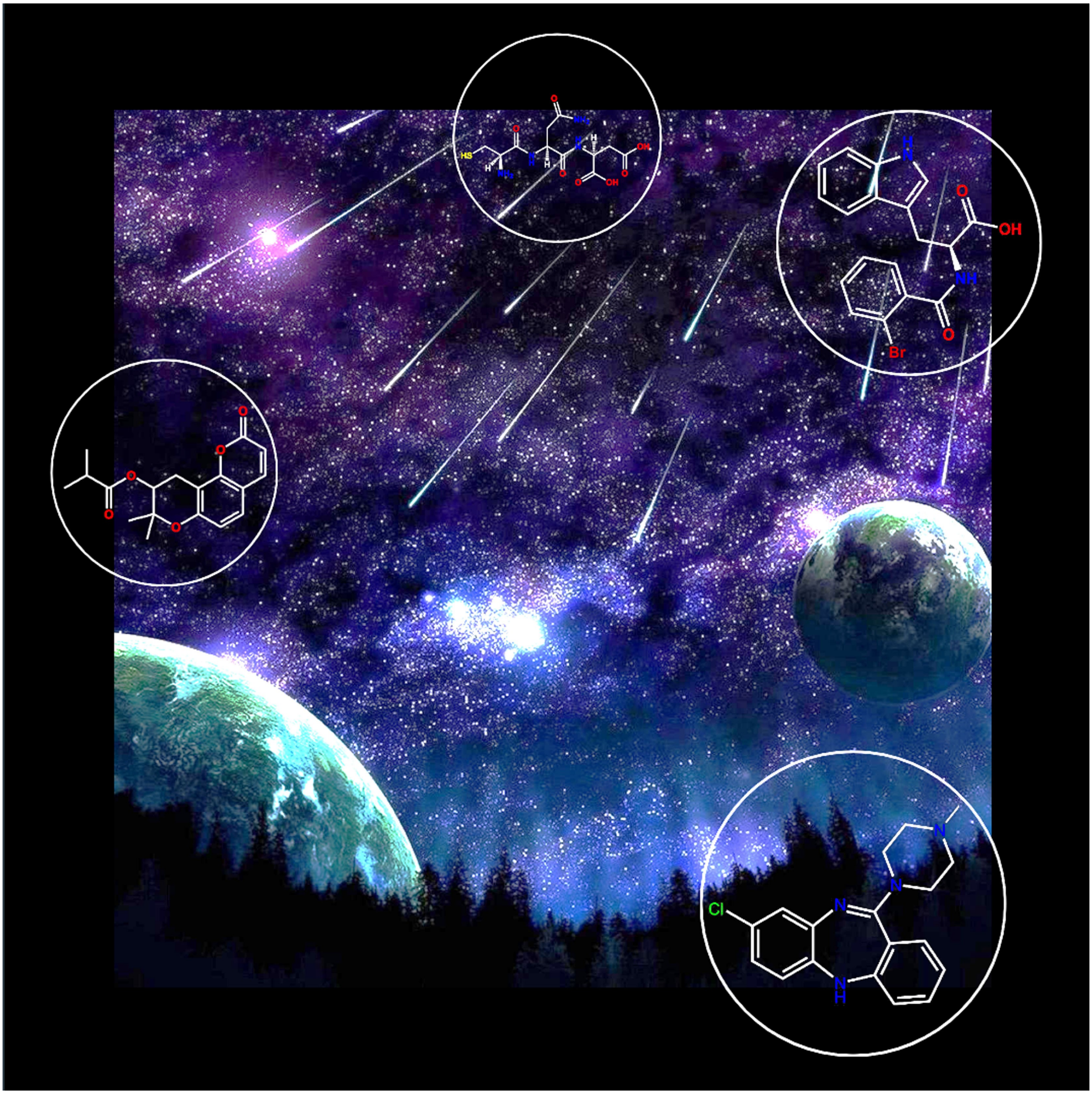
[159] Zhang Z-M, Liu S, Lin K, Luo Y, Perry JJ, et al. Crystal structure of human DNA methyltransferase 1. J Mol Biol 2015;427:2520–31. https://doi.org/10.1016/j.jmb.2015.06.001.

Artículos de difusión y divulgación

doi.org/10.1002/minf.202200116

Chemical Multiverse: An Expanded View of Chemical Space

José L. Medina-Franco,*[a] Ana L. Chávez-Hernández,[a] Edgar López-López,[b] and Fernanda I. Saldívar-González[a]



Abstract: Technological advances and practical applications of the chemical space concept in drug discovery, natural product research, and other research areas have attracted the scientific community's attention. The large- and ultra-large chemical spaces are associated with the significant increase in the number of compounds that can potentially be made and exist and the increasing number of experimental and calculated descriptors, that are emerging that encode the molecular structure and/or property aspects of the molecules. Due to the importance and continued evolution of compound libraries, herein, we discuss definitions proposed in the literature for chemical space and emphasize the convenience, discussed in the literature to use complementary descriptors to obtain a comprehensive view of the chemical space of compound data sets. In this regard, we introduce the term *chemical multiverse* to refer to the comprehensive analysis of compound data sets through several chemical spaces, each defined by a different set of chemical representations. The chemical multiverse is contrasted with a related idea: consensus chemical space.

Keywords: chemical multiverse · chemical space · drug discovery · machine learning · molecular representation · structure-property relationships · ultra-large chemical library · visualization

1 Introduction

The concept of "chemical universe," "chemical space," or "chemical compound space" is associated with a set of all possible molecules described by a multi-dimensional space that represents their functional and structural properties and the relationship of the molecules to each other.[1,2] Although the type of molecules could be any, the chemical space has been typically studied quantitatively and qualitatively with small organic compounds. Practical applications in drug discovery approaches include the study of epigenetic-focused compounds,[3,4] covalent protein kinase inhibitors,[5] human immunodeficiency virus (HIV) protease inhibitors,[6] natural products (NPs),[7,8] and novel compounds from combining fragments or scaffolds, e.g., pseudo-NPs.[9] Also, there are reported applications of the chemical space concept to food and flavor chemicals,[10] peptides,[11] and metal-containing molecules,[12] and virtual and on-demand libraries.[13]

In theory, the multi-dimensional space can be formed by two (or even one) dimensions, e.g., a single or two descriptors that encode a specific set of structural or functional properties. However, depending on the project's goals, the dimensions are typically more than three. Eventually, it could contain hundreds or a few thousands of descriptors, for instance, when using structural fingerprints. Many descriptors demand the implementation of dimensionality-reduction techniques to generate two (2D) – or three-dimensional (3D) visual representations of the multi-dimensional descriptor space. Reduction methods such as t-distributed stochastic neighbor embedding (t-SNE),[14] principal component analysis (PCA),[15] self-organized maps (SOMs),[16] generative topographic mapping (GTM),[17–19] and chemical space networks[20,21] currently are the most frequently used, but there are others. Visual representation of the chemical space has been the focus of several research projects, as further discussed below.

For several decades, the chemical space concept has been of interest in several areas of chemistry, emphasizing drug discovery. With the rapid advances in machine learning and *de novo* design and the number of chemical compounds that exist or could be made, the community has a significant interest in enumerating large- and ultra-large chemical libraries containing billions of chemical structures.[22] Hence, there is an increased interest in studying the huge chemical libraries under the chemical space concept, e.g., systematic and consistent description with novel and existing chemical descriptors, visual representation of the chemical space of libraries with millions of compounds, and several other analyses that can be done based on the multi-dimensional chemical space (examples of the latter are diversity analysis, similarity-based virtual screening, property, and biological activity prediction). For instance, Schmidt et al. and Bellmann et al. recently explored the chemical space of make-on-demand libraries.[13,23] Zabolotna et al. reported the implementation of GTM to efficiently navigate the chemical space of the entire ZINC library of purchasable compounds, relative to the biologically relevant ChEMBL compounds.[19]

There are several reviews focused on chemical space. These reviews, summarized in Table 1, cover applications to drug discovery and NPs research, progress on visualization methods, and approaches to study the chemical space of small organic compounds, emphasizing the application of

[a] J. L. Medina-Franco, A. L. Chávez-Hernández, F. I. Saldívar-González
DIFACQUIM research group, Department of Pharmacy, School of Chemistry, National Autonomous University of Mexico, Mexico City 04510, Mexico
E-mail: jose.medina.franco@gmail.com
medinajl@unam.mx

[b] E. López-López
Department of Pharmacology, Center for Research and Advanced Studies of the National Polytechnic Institute (CINVESTAV), Mexico City 07360, Mexico

public resources. However, the review or research papers using the concept of chemical space often have a narrow or very focused view of the space. For instance, they usually refer to organic (typically small) molecules.[24]

This manuscript's goal is to review definitions proposed in the literature for chemical space, emphasizing the approaches to generate a consensus view of the chemical space. Building upon the developments of others and our group on chemical space, we also propose the term *chemical multiverse* highlighting the convenience of employing multiple descriptors for a comprehensive assessment of the chemical space. After this brief introduction, we discuss the current definitions of chemical space. Section 3 introduces the term of chemical multiverse and compares it with the notion of consensus chemical space. In Section 4 we review the applications discussed in the literature of multiple descriptors to analyze the chemical space of compound data sets. In other words, we present case studies discussed in the literature showing the general applicability of chemical multiverses in diversity analysis, virtual screening, and structure-activity relationships, among other applications.

2 Current Views of Chemical Space

There are several definitions of chemical space proposed in the literature. Table 2 summarizes examples.

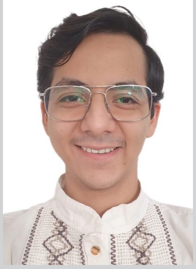
As reviewed in Table 2, several definitions conceptualize the chemical space as a chemical descriptor vector space set by the numerical vector D encoding property or molecular structure aspects as elements of the descriptor vector D.[35] Based on this notion, Figure 1 shows the chemical space concept like "M-multidimensional cartesian space," aka a "chemical space table." Rows represent the number of (n) molecules, and the columns (M) are the number of descriptors or features that encode each molecule. The length of descriptor sets corresponds to the number of dimensions defining the chemical space itself.

3 Chemical Multiverse: General Concept

The chemical space concept implies that a given set of *n* molecules represented with different descriptors would lead to distinct chemical universes e.g., each one is "descriptor universe". Varnek and Baskin have pointed out that "unlike real physical space, a chemical space is not unique: each ensemble of graphs and descriptors defines its own chemical space."[39] It also follows that molecules with very different chemical nature, for instance, organic small-molecules, e.g., lead- and drug-like; peptides; metal-containing compounds, macromolecules, biologics, etc., yield divergent chemical spaces, this be their own nature of the descriptors required to represent the compounds. In addition, the number and type of chemical and biological-related descriptors available from experimental data or computational calculations, e.g., quantum mechanics,[32] are also increasing, yielding the chance to augment the number of possible valid chemical spaces.

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 the Torrey Pines Institute for Molecular Studies in Florida in 2007. In 2013, he moved to the Mayo Clinic and later joined UNAM as 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 an emphasis on drug discovery.

Ana Luisa Chávez-Hernández received her BSc degree in Food Engineering (2016) from the Autonomy Metropolitan University (UAM) and her Master's degree in Chemical Science (2019) from the National Autonomous University of Mexico (UNAM). Currently, she is a Ph.D. student in Chemistry Science under the supervision of Professor José Luis Medina-Franco; her research focuses on the development of libraries from natural products for de novo drug design.



Edgar López-López received is B.S. degree in Clinical Chemistry from the University of Veracruz, Mexico, in 2019, and an M.Sc. degree in pharmacology from the Center for Research and Advanced Studies of the National Polytechnic Institute (CINVESTAV), Mexico, in 2021. He is currently a Ph.D. student, and his research interest includes the design, synthesis, and biological evaluation of anticancer and antiparasitic drugs.

Fernanda I. Saldivar-Gonzalez received her BSc degree in Chemistry Pharmacy and Biology (2017) from the National Autonomous University of Mexico (UNAM). She received the Master's degree in Chemical Sciences in 2019 under the supervision of Professor José Luis Medina-Franco, after spending a research period in the group of Prof. Andrea Trabocchi at the University of Florence, Italy. She is currently a Ph.D. student in Chemistry in the area of pharmacy where she develops her project focused on the design of virtual chemical libraries of antidiabetic compounds.

Table 1. Selected review papers focused on the chemical space of different types of compounds.[a]

Review title	Focus of the review	Ref.
Visualization of the chemical space in drug discovery.	Summarized chemical space visualization, several programs to visualize chemical space for chemical compounds from natural products, drug molecules, and combinatorial libraries, and their application in drug discovery.	[25]
Exploring chemical space for drug discovery using the chemical universe database.	Progress on the authors in searching for bioactive ligands by enumeration and virtual screening of the unknown chemical space of small molecules. The review covers the application of these libraries to the search for NMDA-receptor analogous.	[26]
The chemical space project.	Development of chemical universe databases (GDB), a project to enumerate all possible molecules to generate an unbiased insight into the entire chemical space, taking only simple chemical stability and synthetic feasibility criteria.	[27]
Progress in visual representation of chemical space.	Chemical space concepts. Advances in methods for visualization of chemical space, and older but overlooked methods.	[28]
Reaching for the bright StARs in chemical space.	Visualization methods to explore the chemical space aiming at reaching insightful structure-activity relationships.	[29]
Chemoinformatics in natural product-based drug discovery.	General review of chemoinformatics applied to natural products. It includes analysis, visualization, and navigation of their chemical space.	[30]
Defining and exploring chemical spaces.	Overview of algorithmic approaches to defining and exploring chemical spaces that have the potential to operationalize the process of molecular discovery.	[31]
Ab initio machine learning in chemical compound space.	Machine learning studies aimed at sampling exhaustively the chemical compound space. The review covers novel molecules in general (non-only small organic molecules) and materials.	[32]
Progress on open chemoinformatic tools for expanding and exploring the chemical space.	Recent progress on chemoinformatic tools to expand and characterize the chemical space of compound data sets employing various types of molecular representations, generate visual representations of the chemical space and analyze the SAR of data sets.	[33]
Using deep neural networks to explore chemical space.	Common deep learning methods to explore the chemical space. The review discusses the selection of molecular representation, training for focused chemical space exploration, and considerations for assessing and validating the chemical space coverage.	[34]
Approaches for enhancing the analysis of chemical space for drug discovery.	The current state of chemical space in drug design and discovery. Topics discussed: advances for efficient navigation in chemical space, the use of this concept in assessing the diversity of different data sets, exploring SAR for one or multiple endpoints, and compound library design.	[35]

[a] In chronological order of publication.

Table 2. Examples of definitions of chemical space concepts, as proposed in the literature.[a]

Author(s)	Chemical space definitions	Ref.
Dobson	"All possible small organic molecules, including those present in biological systems".	[36]
Lipinski and Hopkins	"Chemical space can be viewed as being analogous to cosmological universe in its vastness, with chemical compounds populating space instead of stars".	[37]
Reymond, et al.	"Ensemble of all known and possible molecules described by their chemical properties".	[38]
Varnek and Baskin	"The ensemble of graphs or descriptor vectors forms a chemical space in which some relations between the objects must be defined".	[39]
von Lilienfeld, et al.	"The combinatorial set of all compounds that can be isolated and constructed from possible combinations and configurations of N_1 atoms and N_e electrons in real space".	[40]
Virshup et al.	"An M-dimensional cartesian space in which compounds are located by a set of M physicochemical and/or chemoinformatic descriptors".	[2]
Vogt	"Comprehensive collection of all possible small molecules under some reasonable restrictions considering size and composition".	[41]
Huang and Lilienfeld	"Chemical compound space is the set of all theoretically conceivable combinations of chemical elements and (meta-)stable geometries that make up matter".	[32]

[a] In chronological order of publication.

In physics, Everett's multiverse[42] is "a hypothetical collection of potentially diverse observable universes, each of which would comprise everything that is experimentally accessible by a connected community of observers."[43] In

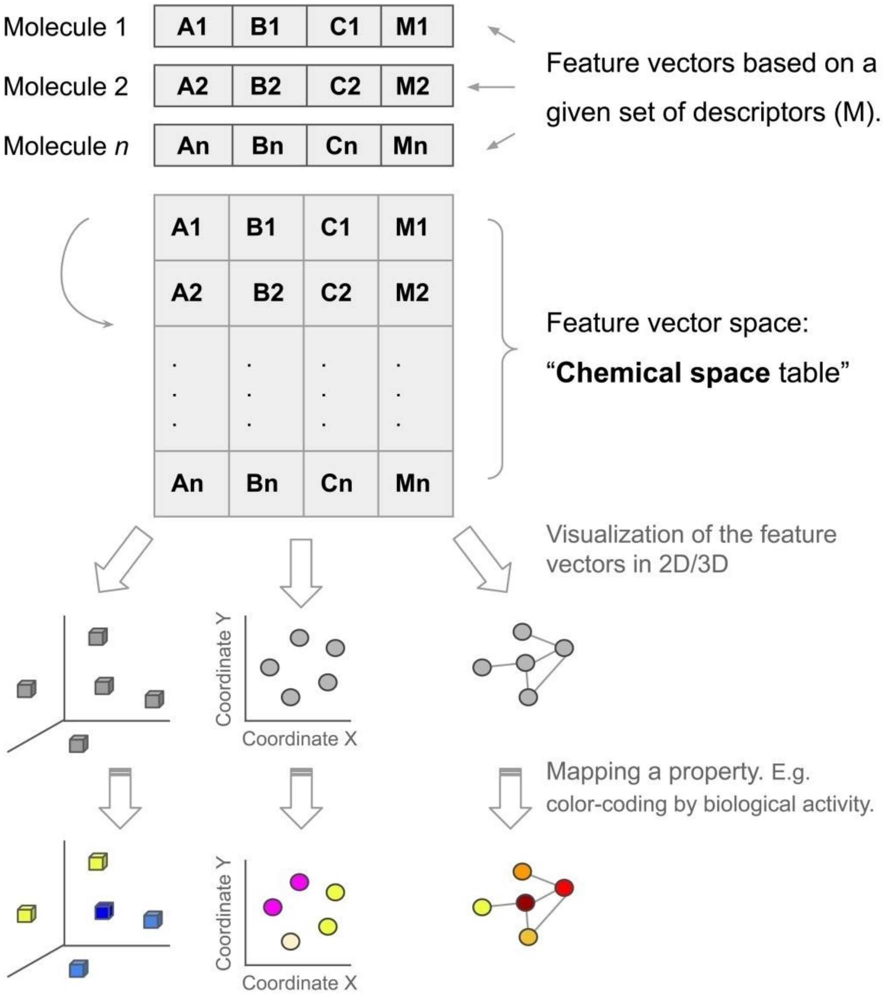

Molecule 1
A1 B1 C1 M1
Molecule 2
A2 B2 C2 M2
Molecule *n*
An Bn Cn Mn
Feature vectors based on a given set of descriptors (M).
A1 B1 C1 M1
A2 B2 C2 M2
An Bn Cn Mn
Feature vector space:
"**Chemical space** table"
Visualization of the feature vectors in 2D/3D
Coordinate Y
Coordinate X
Mapping a property. E.g. color-coding by biological activity.
Coordinate Y
Coordinate X

Figure 1. Schematic view of the chemical space. Each molecule in the compound data set is represented by *M* descriptors that lead to a feature vector space. The group of *n* molecules represented with *M* descriptors form a "chemical space table" that can be represented using different visualization methods. Mapping a property (e.g., biological activity) to the "chemical space table" or the visual representation gives rise to a chemogenomic space that is the basis to do structure-property (activity) relationships.

other words, the multiverse "is a hypothetical group of multiple universes," and regions in the universe detached from one another exhibit distinct properties.[44]

By rough analogy with the cosmic multiverse, here we introduce the term *chemical multiverse* as the group of numerical vectors that describe it differently from the same set of molecules. In other words, a chemical multiverse is a group of multiple chemical spaces, each one defined by a given set of descriptors e.g., a group of "descriptor universes". Furthermore, and maintaining the analogy with the cosmology megaverse, a "chemical megaverse" is the collection of chemical multiverses. This would be given by the several different sets of descriptors that can be used to define a chemical space (Figure 1). As discussed in previous papers (Table 2), different chemical space representations lead to other spaces, and relationships between chemical compounds could be maintained or not.[20]

The concept of the *chemical multiverse* is schematically shown in Figure 2. Three grids ("chemical space tables") encode different chemical spaces, each described by a different set of descriptors. For schematic and illustrative purposes, the blue triangles encode a chemical multiverse defined by Lipinksi's "Rule-of-five" that displays favorable pharmacokinetic properties in terms of absorption and distribution. The orange cylinders represent a chemical space described by molecular fingerprints as Extended Connectivity Fingerprints (ECFP). The pink cubes encode the chemical space defined by constitutional descriptors like carbon, nitrogen, and oxygen atoms, ring counts, and bridgehead atoms. Of course, any other set of descriptors can be used. All three chemical spaces in Figure 3 comprise the chemical multiverse of the data set with *n* atoms. Again, the chemical multiverse could be formed by more than three chemical spaces, depending on how many different sets of descriptors need to be used to meet the study's

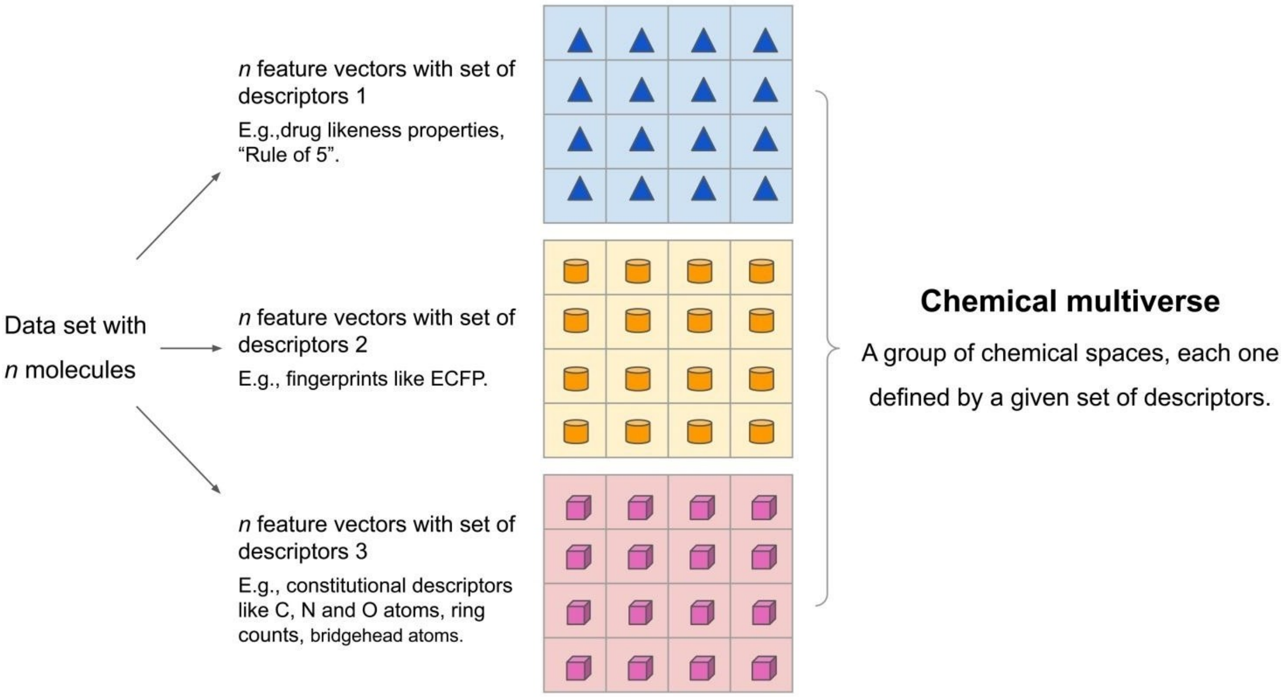

n feature vectors with set of descriptors 1
E.g.,drug likeness properties, "Rule of 5".
Data set with n molecules
n feature vectors with set of descriptors 2
E.g., fingerprints like ECFP.
n feature vectors with set of descriptors 3
E.g., constitutional descriptors like C, N and O atoms, ring counts, bridgehead atoms.
Chemical multiverse
A group of chemical spaces, each one defined by a given set of descriptors.

Figure 2. Schematic and general representation of the concept of the chemical multiverse. The chemical multiverse of the same data set with *n* molecules would be composed of several (shown in figure three for illustrative purpose) alternative chemical spaces, each one defined by a different set of descriptors. The geometric figures represent the encoding of the structures using different descriptors: e.g., drug-likeness properties (blue triangles), fingerprints (orange cylinders), and constitutional descriptors (such as ring counts, carbon, nitrogen, oxygen, and bridgehead atoms) (pink cubes). Depending on the study's goals, a chemical multiverse could contain as many chemical spaces as needed. Each chemical space (in the middle section: "chemical space tables") could be subject to different 2D/3D visual representations of the chemical space.

goals. Section 4 presents case studies, most of them published in the literature, of chemical multiverses using real data sets (albeit the name "chemical multiverse" has not been used in previous publications).

The importance of the chemical multiverse concept relies on the fact that a comprehensive view of the chemical space of a data set should be given by several representations, as opposed to a single one. This is because a single set of descriptors is limited and does not capture all aspects of the chemical structures. The need to consider multiple descriptors for chemoinformatics applications has been broadly recognized, for instance, in similarity searching[45] and diversity analysis.[46] A case in point is the so-called consensus diversity plots that consider at least four types of molecular representation to gain a more comprehensive view of the "global" or total diversity of compound data sets in comparative studies. In consensus diversity plots, types of representation such as scaffolds, structural fingerprints, physicochemical properties, and metrics associated with structural complexity have been employed.[46,47]

3.1 Chemical Multiverse vs. Consensus Chemical Space

The consensus chemical space can be conceptualized as the result of the fusion or combination of the different descriptors into one. Thus, it is highly associated with the concept of data fusion.[48] Under this concept, the consensus chemical space is dependent on the approach to combining the descriptors of the molecules or making hybrid representations.[49] A consensus chemical space can also be seen as the combination of similarity metrics, also reminiscent of data fusion, to generate a unified representation of the chemical space. In contrast, the chemical multiverse, as introduced in the previous section, does not require a fusion of a combination of descriptors for each molecule: the chemical multiverse is composed of a group of alternative representations of the compounds (e.g., using different fingerprints, properties, or descriptors for a given dataset). Figure 3 shows the difference and relationship between the chemical multiverse and consensus chemical space of a given set of *n* molecules.

While the consensus chemical space combines into one several possible universes created by different sets of descriptions, the chemical multiverse herein is conceptualized as a set of parallel universes. Thus, the chemical

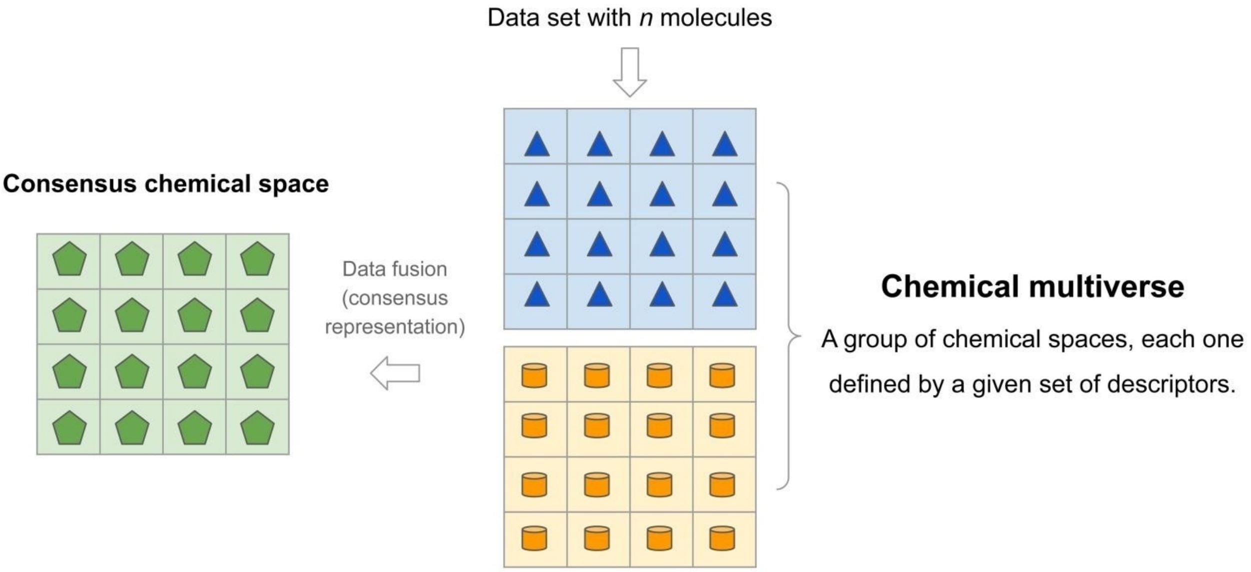

Data set with *n* molecules
Consensus chemical space
Data fusion
(consensus
representation)
Chemical multiverse
A group of chemical spaces, each one
defined by a given set of descriptors.

Figure 3. General comparison of a chemical multiverse with consensus chemical space. While a chemical multiverse of a compound data set is a set of alternative chemical spaces (shown in figure only two, for illustration purposes), a consensus chemical space is the combination of the alternative chemical spaces to yield one single chemical space that is the result of data fusion or combination of the descriptors. The grids with blue triangles and orange cylinders encode two different chemical spaces described by, for instance, drug-likeness and ECFP descriptors, respectively. In this schematic example, the green hexagons represent the fusion or combination of the two descriptors to lead to a new but single consensus chemical space.

multiverse could be a better option to handle the chemical spaces because each one will provide information associated with the particular descriptor used and, therefore, be easier to interpret. In contrast, in a consensus representation, the result could be hard to interpret due to combining several representations into one.

Analogous to the molecular representations,[50] there is no unique and global chemical universe. Whereas a consensus chemical space is an attempt to generate "a single chemical universe," the relevant information from a set of descriptors can be lost due to combining the chemical spaces, as depicted in Figure 3.

4 Examples of Chemical Multiverses

This section reviews case studies of chemical multiverses applied to different data sets of molecules (though the term "chemical multiverse" has not been used before). The chemical multiverses could be analyzed, navigated, and compared, considering the full-dimension space, similar to individual chemical spaces.[30] Even though the chemical space is typically associated with small organic molecules for drug discovery applications (see exemplary proposed definitions of chemical space in Table 2), we emphasize that the chemical space and chemical multiverse could involve other types of chemical compounds. Table 3 summarizes recent studies of the chemical space using multiple structure representations for different aims. The studies are briefly discussed below.

It has been noted that different similarity measures generate different chemical spaces. For instance, in 2014 Medina-Franco and Maggiora illustrated the diversity of nine compound data sets as they are projected into 3D obtained from pairwise structure similarity computed with the Tanimoto coefficient and four fingerprints of different designs.[52] For comparison, the authors generated a single visualization of the chemical space obtained by the mean fusion of the four different sets of similarity values (an example of representation of a consensus chemical space). From the visual representation of the chemical multiverse and consensus chemical space was concluded that, albeit graphically, those compound neighborhoods will not remain invariant to changes in molecular representation.

In separate work, Casciuc et al. showed the complementarity of seven GTMs based on a different descriptor space to classify actives and inactive compounds in ChEMBL. In that work, the authors also compared the performance of virtual screening of the complementarity of different maps vs. a consensus map concluding that "while any single universal map has moderate predictive power, the combination of complementary maps lead to a more robust consensus effect in virtual screening."[17] Also using GTMs, Zabolotna et al. presented "NP-Navigator", a freely available intuitive online tool for visualization and navigation through the chemical space of NPs and NP-like molecules.[8] The different representations generated in NP-Navigator

Table 3. Examples of studies of chemical spaces of compound data sets using multiple representations.[a]

Study aims	Data sets and molecular representations and descriptors	Ref.
Structure-activity relationships	2D representations of the chemical space combined with biological activity using 11 2D and 3D structural representations. The therein generated activity landscapes were used to identify consensus activity cliffs.	[51]
Diversity analysis	3D projections of PCA-based chemical spaces generated from a set of 2250 compounds obtained from nine datasets of 250 compounds each using four fingerprints (atom pairs, MACCS keys, TGD, and piDAPH4) and the Tanimoto coefficient. The visualizations of the chemical space were employed to analyze the diversity of the data sets.	[52]
Virtual screening	Eight universal GTMs each generated with different descriptor vectors (In SIlico design and Data Analysis – ISIDA descriptors), each encoding distinct structural features, were employed as support for predictive classification landscapes.	[17]
Compound library design	Four PCA plots of molecular quantum number fingerprints to assess the quality of the training process in generative models.	[53]
Chemical space navigation	Visualization and navigation through the chemical space of natural products and natural products-like molecules considering chemotype distribution, physicochemical properties, biological activity, and commercial availability.	[8]
Diversity analysis and compound(s) selection	CLNs of 19 chemical libraries used in drug discovery and natural products research were generated using four fingerprints (MACCS keys, RDKit, and ECFP4), and the extended Tanimoto index. CLNs were used to compare the diversity of the data sets.	[54]
Diversity analysis	16 comparative ChEMBL vs. purchasable building blocks (PBB) landscapes using GTMs and ISIDA fragment descriptors. GTMs allowed the identification of the most represented and underrepresented classes of PBBs.	[55]

[a] In chronological order of publication.

allows to efficiently analyze different aspects of NPs such as chemotype distribution, physicochemical properties, biological activity, and commercial availability of NPs.

The use of multiple structure representations to explore the chemical space of compound data sets and its impact to analyze structure-activity relationships under the concept of activity landscapes has been shown and reviewed in the literature.[56] In an early work, 2D and 3D representations were used to identify activity cliffs of a data set of 48 bicyclic guanidines with κ-opioid receptor binding affinity. It was concluded that while some activity cliffs are dependent on the structure representation, there are activity cliffs that are consistent regardless of the representation explored, i.e. consensus activity cliffs.[51]

In order to generate efficient methods to quantify the diversity of large and ultra-large chemical libraries and visualize their mutual relationships in chemical space, Dunn et al. developed Chemical Library Networks (CLNs) based on extended similarity indices.[54] In this work, different CLNs of 19 chemical libraries used in drug discovery and NPs research were generated using MACCS keys (166-bits), RDKit, and ECFP4 fingerprints. The analysis and comparison of the generated CLNs led to the conclusion that the extended Tanimoto index offers the best description of extended similarity in combination with RDKit fingerprints. In subsequent work, Flores-Padilla et al. used CLNs and Constellation plots based on chemical core scaffolds to analyze 11 commercial libraries of different sizes focused on epigenetic targets (with 53443 compounds in total).[3] The chemical space content and diversity analysis based on different descriptors helped to identify the most diverse synthetic-focused screening libraries.

In the design of compound libraries, various representations of chemical space are also commonly used to assess the novelty, pharmaceutical properties, and molecular and shape diversity of the compounds generated. For example, to design combinatorial libraries it is common to analyze the chemical space based on physicochemical properties and compare it with sets of pharmaceutical relevance such as approved drugs. Structural novelty and diversity are also often visualized through chemical space based on molecular fingerprints. Other visualizations that are included in compound library design programs are the Principal Moments of Inertia plots. These graphs are mainly used in diversity-oriented synthesis approaches to ensure the diversity of shapes of the designed compounds.[57]

Recently, Arús-Pous et al. used different PCA plots of molecular quantum number fingerprints in compound library design to assess the quality of the training process in generative models.[53] In that study, the obtained plots showed that when using recurrent neural networks to train models that sample chemical space, complex molecules with many rings and heteroatoms are more difficult to sample than molecules with fewer rings and more carbon atoms.

The encoding of fragments or building blocks, as well as experimental data on reactions in computer-accessible formats, are opening new ways of representing chemical space. Integrating this information into the chemical space representation can facilitate the search for promising compounds in ultra-large data collections and focus the

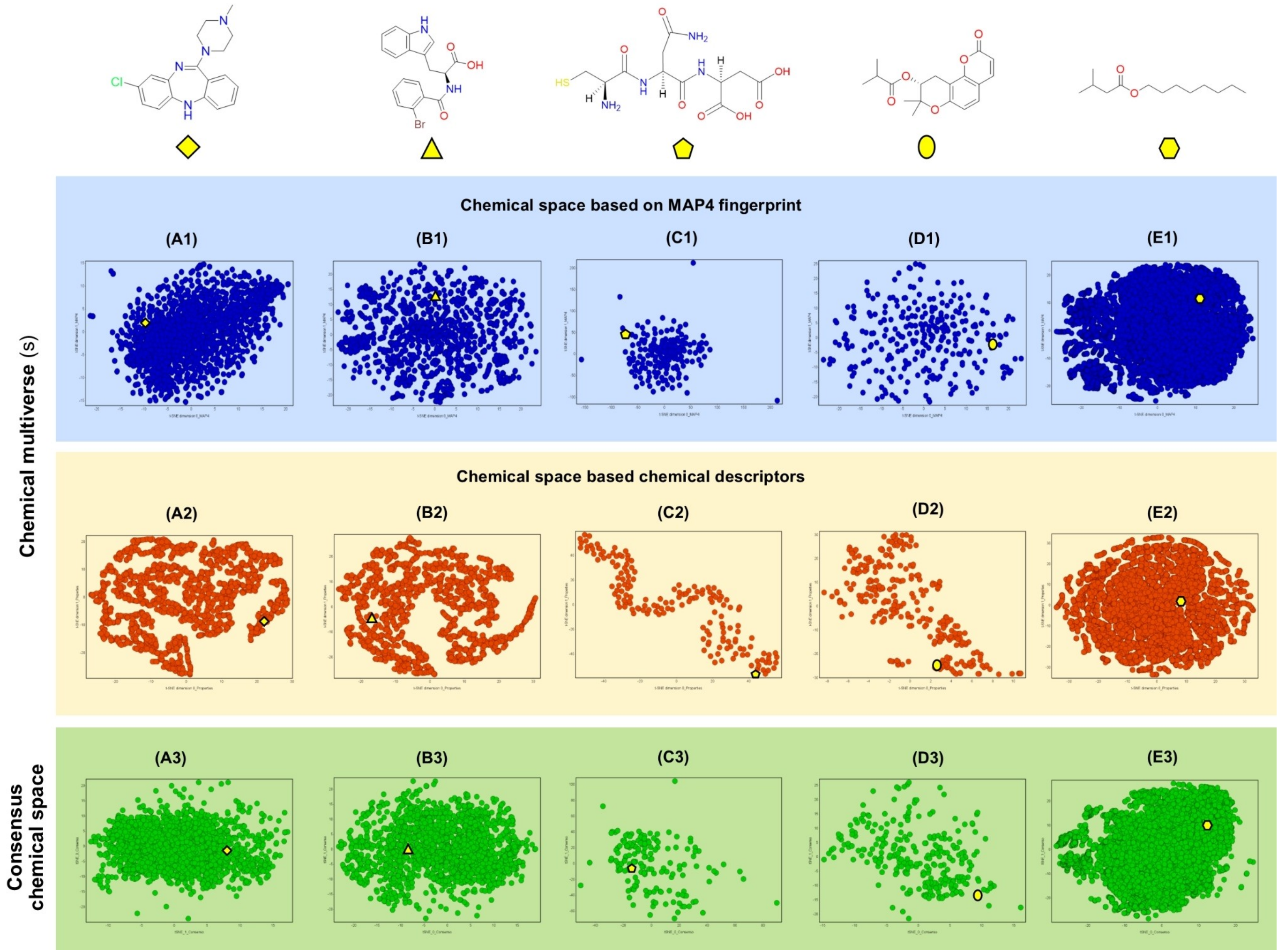


Figure 4. Example of the visual representation of chemical multiverse of four compound data sets: **A**) drug-like (2,403 compounds); **B**) protein-protein interaction inhibitors (2,227 compounds);[65] **C**) anti-MRSA peptides (165);[59] **D**) natural products (285 compounds from BIOFACQUIM database);[66] **E**) food chemicals (21,319 compounds from FooDB).[67] The visual representations were obtained from the t-SNE module implemented in KNIME. The chemical multiverse of each set is compared with a consensus representation of the chemical space obtained from averaging (i.e, average fusion rule) the coordinates of the data points. The upper part of this figure shows examples of representative chemical structures of each data set, each structure is represented by a yellow shape.

synthesis of new compounds in relevant medicinal chemistry spaces.[58] In another application of GTMs, Zabolotna et al. analyzed the diversity of more than 400,000 purchasable building blocks (PBBs) provided by eMolecules.[55] Comparison of PBBs with synthons derived from ChEMBL fragmentation revealed that the internal diversity among members of the same class of synthons is significantly better for ChEMBL-derived synthons, leaving room for the design and improvement of corresponding PBBs. Similarly, the existence of structurally equivalent synthons in ChEMBL can be used to search for alternative synthesis ways in situations where the same structural remainder can be provided by radically different reactivity BBs applicable in different synthetic routes.

Figure 4 shows additional examples of visualization of chemical multiverses for different data sets: different nature of chemical structures and a varied number of molecules. Of note, the purpose of the examples in this figure showing data sets of varying chemical nature is to illustrate the concept of the chemical multiverse. Still, we do not attempt to make a direct comparison between the data sets or discuss their diversity. Figure 4 shows a 2D visualization of the chemical multiverse (given by two chemical spaces, each obtained with a different set of descriptors) and a representative consensus chemical space of five data sets from various sources: small organic drug-like compounds, protein–protein interaction inhibitors, anti-*Staphylococcus aureus* methicillin-resistant (MRSA) peptides from the anti-microbial peptide database 3,[59] natural products, and food chemicals. The coordinates for each graph in Figure 4 were created using the "tSNE module"[14] of KNIME software (version 4.3.4).[60] Figure 4A1–E1 (top part of the figure, in blue) shows a visualization of the chemical space based on the recently developed "MinHashed atom-pair fingerprint up to a diameter of four bonds" (MAP4) fingerprint (2048 bits).[61] Figure 4A2–E2 (middle of the figure, in orange) shows the visual representation of the chemical space using a different set of chemical descriptors: SlogP, TPSA, MW, Rotatable bonds, NumHBD, NumHBA, NumStereocenters, FractionCSP3, that were calculated with the RDKIt module implemented in KNIME.[62] In this illustrative example, both chemical space representations in blue and orange of the five different data sets (A–E) shows a visualization of the chemical multiverse of the sets, in this case using t-SNE (other approaches to visualize the chemical space or chemical multiverses can be employed, as described in Section 3). The chemical multiverses for each data set, e.g., comparing the pair of chemical spaces A1—A2, B1—B2, C1—C2, D1—D2, and E1—E2, clearly show the dependence of the chemical space on the structure representation, in these cases MAP4 fingerprints and chemical descriptors. This is particularly dramatic for data sets such as peptides and NPs.

For comparison, Figure 4A3–E3 (bottom part of the figure, in green) illustrates an example of a visualization of a consensus chemical space of the same five data sets. For this example, the consensus representation was generated by taking the average of the t-SNE coordinates of each data set. Several other combinations of descriptors or fusion rules could be employed and finding "the best" one would not be straightforward. Of note, exploring and comparing different approaches to combine descriptors is beyond the scope of this review (in general) and the Figure 4 (in particular). As discussed in Section 3.1, the consensus representation of the chemical space would be directed to generating a "global" perception of the chemical universe, which is also valuable. To illustrate further the interdependence of data between the multiverse chemical spaces and consensus representation, the upper part of Figure 4 shows representative chemical structures of each data set. In general, the change in the relative position of each chemical structure in their chemical multiverse is condensed on the consensus representation of each one. However, the interpretation of the consensus chemical space becomes more complicated if one wants to associate the structure of property features that distinguish the compound data sets.

The interactive visualizations of chemical spaces were generated using DataWarrior software.[63,64] The visualizations are available on Figshare at https://doi.org/10.6084/m9.figshare.20483958.

In Figures 4 alternative and more molecular representations could be used to illustrate the concept of chemical multiverses of the various data sets in addition to the published examples reviewed in Table 3. For illustration, only a few molecular representations were used.

The survey of published case studies (Table 3) shows the relevance of using several descriptors for various type of applications, namely structure-activity relationships, diversity analysis, virtual screening, compound library design, visual representation of the chemical space, and compound selection. The examples in Figure 4 are intended to further illustrate that chemical multiverses can be applied to virtually any type of molecular structure such as small molecules as drug candidates, NPs, peptides, inhibitors of protein—protein interactions, peptides, and food chemicals, that have distinct structural features.

5 Summary and Outlook

The number and type of chemical structures relevant to drug discovery and other related applications are dramatically increasing. Similarly, the number and type of experimental or calculated descriptors also augment. The continued increase in the number of compounds and descriptors encourages novel ways to interact with the chemical space beyond the traditional medicinally relevant chemical space built based on drug- or lead-like properties of the pharmaceutical interest and the conventional organic small-molecules. Herein, we introduce the term *chemical multiverse* as a group of chemical spaces, each one defined by a given set of descriptors. By its own nature, a chemical multiverse provides more information about a single chemical space defined by a specific molecular representation. In this review, we have shown that the use of multiple descriptors to study the chemical space has been implemented in several studies as a single type of descriptor is not enough to capture all the required features for structure-activity relationships, or other applications such as diversity analysis, virtual screening, compound library design, visual representation of the chemical space, and compound selection. A "chemical megaverse" is the collection of chemical multiverses, and this would be given by the several different groups of descriptors that can be used to define a chemical space.

For some specific compound data sets, for instance, metal-containing molecules and other types of complex molecules, it remains to determine the chemical multiverses consistently: very much like in cosmology, to explore uncharted regions of the chemical space. Also, it remains to explore how to move from one universe to another (aka, navigating between chemical spaces or *navigating the chemical multiverse* of a compound data set), similar to physics or astronomy, but we can find a point where each multiverse connects or overlaps, and we can observe these points in the visualizations of chemical spaces.

Beyond this review paper that surveys studies that use different descriptors for various drug discovery applications, it remains to compare, in a research study, different ways to combine descriptors and, in general, generate consensus representations of chemical spaces.

As illustrated in this manuscript, same as the chemical space concept, the chemical multiverse and chemical megaverse have applications in several research areas, including drug discovery and beyond.

Abbreviations

2D, two-dimensional, 3D, three-dimensional; CLNs, Chemical Library Networks; ECFP, extended connectivity fingerprint; FDA, Food and Drug Administration; GTM, Generative Topographic Mapping; HIV, human immunodeficiency virus; MAP4, MinHashed atom-pair fingerprint up to a diameter of four bonds; MRSA, anti-*Staphylococcus aureus* methicillin-resistant; NPs, natural products; PCA, principal component analysis; Rule of five, favorable pharmacokinetic properties in terms of absorption and distribution, described by molecular weight (MW) ≤ 500, partition coefficient octanol/water (SlogP) ≤ 5, hydrogen bond acceptors (HBA) ≤ 10, hydrogen bond donors (HBD) ≤ 5; SOM, self-organized map; SAR, structure-activity relationships; t-SNE, t-distributed stochastic neighbor embedding.

Author Contribution Statement

All authors have contributed equally to the present manuscript.

Acknowledgements

A.L.-C.-H., E.L.-L., and F.I.S.-G are grateful to Consejo Nacional de Ciencia y Tecnología (CONACyT), Mexico, for the Ph.D. scholarship numbers, 847870, 814741, and 848061, respectively. We also thank the funding of DGAPA, UNAM, Programa de Apoyo a Proyectos de Investigación e Innovación Tecnológica (PAPIIT), grant No. IN201321.

Conflict of Interest

None declared.

Data Availability Statement

The data that supports the findings of this study are freely available on Figshare at https://doi.org/10.6084/m9.figshare.19768174.

References

[1] G. M. Maggiora, in *Foodinformatics: Applications of Chemical Information to Food Chemistry* (Eds.: K. Martinez-Mayorga, J. L. Medina-Franco), Springer International Publishing, Cham, **2014**, pp. 1–81.

[2] A. M. Virshup, J. Contreras-García, P. Wipf, W. Yang, D. N. Beratan, *J. Am. Chem. Soc.* **2013**, *135*, 7296–7303.

[3] E. A. Flores-Padilla, K. E. Juárez-Mercado, J. J. Naveja, T. D. Kim, R. Alain Miranda-Quintana, J. L. Medina-Franco, *Mol. Inf.* **2021**, *41*, e2100285.

[4] D. L. Prado-Romero, J. L. Medina-Franco, *ACS Omega* **2021**, *6*, 22478–22486.

[5] A. Yoshimori, F. Miljković, J. Bajorath, *Molecules* **2022**, *27*, 570.

[6] A. L. Chávez-Hernández, K. E. Juárez-Mercado, F. I. Saldívar-González, J. L. Medina-Franco, *Biomol. Eng.* **2021**, *11*,1805.

[7] Y. Chen, M. Garcia de Lomana, N.-O. Friedrich, J. Kirchmair, *J. Chem. Inf. Model.* **2018**, *58*, 1518–1532.

[8] Y. Zabolotna, P. Ertl, D. Horvath, F. Bonachera, G. Marcou, A. Varnek, *Mol. Inf.* **2021**, *40*, e2100068.

[9] M. Grigalunas, S. Brakmann, H. Waldmann, *J. Am. Chem. Soc.* **2022**, *144*, 3314–3329.

[10] L. Ruddigkeit, J.-L. Reymond, in *Foodinformatics: Applications of Chemical Information to Food Chemistry* (Eds.: K. Martinez-Mayorga, J. L. Medina-Franco), Springer International Publishing, Cham, **2014**, pp. 83–96.

[11] A. Capecchi, J.-L. Reymond, *Med. Drug Discovery* **2021**, *9*, 100081.

[12] E. Meggers, *Curr. Opin. Chem. Biol.* **2007**, *11*, 287–292.

[13] R. Schmidt, R. Klein, M. Rarey, *J. Chem. Inf. Model.* **2022**, *62*, 2133–2150.

[14] L. van der Maaten, *J. Mach. Learn. Res.* **2008**, *1*, 1–48.

[15] I. T. Jolliffe, *Principal Component Analysis*, Springer Science & Business Media, **2002**.

[16] T. Kohonen, in *Self-Organizing Maps* (Ed.: T. Kohonen), Springer Berlin Heidelberg, Berlin, Heidelberg, **1995**, pp. 77–130.

[17] I. Casciuc, Y. Zabolotna, D. Horvath, G. Marcou, J. Bajorath, A. Varnek, *J. Chem. Inf. Model.* **2019**, *59*, 564–572.

[18] D. Horvath, G. Marcou, A. Varnek, *Drug Discovery Today Technol.* **2019**, *32–33*, 99–107.

[19] Y. Zabolotna, A. Lin, D. Horvath, G. Marcou, D. M. Volochnyuk, A. Varnek, *J. Chem. Inf. Model.* **2021**, *61*, 179–188.

[20] G. M. Maggiora, J. Bajorath, *J. Comput.-Aided Mol. Des.* **2014**, *28*, 795–802.

[21] A. de la Vega de León, J. Bajorath, *Future Med. Chem.* **2016**, *8*, 1769–1778.

[22] W. A. Warr, M. C. Nicklaus, C. A. Nicolaou, M. Rarey, *J. Chem. Inf. Model.* **2022**, *62*, 2021–2034.

[23] L. Bellmann, P. Penner, M. Rarey, *J. Chem. Inf. Model.* **2021**, *61*, 238–251.

[24] W. P. Walters, *J. Med. Chem.* **2019**, *62*, 1116–1124.

[25] J.-L. Medina-Franco, K. Martinez-Mayorga, M. Giulianotti, R. Houghten, C. Pinilla, *Curr. Comput.-Aided Drug Des.* **2008**, *4*, 322–333.

[26] J.-L. Reymond, M. Awale, *ACS Chem. Neurosci.* **2012**, *3*, 649–657.

[27] J.-L. Reymond, *Acc. Chem. Res.* **2015**, *48*, 722–730.

[28] D. I. Osolodkin, E. V. Radchenko, A. A. Orlov, A. E. Voronkov, V. A. Palyulin, N. S. Zefirov, *Expert Opin. Drug Discovery* **2015**, *10*, 959–973.

[29] J. L. Medina-Franco, J. J. Naveja, E. López-López, *Drug Discovery Today* **2019**, *24*, 2162–2169.

[30] Y. Chen, J. Kirchmair, *Mol. Inf.* **2020**, *39*, e2000171.

[31] C. W. Coley, *Trends Chem.* **2021**, *3*, 133–145.

[32] B. Huang, O. A. von Lilienfeld, *Chem. Rev.* **2021**, *121*, 10001–10036.

[33] J. L. Medina-Franco, N. Sánchez-Cruz, E. López-López, B. I. Díaz-Eufracio, *J. Comput.-Aided Mol. Des.* **2022**, *36*, 341–354, DOI 10.1007/s10822-021-00399–1.

[34] M. Vogt, *Expert Opin. Drug Discovery* **2022**, *17*, 297–304.

[35] F. I. Saldívar-González, J. L. Medina-Franco, *Expert Opin Drug Discov.* **2022**, *17*, 789–798, DOI: 10.1080/17460441.2022.2084608.

[36] C. M. Dobson, *Nature* **2004**, *432*, 824–828.

[37] C. Lipinski, A. Hopkins, *Nature* **2004**, *432*, 855–861.

[38] J.-L. Reymond, R. van Deursen, L. C. Blum, L. Ruddigkeit, *MedChemComm* **2010**, *1*, 30–38.

[39] A. Varnek, I. I. Baskin, *Mol. Inf.* **2011**, *30*, 20–32.

[40] O. A. von Lilienfeld, *Int. J. Quantum Chem.* **2013**, *113*, 1676–1689.

[41] M. Vogt, *Expert Opin. Drug Discovery* **2020**, *15*, 523–525.

[42] H. Everett, Hugh Everett Theory of the Universal Wavefunction, Thesis, Princeton University, **1957**.

[43] A. Aguirre, *Encyclopedia Britannica* **2022**.

[44] H. Kragh, *Ann. Sci.* **2009**, *66*, 529–551.

[45] R. P. Sheridan, S. K. Kearsley, *Drug Discovery Today* **2002**, *7*, 903–911.

[46] M. González-Medina, F. D. Prieto-Martínez, J. R. Owen, J. L. Medina-Franco, *J. Cheminf.* **2016**, *8*, 63.

[47] J. J. Naveja, M. P. Rico-Hidalgo, J. L. Medina-Franco, *F1000Research* **2018**, *7* (Chem Inf Sci):993. DOI: 10.12688/f1000research.15440.2.

[48] M. Whittle, V. J. Gillet, P. Willett, J. Loesel, *J. Chem. Inf. Model.* **2006**, *46*, 2206–2219.

[49] B. Nisius, J. Bajorath, *Chem. Biol. Drug Des.* **2010**, *75*, 152–160.

[50] D. S. Wigh, J. M. Goodman, A. A. Lapkin, *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2022**, in press, DOI 10.1002/wcms.1603.

[51] J. L. Medina-Franco, K. Martínez-Mayorga, A. Bender, R. M. Marín, M. A. Giulianotti, C. Pinilla, R. A. Houghten, *J. Chem. Inf. Model.* **2009**, *49*, 477–491.

[52] J. L. Medina-Franco, G. M Maggiora, In *Chemoinformatics for Drug Discovery* (Ed. J. Bajorath), John Wiley & Sons, Inc. **2014**, pp. 343–399.

[53] J. Arús-Pous, T. Blaschke, S. Ulander, J.-L. Reymond, H. Chen, O. Engkvist, *J. Cheminf.* **2019**, *11*, 20.

[54] T. B. Dunn, G. M. Seabra, T. D. Kim, K. E. Juárez-Mercado, C. Li, J. L. Medina-Franco, R. A. Miranda-Quintana, *J. Chem. Inf. Model.* **2022**, *62*, 2186–2201.

[55] Y. Zabolotna, D. M. Volochnyuk, S. V. Ryabukhin, D. Horvath, K. S. Gavrilenko, G. Marcou, Y. S. Moroz, O. Oksiuta, A. Varnek, *J. Chem. Inf. Model.* **2022**, *62*, 2171–2185.

[56] J. L. Medina-Franco, *J. Chem. Inf. Model.* **2012**, *52*, 2485–2493.

[57] F. I. Saldívar-González, J. L. Medina-Franco, in *Small Molecule Drug Discovery* (Eds.: A. Trabocchi, E. Lenci), Elsevier, **2020**, pp. 83–102.

[58] M. Rarey, M. C. Nicklaus, W. Warr, *J. Chem. Inf. Model.* **2022**, *62*, 2009–2010.

[59] G. Wang, X. Li, Z. Wang, *Nucleic Acids Res.* **2016**, *44*, D1087–93.

[60] "KNIME," can be found under https://www.knime.com/ (last accessed 11 July, 2022).

[61] A. Capecchi, D. Probst, J.-L. Reymond, *J. Cheminf.* **2020**, *12*, 43.

[62] G. Landrum, "RDKit," can be found under http://www.rdkit.org, (last accessed 11 July, 2022).

[63] T. Sander, J. Freyss, M. von Korff, C. Rufener, *J. Chem. Inf. Model.* **2015**, *55*, 460–473.

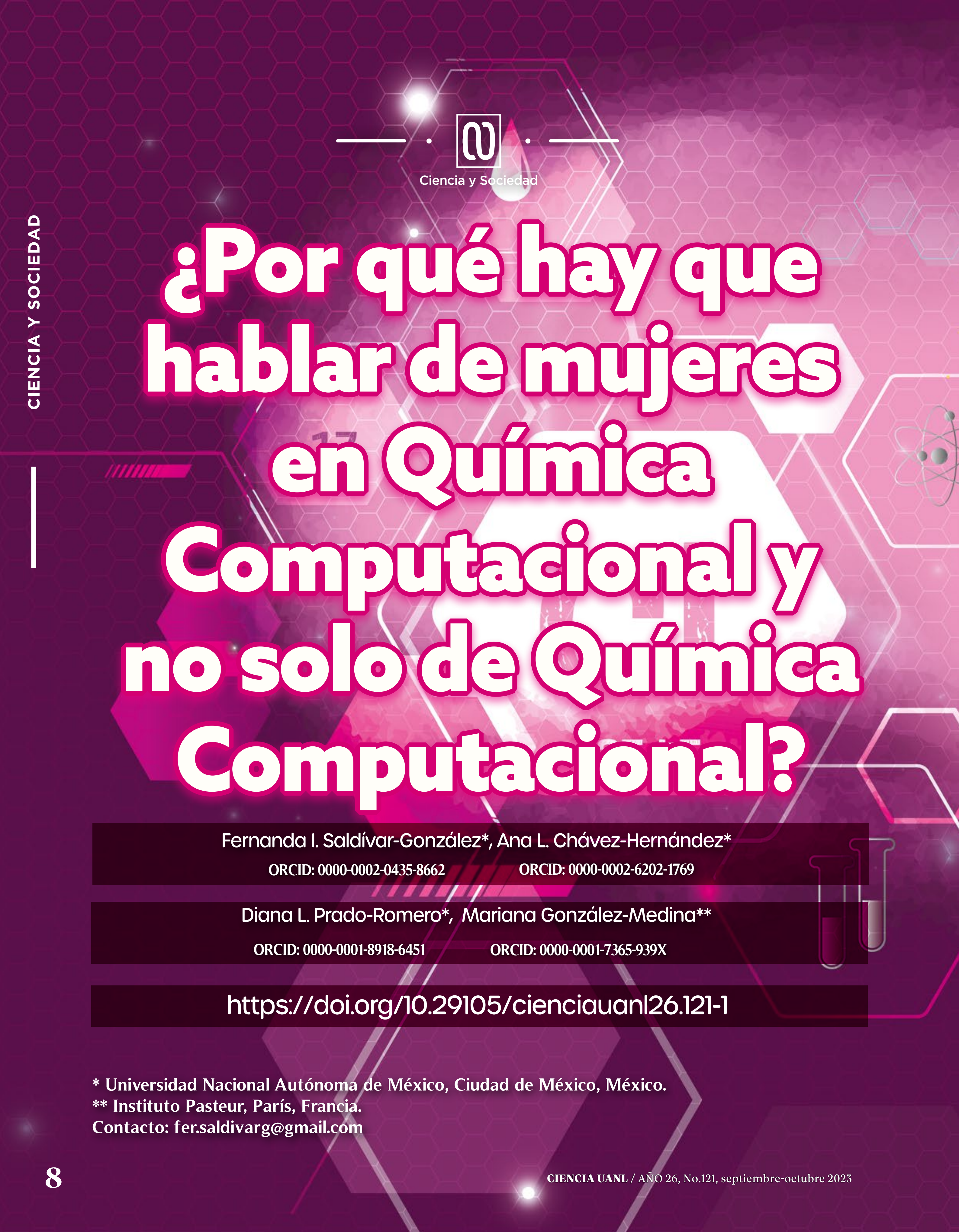
[64] E. López-López, J. J. Naveja, J. L. Medina-Franco, *Expert Opin. Drug Discovery* **2019**, *14*, 335–341.

[65] B. I. Díaz-Eufracio, O. Palomino-Hernández, A. Arredondo-Sánchez, J. L. Medina-Franco, *Mol. Inf.* **2020**, *39*, e2000035.

[66] B. A. Pilón-Jiménez, F. I. Saldívar-González, B. I. Díaz-Eufracio, J. L. Medina-Franco, *Biomol. Eng.* **2019**, *9*, 31.

[67] "FooDB," can be found under http://www.foodb.ca, (last accessed 11 July, 2022).

Received: May 19, 2022
Accepted: August 1, 2022
Published online on August 23, 2022



¿Por qué hay que hablar de mujeres en Química Computacional y no solo de Química Computacional?

Fernanda I. Saldívar-González*, Ana L. Chávez-Hernández*

ORCID: 0000-0002-0435-8662 ORCID: 0000-0002-6202-1769

Diana L. Prado-Romero*, Mariana González-Medina**

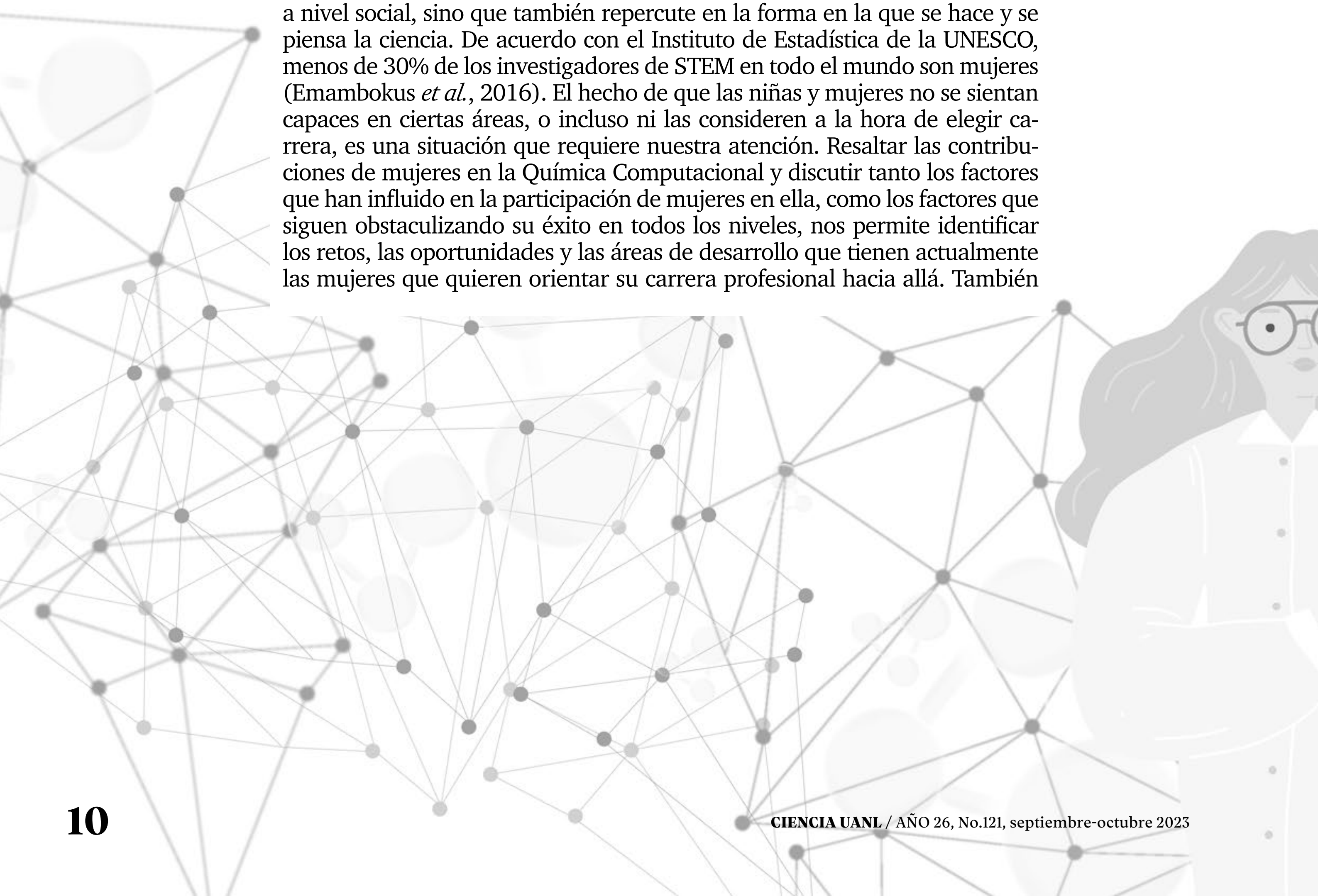
ORCID: 0000-0001-8918-6451 ORCID: 0000-0001-7365-939X

https://doi.org/10.29105/cienciauanl26.121-1

* Universidad Nacional Autónoma de México, Ciudad de México, México.
** Instituto Pasteur, París, Francia.
Contacto: fer.saldivarg@gmail.com

Cuando se habla de Química, lo más común es imaginar a alguien dentro de un laboratorio, portando *goggles*, vistiendo una bata y trabajando con matraces (el llamado "laboratorio húmedo" o *wet lab*). Sin embargo, también existen hombres y mujeres que están detrás de una computadora haciendo experimentos, pero con algoritmos (el llamado "laboratorio seco" o *dry lab*). La Química Computacional es una disciplina que se nutre en gran medida de datos experimentales generados en un laboratorio húmedo. La idea de usar computadoras es transformar estos datos químicos (llámense reacciones químicas, compuestos químicos, datos de actividad biológica, etc.) en información y ésta en conocimiento, lo cual permite reducir costos y eficientar procesos. Por esa razón ha tenido un gran impacto en la sociedad y cada vez hay un mayor número de aplicaciones que se ven reflejadas en un incremento de artículos y publicaciones científicas (Damm-Ganamet *et al.*, 2020).

Pero, ¿por qué es necesario enfatizar la labor que han hecho las mujeres en la Química Computacional y no sólo hablar de esta disciplina? La baja representación de mujeres en espacios STEM (del inglés ciencia, tecnología, ingeniería y Matemáticas) es un problema multifactorial que no sólo afecta a nivel social, sino que también repercute en la forma en la que se hace y se piensa la ciencia. De acuerdo con el Instituto de Estadística de la UNESCO, menos de 30% de los investigadores de STEM en todo el mundo son mujeres (Emambokus *et al.*, 2016). El hecho de que las niñas y mujeres no se sientan capaces en ciertas áreas, o incluso ni las consideren a la hora de elegir carrera, es una situación que requiere nuestra atención. Resaltar las contribuciones de mujeres en la Química Computacional y discutir tanto los factores que han influido en la participación de mujeres en ella, como los factores que siguen obstaculizando su éxito en todos los niveles, nos permite identificar los retos, las oportunidades y las áreas de desarrollo que tienen actualmente las mujeres que quieren orientar su carrera profesional hacia allá. También

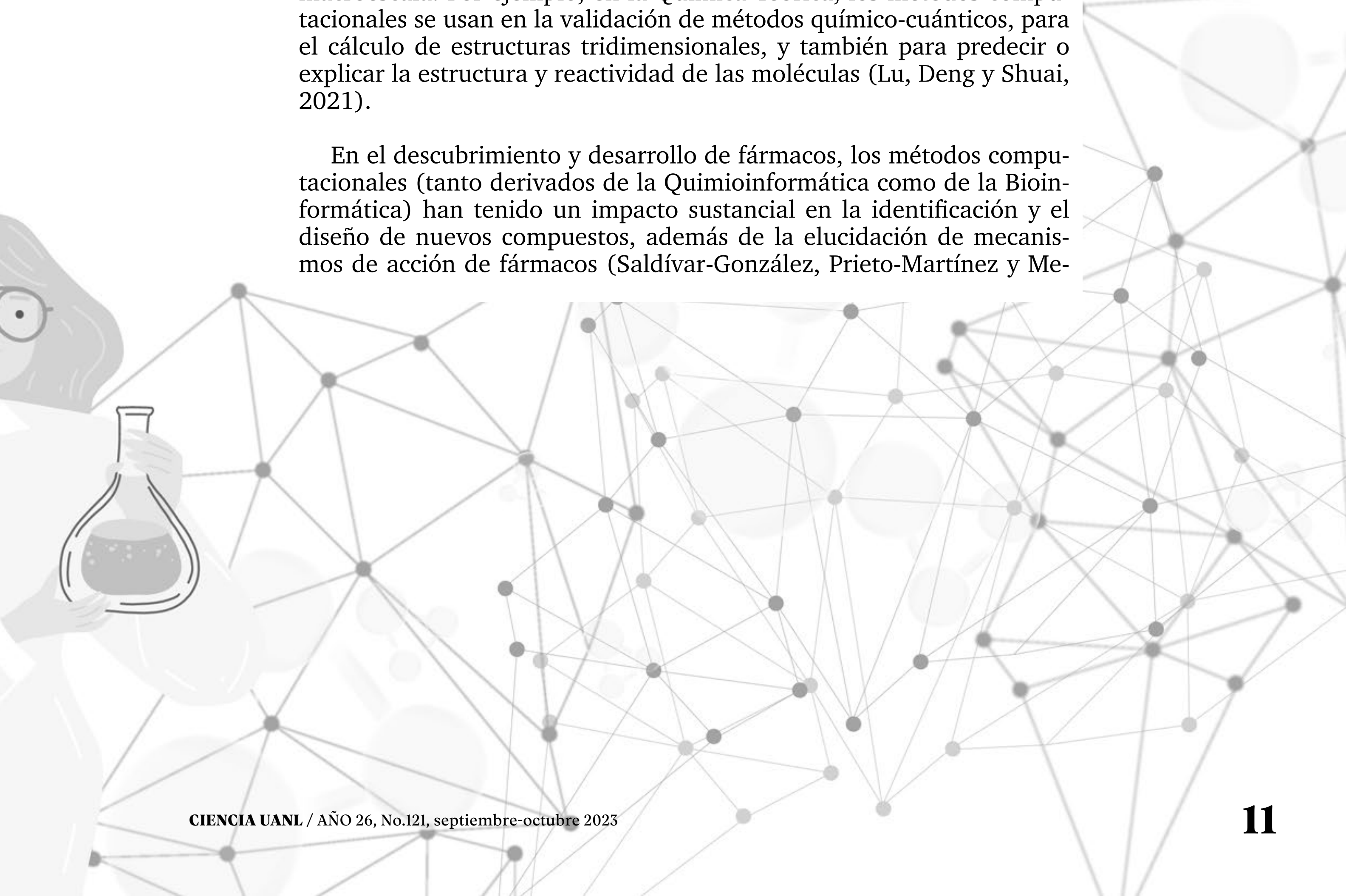


es importante dar difusión a este tema para avanzar en la construcción de una ciencia con perspectiva de género en la que las aplicaciones científicas beneficien a una mayor parte de la sociedad.

¿QUÉ ES LA QUÍMICA COMPUTACIONAL Y QUÉ APLICACIONES TIENE?

Existen tantas definiciones de Química Computacional que pueden llegar a ser confusas, debido a que, por mucho tiempo, el término fue usado para describir lo que ahora representa la Química Teórica. Otras definiciones se enfocan demasiado en los métodos utilizados, excluyendo campos como la Quimioinformática. Para usos prácticos, y para dar una visión más amplia de lo que representan los métodos computacionales en Química, utilizaremos la definición del Dr. Gabriel Cuevas, quien la menciona como una disciplina que comprende todos aquellos aspectos de la investigación en Química que se benefician de la aplicación de las computadoras (Cuevas, 2005). La figura 1 resume algunas áreas que comprende y sus principales aplicaciones. Como podemos ver, aborda cuestionamientos desde el nivel electrónico y atómico hasta el nivel macroescala. Por ejemplo, en la Química Teórica, los métodos computacionales se usan en la validación de métodos químico-cuánticos, para el cálculo de estructuras tridimensionales, y también para predecir o explicar la estructura y reactividad de las moléculas (Lu, Deng y Shuai, 2021).

En el descubrimiento y desarrollo de fármacos, los métodos computacionales (tanto derivados de la Quimioinformática como de la Bioinformática) han tenido un impacto sustancial en la identificación y el diseño de nuevos compuestos, además de la elucidación de mecanismos de acción de fármacos (Saldívar-González, Prieto-Martínez y Me-



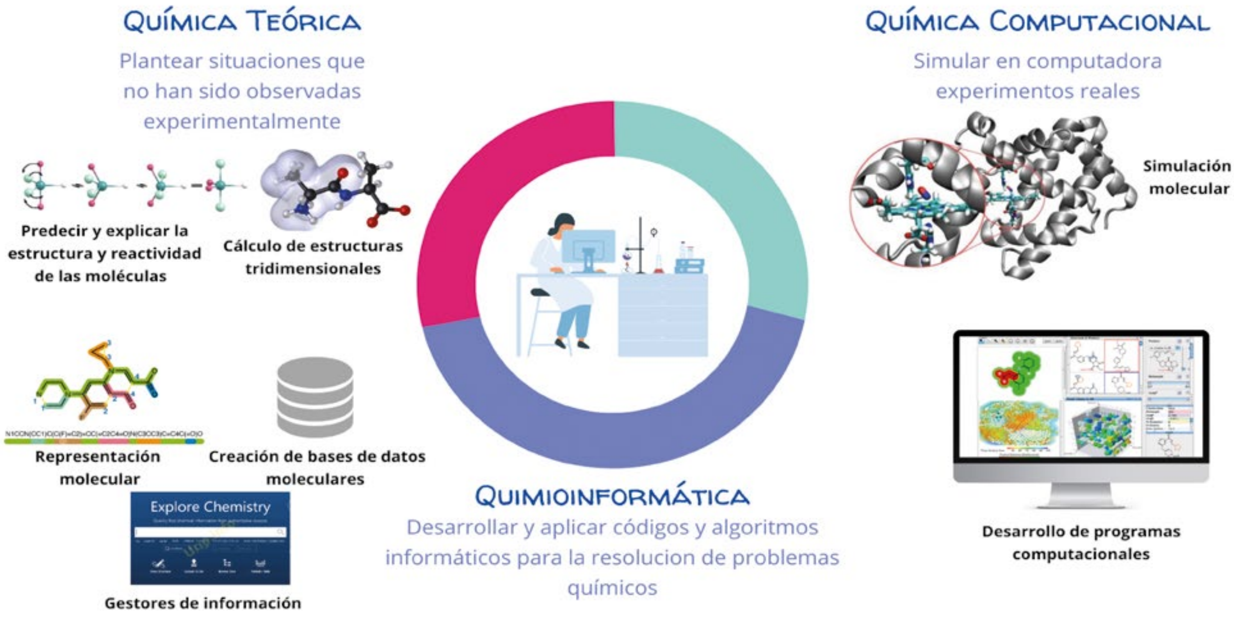

Química Teórica
Plantear situaciones que no han sido observadas experimentalmente
Predecir y explicar la estructura y reactividad de las moléculas
Cálculo de estructuras tridimensionales
Química Computacional
Simular en computadora experimentos reales
Simulación molecular
Representación molecular
Creación de bases de datos moleculares
Explore Chemistry
Gestores de información
Quimioinformática
Desarrollar y aplicar códigos y algoritmos informáticos para la resolucion de problemas químicos
Desarrollo de programas computacionales

Figura 1. Áreas de la Química Computacional y sus principales aplicaciones.

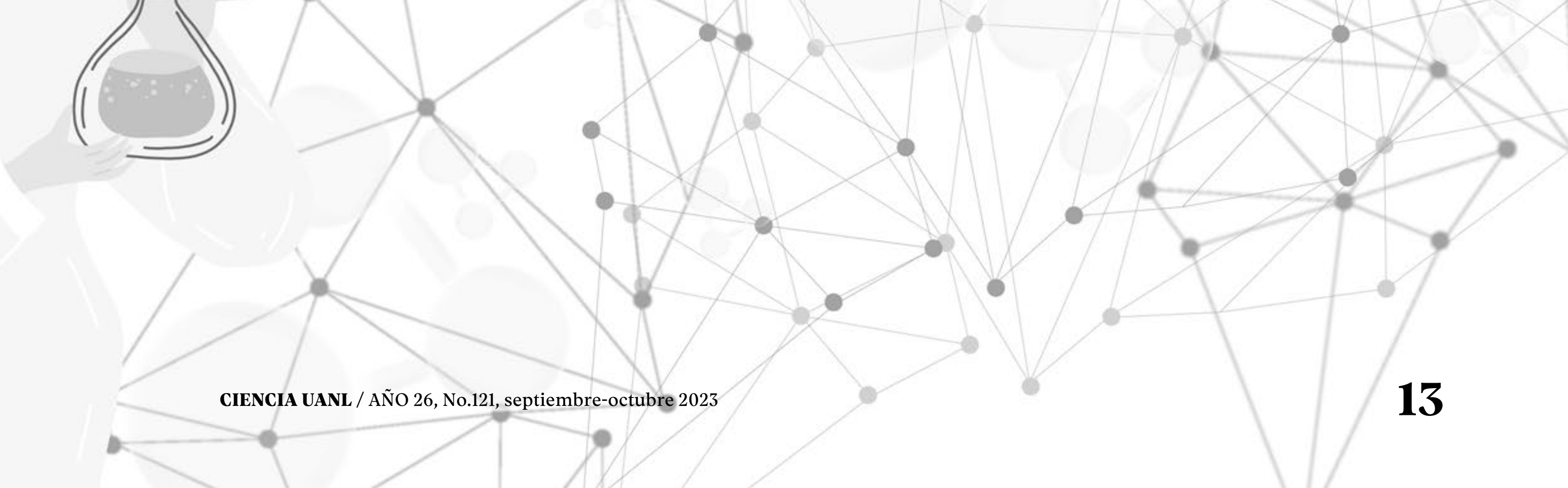
dina-Franco, 2017). A nivel industrial, la Química Computacional se usa para caracterizar y diseñar nuevos materiales, así como en el desarrollo de rutas de síntesis química más eficientes a través de herramientas de inteligencia artificial (IA) (Lu, Deng y Shuai, 2021).

¿QUÉ PAPEL JUEGAN LAS MUJERES EN LA QUÍMICA COMPUTACIONAL?

Históricamente, las mujeres han tenido una presencia muy activa en sectores como la informática y las telecomunicaciones. En la década de los sesenta, las mujeres constituían la mayor parte de la fuerza laboral informática. Sin embargo, esto cambió con la aparición de las computadoras de escritorio, que popularizaron esta actividad como algo exclusivo para hombres. Tan sólo en EUA, el número de mujeres graduadas en informática pasó de 37% en 1984 a 18% en 2017 (White, 2017). Esta disparidad entre hombres y mujeres es aún más grande conforme va aumentando el nivel educativo. A pesar de las barreras de su época, mujeres como las doctoras Margaret Dayhoff, Yvonne Martin y Arianna Wright Rosenbluth, lograron posicionarse dentro del círculo computacional, abriendo camino para la aplicación de la computación en áreas de la ciencia como Biología y Química. La Dra. Dayhoff sentó las bases de la Bioinformática (Gauthier *et al.*, 2019), mientras que la Dra. Martin fue una de las primeras defensoras de la Química Computacional y su uso en el diseño de fármacos (Stouch, 2009). Por su parte, la Dra. Wright Rosenbluth ayudó a crear Metrópolis, uno de los algoritmos más importantes que se usa en modelado molecular.

Afortunadamente, el panorama actual en la Biología y la Química Computacional es mejor que en los sectores exclusivamente computacionales. En los últimos 40 años, el número de afiliaciones y la presencia de mujeres en congresos en esta división y en diseño de fármacos asistido por computadora (DIFAC) ha aumentado a 25% (Holloway y McGaughey, 2018). En particular, en el DIFAC se estima que esta cifra ha aumentado a 38% (hasta 2017), comparado con 13% en 1989 y 1% en 1975. Esto se debe en gran parte a que son divisiones relativamente nuevas, además de la disminución de limitaciones por normas sociales obsoletas respecto a roles de género. También, la naturaleza multidisciplinaria, en donde convergen egresados de carreras como Química, Biología, Bioquímica, Farmacia e incluso algunas ingenierías, ha disminuido esta brecha. Otros factores que contribuyen a un crecimiento del número de mujeres son la mayor flexibilidad que permite el trabajo vía remota y las mentorías positivas. Estas últimas pueden proporcionar perspectivas importantes en el área de trabajo, estilos de vida y valores reflejados.

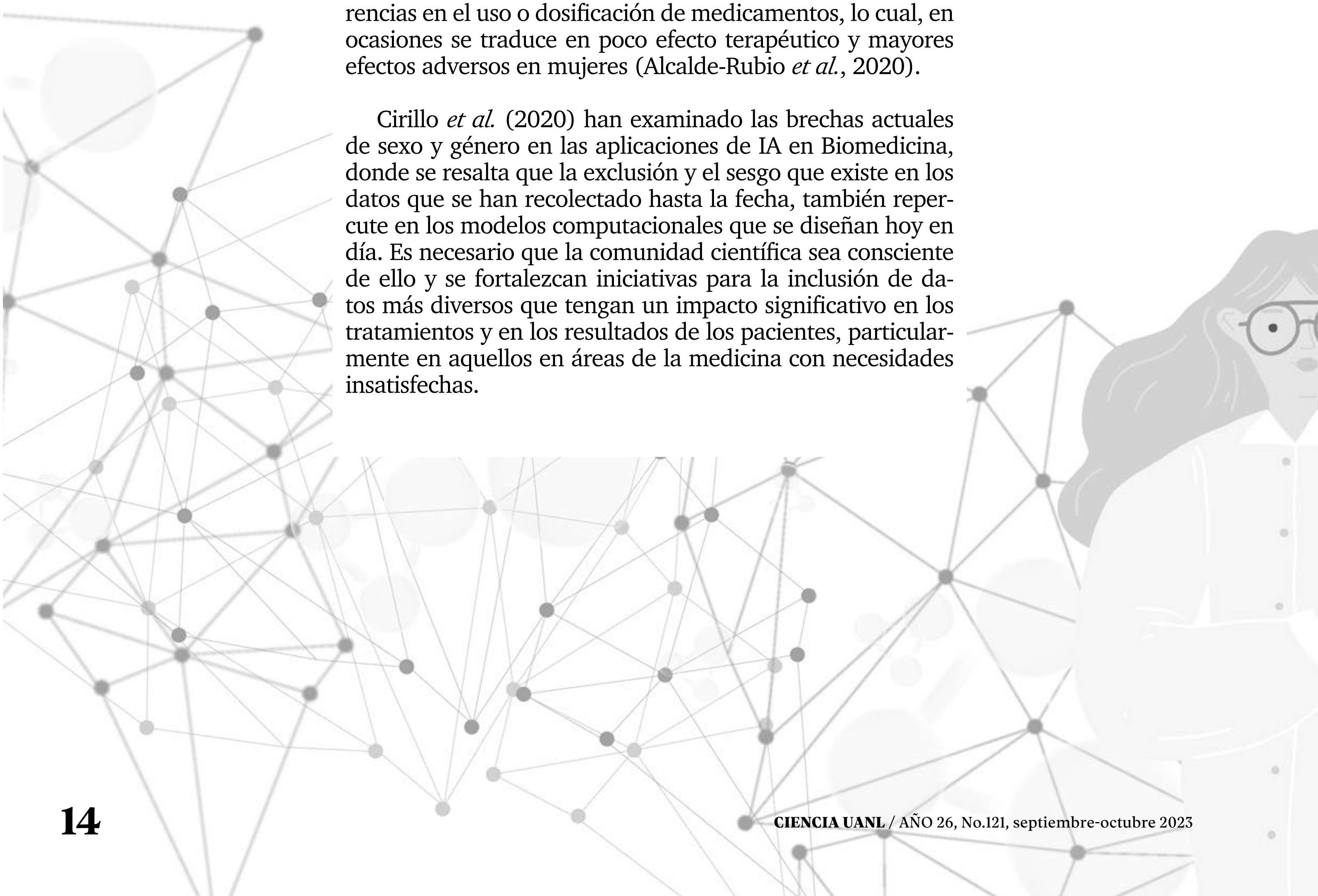
Recientemente, las revistas científicas han tomado iniciativas para reconocer y celebrar el trabajo de mujeres en la Química Computacional y han organizado conversatorios y volúmenes o colecciones especiales (ver en línea: *Women in Computational Chemistry y Women in Artificial Intelligence in the Life Sciences*), que exponen investigaciones, cifras, opiniones y experiencias



personales de mujeres en esa especialidad. Entre los retos que aún quedan por afrontar se recalcan: obstáculos culturales y sociales (mayoritariamente en mujeres de primera generación), sexismo, malas políticas de licencia de maternidad/paternidad, discrepancias salariales y el famoso techo de cristal.

En la actualidad, es importante que las mujeres estemos ejerciendo como profesionistas en nuestra sociedad y fungiendo como pioneras en disciplinas emergentes como el DIFAC y en las aplicaciones que ofrecen tecnologías como la IA. La presencia de mujeres, al menos en estas disciplinas, ha enriquecido la visión que se tiene respecto al manejo y tratamiento de ciertas enfermedades. Por ejemplo, se ha evidenciado la falta de representación de mujeres en estudios clínicos, lo que repercute en el espectro de efectos adversos que se estudian o contemplan al lanzar un nuevo medicamento, o bien, al darle un nuevo uso (Carrasco *et al.*, 2022). A pesar de que se ha demostrado que la respuesta a fármacos es distinta en hombres y mujeres, no se han establecido diferencias en el uso o dosificación de medicamentos, lo cual, en ocasiones se traduce en poco efecto terapéutico y mayores efectos adversos en mujeres (Alcalde-Rubio *et al.,* 2020).

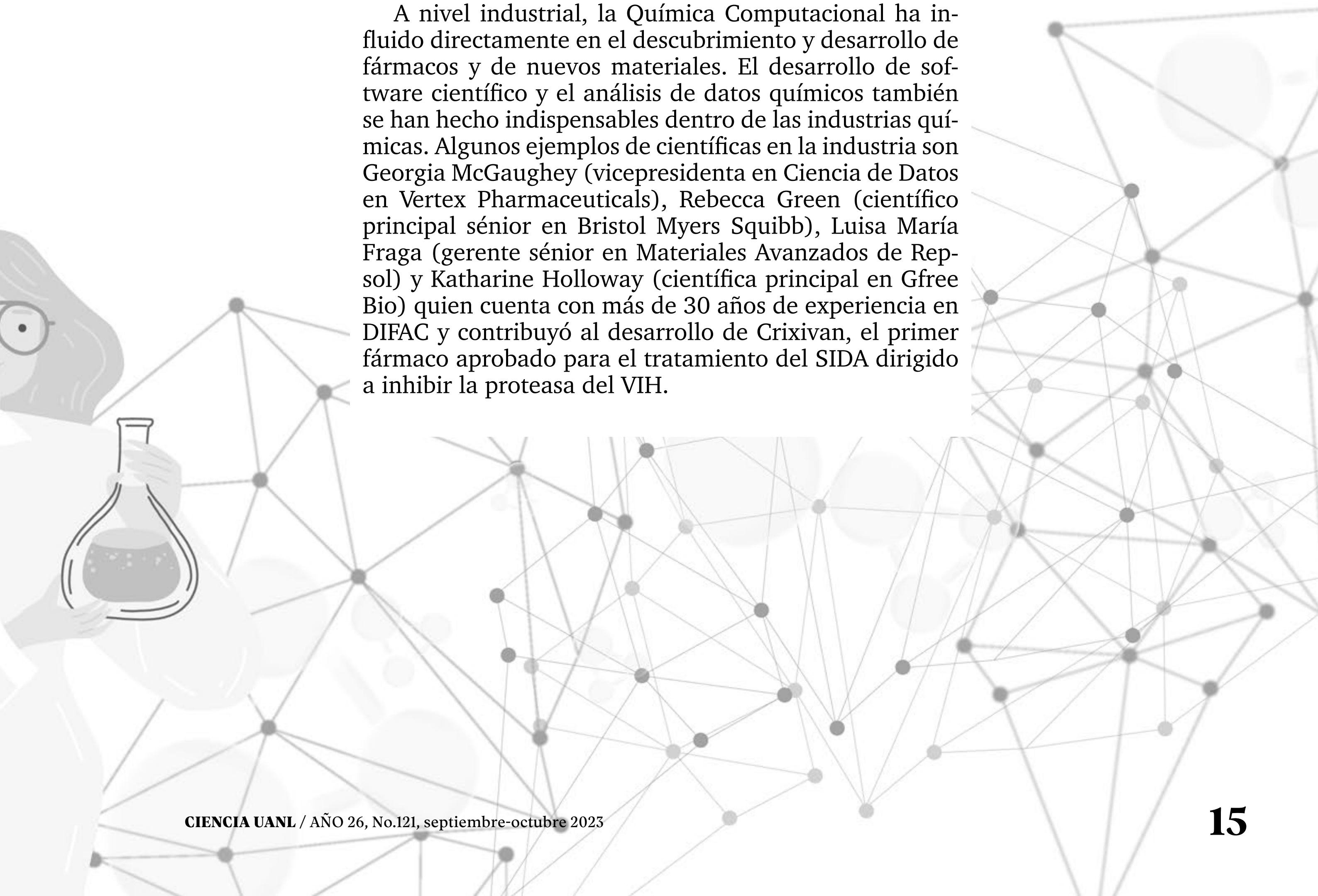
Cirillo *et al.* (2020) han examinado las brechas actuales de sexo y género en las aplicaciones de IA en Biomedicina, donde se resalta que la exclusión y el sesgo que existe en los datos que se han recolectado hasta la fecha, también repercute en los modelos computacionales que se diseñan hoy en día. Es necesario que la comunidad científica sea consciente de ello y se fortalezcan iniciativas para la inclusión de datos más diversos que tengan un impacto significativo en los tratamientos y en los resultados de los pacientes, particularmente en aquellos en áreas de la medicina con necesidades insatisfechas.



HALLAR TU CAMINO A TRAVÉS DE UN ESPACIO DE POSIBILIDADES

Existe una idea falsa de que una vez que decides estudiar una carrera no hay vuelta atrás y harás eso toda tu vida, o bien, que si decides realizar estudios de posgrado tu único camino será la academia. Actualmente, el desarrollo laboral para una mujer que enfoca su carrera profesional en la Química Computacional es muy amplio, aunque probablemente limitado en términos geográficos. Para mostrar que las trayectorias profesionales no siempre son y ni tienen que ser lineales, aquí mostramos ejemplos de mujeres con esta formación desarrollándose profesionalmente en diferentes campos, desde la industria y la academia hasta en emprendimientos y política. Para tener un panorama más amplio de las investigaciones de estas mujeres puedes consultar el siguiente directorio: http://iopenshell.usc.edu/wtc/resources.html

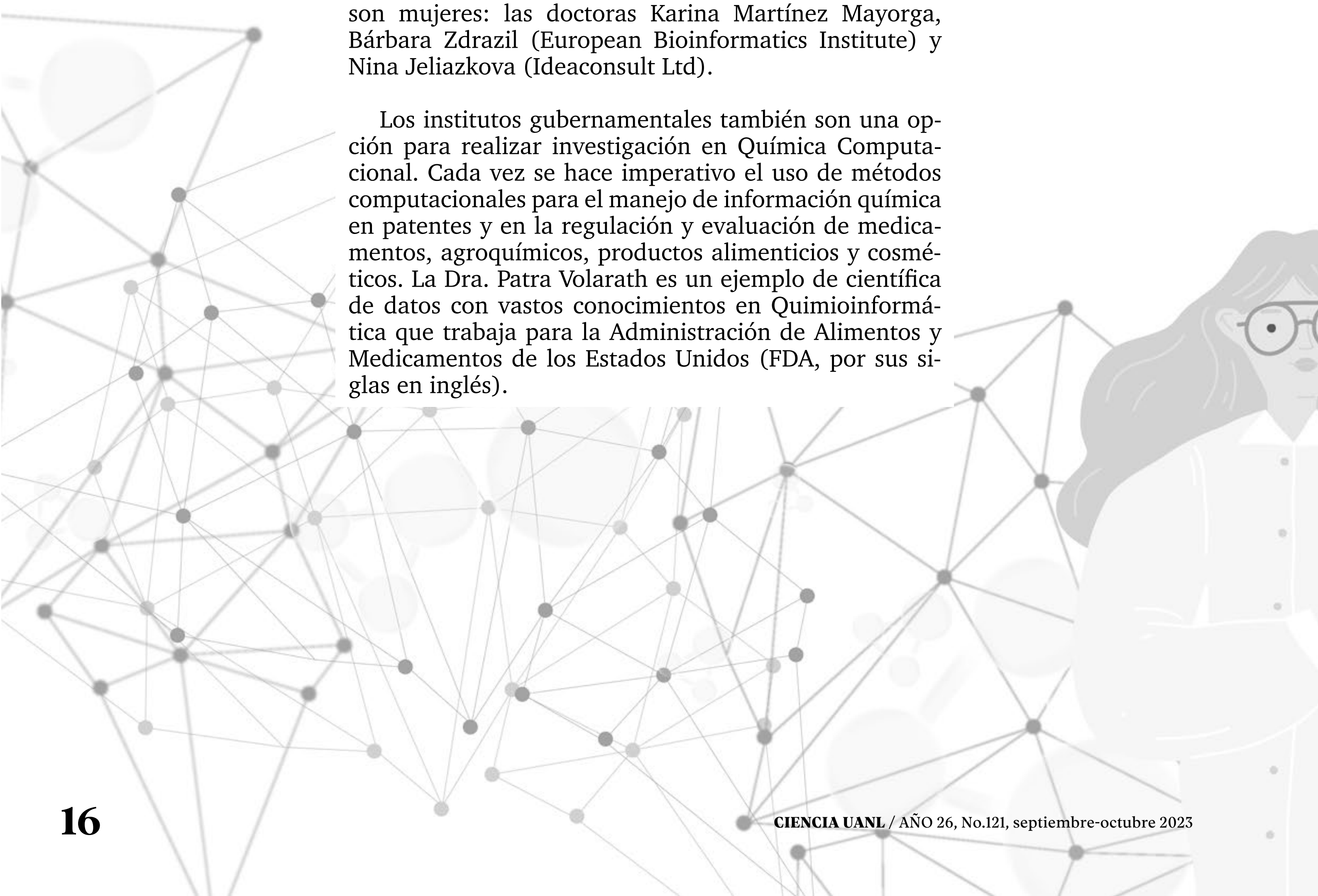
A nivel industrial, la Química Computacional ha influido directamente en el descubrimiento y desarrollo de fármacos y de nuevos materiales. El desarrollo de software científico y el análisis de datos químicos también se han hecho indispensables dentro de las industrias químicas. Algunos ejemplos de científicas en la industria son Georgia McGaughey (vicepresidenta en Ciencia de Datos en Vertex Pharmaceuticals), Rebecca Green (científico principal sénior en Bristol Myers Squibb), Luisa María Fraga (gerente sénior en Materiales Avanzados de Repsol) y Katharine Holloway (científica principal en Gfree Bio) quien cuenta con más de 30 años de experiencia en DIFAC y contribuyó al desarrollo de Crixivan, el primer fármaco aprobado para el tratamiento del SIDA dirigido a inhibir la proteasa del VIH.



En la academia, específicamente en el campo de la dinámica molecular, resaltan los trabajos de las doctoras Teresa Head-Gordon, Rommie E. Amaro y Zoe Cournia, esta última también desarrolló SME Ingredio, una *app* para teléfonos móviles que informa a los consumidores sobre los peligros potenciales de los ingredientes químicos en los productos alimenticios y cosméticos. En América Latina, científicas que destacan son las doctoras Carolina Horta Andrade (Universidade Federal de Goiás, Brasil), Fernanda Duarte (Chile, actualmente investigadora en la Universidad de Oxford, Reino Unido), Karina Martínez Mayorga (Instituto de Química, UNAM, México) y Laura Domínguez Dueñas (Facultad de Química, UNAM, México).

Además de la investigación, es común que científicas se desempeñen como revisoras o editoras de revistas científicas, un ejemplo sobresaliente es el del *Journal of Chemoinformatics*, en donde tres de los cuatro editores son mujeres: las doctoras Karina Martínez Mayorga, Bárbara Zdrazil (European Bioinformatics Institute) y Nina Jeliazkova (Ideaconsult Ltd).

Los institutos gubernamentales también son una opción para realizar investigación en Química Computacional. Cada vez se hace imperativo el uso de métodos computacionales para el manejo de información química en patentes y en la regulación y evaluación de medicamentos, agroquímicos, productos alimenticios y cosméticos. La Dra. Patra Volarath es un ejemplo de científica de datos con vastos conocimientos en Quimioinformática que trabaja para la Administración de Alimentos y Medicamentos de los Estados Unidos (FDA, por sus siglas en inglés).



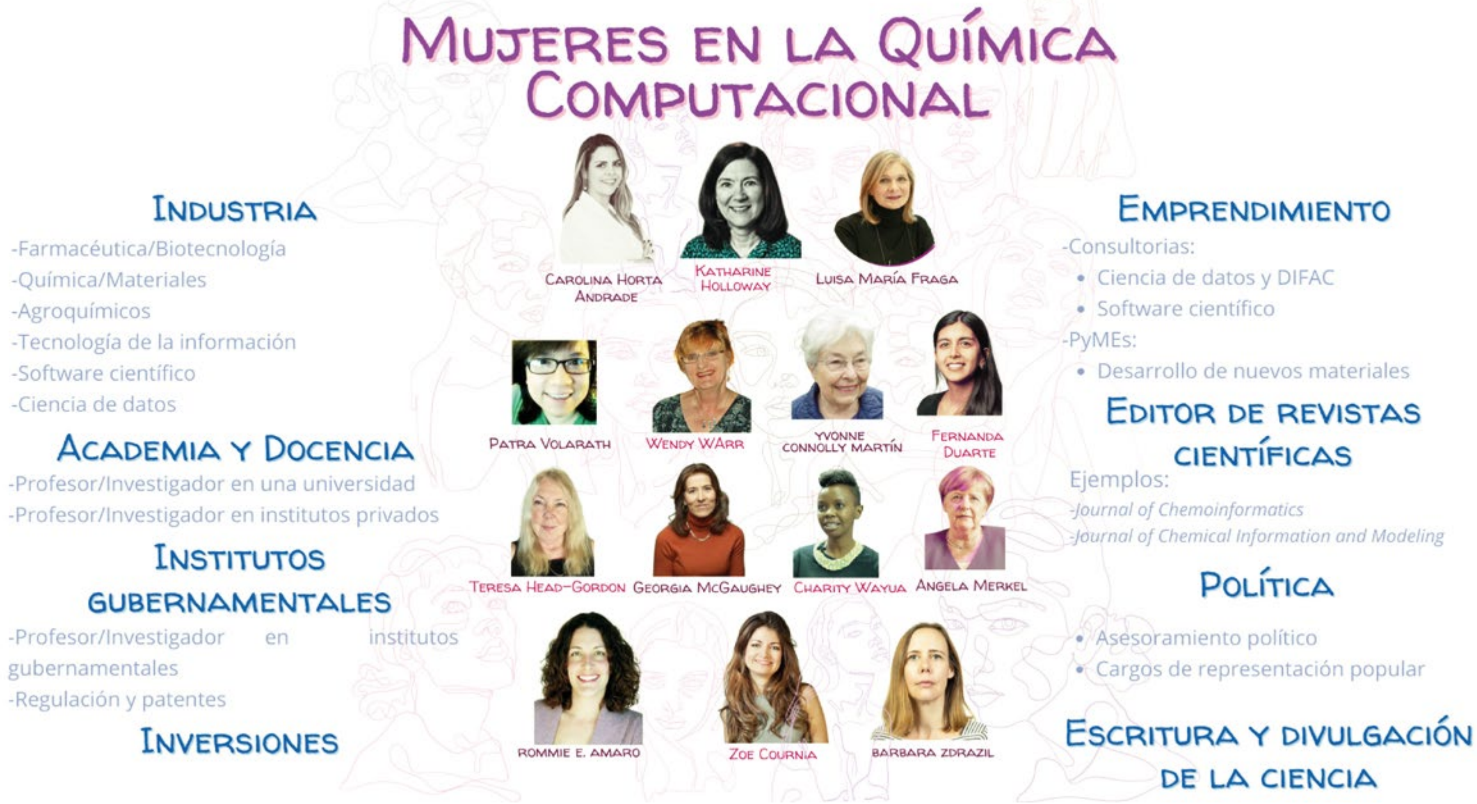

Mujeres en la Química Computacional
Industria
-Farmacéutica/Biotecnología
-Química/Materiales
-Agroquímicos
-Tecnología de la información
-Software científico
-Ciencia de datos
Academia y Docencia
-Profesor/Investigador en una universidad
-Profesor/Investigador en institutos privados
Institutos gubernamentales
-Profesor/Investigador en institutos gubernamentales
-Regulación y patentes
Inversiones
Carolina Horta Andrade
Katharine Holloway
Luisa María Fraga
Patra Volarath
Wendy Warr
Yvonne Connolly Martín
Fernanda Duarte
Teresa Head-Gordon
Georgia McGaughey
Charity Wayua
Angela Merkel
Rommie E. Amaro
Zoe Cournia
Barbara Zdrazil
Emprendimiento
-Consultorias:
• Ciencia de datos y DIFAC
• Software científico
-PyMEs:
• Desarrollo de nuevos materiales
Editor de revistas científicas
Ejemplos:
-Journal of Chemoinformatics
-Journal of Chemical Information and Modeling
Política
• Asesoramiento político
• Cargos de representación popular
Escritura y divulgación de la ciencia

Figura 2. Espacios para el desarrollo laboral en Química Computacional (adaptado de Dong, 2020).

Por otra parte, las consultorías y las pequeñas y medianas empresas (Pymes) son ejemplos de emprendimientos. Destacamos Wendy Warr & Associates, empresa fundada en 1992 por la Dra. Wendy Warr, quien tiene cerca de 50 años de experiencia en Quimioinformática y Química Computacional y casi 20 de éstos en la industria farmacéutica. Otro ejemplo es pinely.co, empresa fundada por Rachelle Choueiri y Shabnam Safaei, que utiliza la computación cuántica y las simulaciones clásicas para optimizar sistemas químicos y desarrollar materiales y catalizadores más sostenibles.

En la política y en las inversiones también hay casos de mujeres muy sobresalientes. Angela Merckel, excanciller alemana, obtuvo su doctorado en Química Cuántica, pero se interesó en la política con la caída del muro de Berlín. La Dra. Charity Wayua (directora de Investigación sénior de IBM) puso en práctica sus habilidades como investigadora del cáncer en un paciente especial: el gobierno de su Kenia natal. Ella contribuyó a mejorar drásticamente el proceso para abrir nuevos negocios en su país, favoreciendo su crecimiento económico, nuevas inversiones y el reconocimiento del Banco Mundial.

CONCLUSIONES

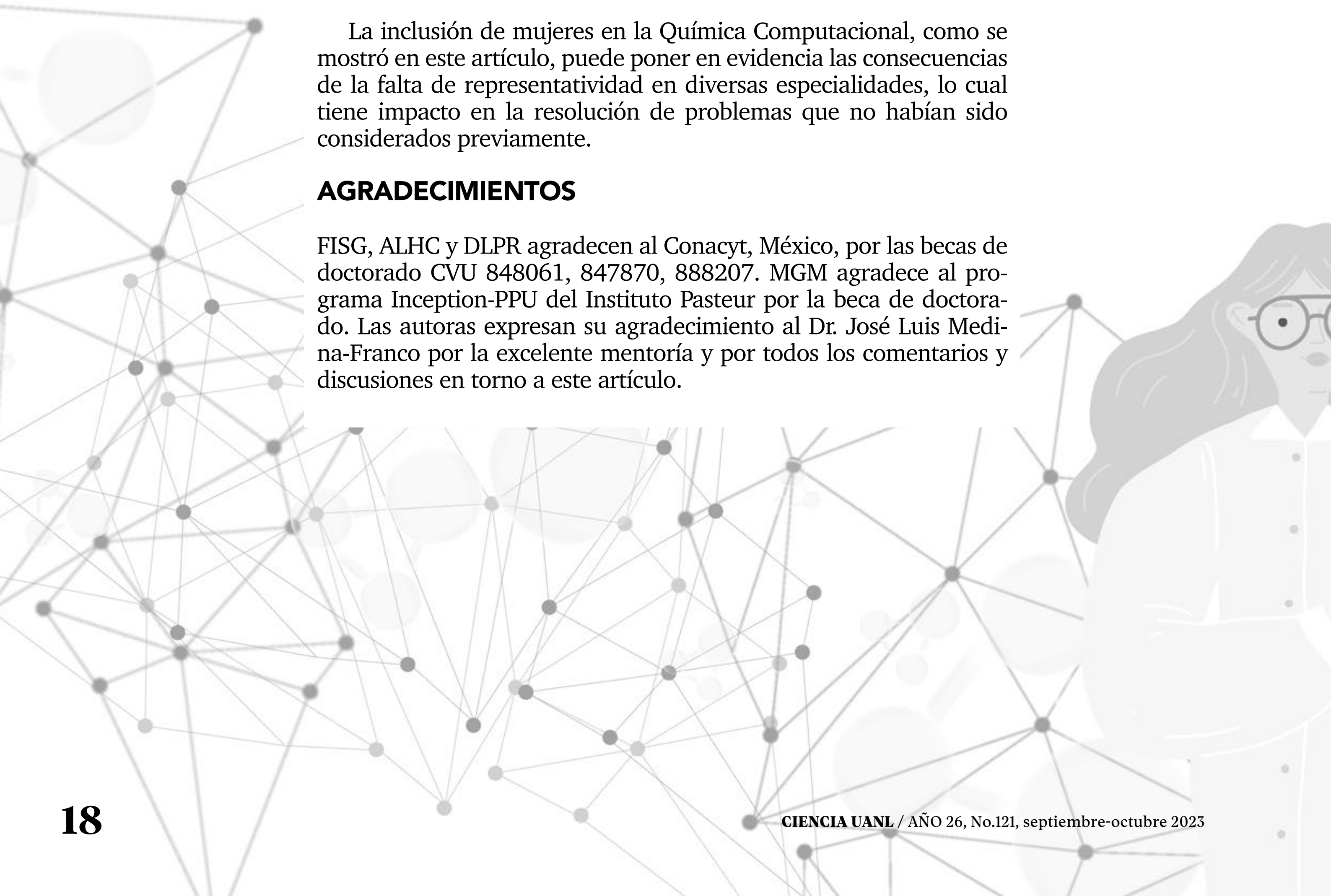
Las mujeres podemos y debemos continuar aventurándonos en las áreas emergentes del STEM. Muy importante: si aspiramos a enriquecer la visión que se tiene respecto a la ciencia y sus aplicaciones en la sociedad, se deben realizar esfuerzos que incentiven la diversidad de género en sectores STEM. Como ciudadanos podemos contribuir en esta tarea externando nuestras opiniones y "alzando la voz" ante situaciones no equitativas o injustas. Mujeres científicas podemos compartir experiencias e impulsar la participación de más mujeres hacia una igualdad de género. Colegas hombres también pueden ayudar informándose y visibilizando el trabajo de sus colegas mujeres (más allá de sólo cubrir cuotas para congresos o atraer recursos a sus laboratorios).

Otra forma es evitando el *mansplaining*, respetando y no ocupando lugares como ponentes en foros y espacios destinados a compartir experiencias de mujeres científicas. Finalmente, las instituciones educativas, gubernamentales y las empresas deben asumir compromisos que aseguren espacios para las mujeres, e implementar políticas de equidad de género que garanticen que las mujeres puedan tener visibilidad, éxito y reconocimiento en todos los niveles.

La inclusión de mujeres en la Química Computacional, como se mostró en este artículo, puede poner en evidencia las consecuencias de la falta de representatividad en diversas especialidades, lo cual tiene impacto en la resolución de problemas que no habían sido considerados previamente.

AGRADECIMIENTOS

FISG, ALHC y DLPR agradecen al Conacyt, México, por las becas de doctorado CVU 848061, 847870, 888207. MGM agradece al programa Inception-PPU del Instituto Pasteur por la beca de doctorado. Las autoras expresan su agradecimiento al Dr. José Luis Medina-Franco por la excelente mentoría y por todos los comentarios y discusiones en torno a este artículo.



REFERENCIAS

Alcalde-Rubio, L. *et al.* (2020). Gender disparities in clinical practice: are there any solutions? Scoping review of interventions to overcome or reduce gender bias in clinical practice. *International Journal for Equity in Health.* 19(1): 166.

Carrasco, BO., *et al.* (2022). Drug repositioning with gender perspective focused on Adverse Drug Reactions. *bioRxiv.* Disponible en: https://doi.org/10.1101/2022.07.22.501091

Cirillo, D., Catuara-Solarz, S., Morey, C., *et al.* (2020). Sex and gender differences and biases in artificial intelligence for biomedicine and healthcare. *npj Digital Medicine.* 3(81):1-10.

Cuevas, G. (2005). Química computacional. *Revista Ciencias Matemáticas.* 56(2):33-42.

Damm-Ganamet, K.L., et al. (2020). Breaking the glass ceiling in simulation and modeling: Women in pharmaceutical discovery. Journal of Medicinal Chemistry. 63(5):1929-1936.

Dong, S. (2020). *For students wondering what they can do with a PhD in computational/theoretical chemistry (or in STEM fields in general), I made an (incomplete) list of career paths based on real-life examples. Please forgive me for mixing "roles" and "industry" here* [Twitter] 25 octubre. Disponible en: https://twitter.com/sijia_dong/status/1320445078649868288.

Emambokus, N., *et al.* (2016). Women in Science. *Cell metabolism.* 23(5):747-748.

Gauthier, J., *et al.* (2019). A brief history of bioinformatics. *Briefings in bioinformatics.* 20(6):1981-1996.

Holloway, M.K., y McGaughey, G.B. (2018). Computational Chemistry: A Rising Tide of Women. *Journal of chemical information and modeling.* 58(5):911-915.

Lu, Y., Deng, G., y Shuai, Z. (2021). Future directions of chemical theory and computation. *Journal of Macromolecular Science*, Part A: Pure and Applied Chemistry. 93(12):1423-1433.

Saldívar-González, F., Prieto-Martínez, F.D., y Medina-Franco, J.L. (2017). Descubrimiento y desarrollo de fármacos: un enfoque computacional. *Educación Química.* 28(1):51-58.

Stouch, T.R. (2009). A well deserved honor: Yvonne Martin, 2009 recipient of the Herman Skolnik Award. *Journal of Computer-Aided Molecular Design.* 23(12):829-830.

White, G.B. (2017). Melinda Gates: The Tech Industry Needs to Fix Its Gender Problem-Now. *The Atlantic.* (16 March). Disponible en: https://www.theatlantic.com/business/archive/2017/03/melinda-gates-tech/519762/

Descarga aquí nuestra versión digital.

Bajorath *et al. Journal of Cheminformatics* (2022) 14:82
https://doi.org/10.1186/s13321-022-00661-0

Journal of Cheminformatics

MEETING REPORT

Open Access

Chemoinformatics and artificial intelligence colloquium: progress and challenges in developing bioactive compounds

Jürgen Bajorath[1], Ana L. Chávez-Hernández[2], Miquel Duran-Frigola[3,4], Eli Fernández-de Gortari[5], Johann Gasteiger[6], Edgar López-López[2,7], Gerald M. Maggiora[8], José L. Medina-Franco[2*], Oscar Méndez-Lucio[9], Jordi Mestres[10,11], Ramón Alain Miranda-Quintana[12], Tudor I. Oprea[13,14,15,16], Fabien Plisson[17], Fernando D. Prieto-Martínez[18], Raquel Rodríguez-Pérez[19], Paola Rondón-Villarreal[20], Fernanda I. Saldívar-Gonzalez[2], Norberto Sánchez-Cruz[10,21] and Marilia Valli[22]

Abstract

We report the main conclusions of the first Chemoinformatics and Artificial Intelligence Colloquium, Mexico City, June 15–17, 2022. Fifteen lectures were presented during a virtual public event with speakers from industry, academia, and non-for-profit organizations. Twelve hundred and ninety students and academics from more than 60 countries. During the meeting, applications, challenges, and opportunities in drug discovery, de novo drug design, ADME-Tox (absorption, distribution, metabolism, excretion and toxicity) property predictions, organic chemistry, peptides, and antibiotic resistance were discussed. The program along with the recordings of all sessions are freely available at https://www.difacquim.com/english/events/2022-colloquium/.

Keywords: ADME profile, Antibiotic resistance, Artificial intelligence, Career development, Drug discovery, Machine learning, Ligand-based drug design, Natural products, Peptides, Structure-based drug design, Virtual screening

Introduction

In the setting of a growing number of applications and developments of computational approaches to drug discovery and related fields, of an increasing frequency of virtual meetings [1, 2], and of efforts to enhance the education of students [3, 4], the first Chemoinformatics and Artificial Intelligence (AI) Colloquium organized by a Latin American country was held in Mexico City, June 15–17, 2022. The virtual meeting featured talks by 15 international experts. Table 1 presents the full program. The speakers, eight of which were from Latin American Countries or of Latin American origin, have a broad perspective as they work in academia, large pharmaceutical companies, new start-ups, public research institutions and non- profit organizations.

Twelve hundred and ninety participants, from more than 67 countries, including México, India, Colombia, Brazil, Perú, United States, Cameroon, Ecuador, Argentina, and Germany, had access to the talks through Zoom, YouTube, and the Facebook channels of the School of Chemistry at the Universidad Nacional Autónoma de México (UNAM). The group of participants was made up 659 students (51.1%), 242 academics (18.8%), 236 researchers (18.3%), 119 industry professionals (9.2%), and 34 with other non-disclosed profiles (2.6%) from more than 40 institutions in Mexico and other countries.

The meeting was hosted by the Department of Pharmacy in UNAM's School of Chemistry. Recordings of all

*Correspondence: medinajl@unam.mx

[2] DIFACQUIM Research Group, Department of Pharmacy, School of Chemistry, National Autonomous University of Mexico, 04510 Mexico City, Mexico
Full list of author information is available at the end of the article

Table 1 Program of the chemoinformatics and artificial intelligence colloquium and related links

Speaker[a]	Affiliation (country)	Lecture[b]	Related links and references
Johann Gasteiger	University of Erlangen- Nuremberg (Germany)	Chemistry in times of artificial intelligence	[4–7]
Marilia Valli	University of São Paulo (Brazil)	Brazilian biodiversity chemical space into NuBBE database	[8]
Fernando Prieto D. Prieto-Martínez	National Autonomous University of México (Mexico)	A bird's eye view of AI in structure-based drug design	[9–11]
Paola Rondón-Villarreal	Industrial University of Santander. Currently Universidad de Santander (Colombia)	Machine learning in virtual screening and peptide's design	[12]
Fabien Plisson	Center for Research and Advanced Studies of the National Polytechnic Institute (CINVESTAV-IPN) (Mexico)	Probing the limits in AI-driven peptide design	[13]
Miquel Duran-Frigola	Ersilia Open Source Initiative (UK)	Ersilia, a hub of AI/ML models for infectious and neglected tropical diseases	[14, 15]
Eli Fernández-de Gortari	International Iberian Nanotechnology Laboratory (INL) (Portugal)	The role of generated chemical space in ML-based virtual screening	[16–18]
Norberto Sánchez-Cruz	Chemotargets, LLC (Spain); National Autonomous University of México (Mexico)	Deep graph learning for protein-fragment binding predictions	[19]
Raquel Rodríguez-Pérez	Novartis (Switzerland)	Machine learning for the prediction of ADME properties in pharmaceutical industry	[20, 21]
Jordi Mestres	Chemotargets, LLC (Spain)	Challenges and benefits of integrating the preclinical-to-postmarketing safety data continuum	[19]
Gerald M. Maggiora	University of Arizona (USA)	Development of a soft rule of five	[22]
Ramón A. Miranda-Quintana	University of Florida (USA)	Extended similarity analysis: from pair of molecules, to chemical space and beyond	[23, 24]
Jürgen Bajorath	University of Bonn (Germany)	DeepSARM: From structural and SAR analysis to compound design and optimization	[25, 26]
Oscar Méndez-Lucio	Recursion Pharmaceuticals (USA)	Geometric deep learning for structure-based drug design	[27]
Tudor I. Oprea	Roivant Sciences (USA)	Learning from machine learning: some lessons from a gene-centric Alzheimer's model	[28, 29]

[a] In order of presentation

[b] Each lecture is associated with the references given in the far-right column and vice-versa

talks and the full program are freely available at https://www.difacquim.com/english/events/2022-colloquium/. The following sections summarize the key developments presented and discussed during the meeting. The content is organized into six sections: following the introduction, the effectiveness and challenges of chemoinformatics and AI methods are considered, followed by a discussion of the opportunities afforded by these methods, general insights, and an overview of the material. The report ends with a discussion of the overall conclusions.

Challenges of chemoinformatics and AI methods

Professor Johan Gasteiger, the first speaker in the Colloquium, stated three of the fundamental questions in chemistry: (1) what structure do I need for a certain property?, (2) how do I make this structure?, and (3) how do I synthesize this and characterize this compound? Answers to the first question involve structure-property or structure-activity relationships, to the second question involve synthesis design, and to the third question involve reaction prediction and structure elucidation. In many instances, answers to these questions can be found in the vast amount of data stored in publicly accessible databases, which contain information on millions of compounds, their structures and reactions, as well as many of their chemical and biological properties. Because of the size and complexity of this data, chemoinformatics tools are essential if one is to utilize this information effectively in order to answer important chemical questions (*vide supra*) [4].

Inductive learning, i.e. learning from examples, is an important mode of learning in chemistry, which typically arises in the interpretation and analysis of data. The objective of most artificial intelligence (AI) methods is to emulate human reasoning by machine or automated processes. Thus, inductive learning methods such as machine learning (ML) and deep learning (DL), have many applications in chemistry. In fact, the application of AI, specifically artificial neural networks (ANN), in chemistry and drug design has a long history [7]. Recent developments in AI methods have led to a resurgence and increased interest in this field. Sufficient knowledge and correct application (beyond the hype) are necessary, particularly for students, early career researchers, and investigators interacting with computational chemists or data scientists [30]. It is clear that AI has applications in many areas of chemistry such as property prediction, reaction prediction, synthesis planning, structure elucidation, drug design, food chemistry, agrochemistry, risk assessment of chemicals, development of cosmetic products, material science, and process control [6, 31]. Because of the wide spectrum of applications of chemoinformatics and AI in chemistry, the colloquium was centered on three major areas: identifying and developing small molecules as drug candidates, peptides, and natural products [32]. The following subsections summarize the challenges that were discussed during the meeting.

Data issues

Data is a cornerstone for the generation of information and knowledge. Hence, data quantity and quality are vital to the development and performance of chemoinformatics and AI methods. Thus, academia, start-ups, and industry should, as a scientific community, prioritize access to data, which is as balanced and complete as possible. For example, activity data associated with ligand-target interactions should also include data associated with inactive ligands in order to capture weak or non-existent interaction data. In that way researchers will be able to access the full spectrum of available knowledge [33]. Moreover, such a "holistic" viewpoint would help cope with the data imbalance present in many drug design and compound optimization campaigns.

Data curation and the construction of reliable databases are major issues that also need to be addressed. Poorly curated databases complicate the assessment of the predictive performance of AI models. Combining efforts could, however, facilitate access to new and interesting data. Examples include natural products, metallodrugs, safety, preclinical, and toxicological databases, which complement the current data available in the public domain and offer new perspectives on the known data [34–36]. There are, however, potential conflicts of interest related to the publication of sensitive data associated with intellectual property. For example, post-marketing (pharmacovigilance) data that might be biased related to the time and clarity of data shared.

Technical challenges

One of the most important issues in chemoinformatics is how to compare molecules. There are two equally important aspects to this issue: (1) how to represent the information in a molecular structure in a computationally appropriate form and (2) how to determine the structural relationship of one molecule to another using this information. In the first instance, a common approach in widespread use today is the development of 'vectorized' representations of molecular structure such as that exemplified by Extended Connectivity Fingerprints (ECFP) [37] or MACCS key fingerprints [38], that represent the structural features of molecules as binary vectors whose components are based on the presence or absence of specific substructural features. In addition, SMILES sequences and molecular graphs are being used as features for the most recent neural networks architectures. Many of these and closely related methods provide a basis for developing all manner of AI models. An important caveat regarding these approaches is that they deal almost exclusively with 2D molecular structures. Three-dimensional structural features, such as multiple conformations, are rarely treated for a variety of reasons.

Once the structural information has been appropriately represented, the issue now becomes how to compare molecular structures. Traditionally, this has been done based on assessments of the *structural similarity* [39] of pairs of molecules, using any one of a number of similarity measures (*aka* similarity functions or coefficients), the most popular being that developed by Jaccard and Tanimoto [40, 41]. Recently, Miranda et al. have developed a new, highly efficient method, which facilitates comparison of multiple molecules simultaneously [22, 23], opening up new possibilities in drug research.

Unfortunately, molecular similarities are representation dependent. Thus, different structural representations will typically lead to different similarity values, even if the same similarity function is used. Although this appears to be a severe limitation of structural similarity methods, in many instances they appear to produce reasonable results in similarity-based database searches, which lie at the heart of LBDD methods [42], which are described in greater detail in "Ligand and structure-based drug design methods" and "Ligand-based drug-design opportunities".

Molecular similarity provides a suitable basis for constructing *chemical spaces*, which play an important role in LBDD. Chemical spaces are composed of a set of molecules and the set of pairwise similarities relating them to each other. Thus, they are dependent upon the molecular representation and similarity measure used in their construction, and they are, of course, also subject to the lack of invariance of all structural similarity measures.

Chemical spaces are typically represented in two ways, coordinate-based and network-based. Coordinate-based chemical spaces are generally of high-dimension, and thus are subject to the 'Curse of Dimensionality' [43, 44].

Lower-dimensional subspaces, in many instances, are employed for the purpose of visualization, however, with a concomitant loss of information.

Chemical space networks (CSN) provide an alternative representation that is not afflicted by the Curse of Dimensionality [45]. This combined with the availability of efficient algorithmic methods for characterizing the properties of very large networks, such as the Internet, make CSNs the preferred means for representing very large chemical spaces. Although it is difficult to perceive relationships visually in very large chemical spaces represented by CSNs, the important point here is that the structure of network data facilitates its analysis.

Chemical spaces lie at the heart of LBDD (see "Ligand and structure-based drug design methods" and "Ligand-based drug-design opportunities" for a fuller discussion), but because of their representation dependence they are not unique. However, as noted earlier, this may not in many instances materially affect the effectiveness of ligand-based searches of chemical spaces [42]. Maggiora has provided a relatively comprehensive discussion of molecular representations, similarity measures, and chemical spaces, which should be consulted for more details [46].

Chemical Checker [15] signatures were proposed in order to facilitate the conversion of bioactivity data to a format readily amenable to ML methods. The concept of chemical space is continuing to evolve. Its application has been extended to data visualization and to the study of structure-property relationships, lead optimization, data fusion, and data-driven decision making, to name a few applications. However, many different types of descriptors are available to represent different classes of compounds, e.g., natural products, peptides, metallodrugs, drug-like, and lead compounds. The extensive list of possible molecular representations raises a significant question, viz. "what are the most suitable descriptors for my dataset?" In specific cases, the answer combines different kinds of features or types of data such as chemical or topological features, and physical or biological data. However, it is not easy to collect, order, and organize such heterogeneous information. In order to enter an era where chemical and biological spaces are integrated, the development of new methodologies is required for assessing chemical and biological similarity and for handling genes, proteins, omics data, and chemical data in a consistent manner [47].

Another challenge is the implementation of filters to select molecules according to pre-defined rules such as Lipinski's Rule of Five (Ro5). Maggiora discussed the importance of 'soft' methods for selecting compounds according to Ro5. Zadeh et al. define soft methods as an emerging computational approach that parallels the remarkable ability of the human mind to reason and learn in an environment of uncertainty and imprecision. Such methods tend to produce more realistic molecular property relationships as discussed by Maggiora and coworkers [22].

Ligand and structure-based drug design methods

LBDD methods focus entirely on the structure of the ligand. By contrast, SBDD methods focus on the structure of both the ligand and the binding site in its target proteins and/or nucleic acids. Thus, obtaining data in the latter case is typically more difficult.

Because of the greater availability of data on ligand structure, AI methods are more effective, enabling the study of very large volumes of diverse data in LBDD studies. SBDD approaches, on the other hand, have not yet fully explored the utility of AI, although a significant amount of research is currently in progress. One reason for this is the availability of structural data needed in SBDD studies, which require data on the ligand and on its binding site. By comparison, structure, activity, and physicochemical data typically required in LBDD studies, is considerably more available. Because of the limitations of current computational methods, generation of fully reliable 3D conformational states or binding modes is not possible in all cases, although significant strides have been made in computational docking methods, some of which are now capable of docking more than a billion compounds to a given binding site [48, 49]. In addition, recent progress in AI-driven de novo protein structure prediction (see below) has provided an unprecedented wealth of putatively reliable structural templates, with coverage recently approaching the entire protein universe [10, 50, 51].

General challenges

A current limitation of computational approaches in academic settings is related to the relatively limited amount of computational processing capacity. However, over the next few years accessibility to cost effective, highly efficient hardware could increase dramatically, reducing budgetary and time requirements for developing and evaluating new ML algorithms. Other essential challenges discussed during the meeting included the application of chemoinformatics and AI methods to better understand unexplored, rare, and neglected diseases. More consistent communication and collaboration between academia, start-ups, and large industries is also desirable in order to foster a viable synergy and help the transfer of in silico knowledge ultimately to the clinic.

Opportunities for chemoinformatics and AI methods

Ligand-based drug-design opportunities

In addition to in vitro and in vivo methods, in silico methods can enhance serendipity and help to rationalize phenomena that experimental methods alone cannot explain. For example, serendipity in drug design can lead to unexpected but potentially positive results, as exemplified by the discovery of Lyrica (pregabalin) [52]. An excellent opportunity for ligand-based methods to enhance compound comparisons is through the addition or augmentation [15] of chemical and physicochemical property data, of in vitro, in vivo, and 'omics' biological data, and of preclinical, clinical, and post-marketing pharmacovigilance data. The added information would support the development of a comprehensive similarity searching capability that would likely, in specific instances, be able to identify chemical mimetics capable of reverting disease signatures. For example, drug-design procedures might be developed for reversing (or preventing) molecular pathway alterations or for predicting toxicity or safety issues for marketed drugs [53].

Two new applications, Extended Similarity Indices [23, 24] and the structure–activity relationships Matrix (SARM) approach and its deep learning extension (DeepSARM) [25], were presented at the Colloquium by Quintana (Talk 12) and Bajorath (Talk 13), respectively. These applications support multiple procedures such as analog series identification (fragmentation?), analysis of de novo drug-design signatures, similarity searching, and visualization of SAR and chemical spaces.

Structure-based drug-design opportunities

Over the past few decades, SBDD has attained a significant degree of maturity. This is especially true with regard to structure-based virtual screening, which has made remarkable progress despite its intrinsic limitations [54, 55]. In recent years, DL has been used in attempts to further improve the performance of SBDD methods. Perhaps the most well-known example of this is the usage of DL for protein structure prediction. *De novo* structure prediction with Alphafold [10] RoseTTAfold [50], or other programs [51, 56] has yielded many protein models of near-experimental accuracy which has further expanded the opportunities and the applicability domain of homology modeling. Protein models are now increasingly used for prediction of many biophysical properties [57].

Other uses of AI in SBDD include, but are not limited to, potential energy functions that are similar to quantum-chemical descriptions (ANAKIN-ME) [9]. For example, DFT-like interaction potentials at the computational cost of a geometrical optimization with molecular mechanics; force field development [58]; enhanced sampling by means of collective variables [59]; Boltzmann generators trained to identify transition states [60]; protein-ligand interaction fingerprints [61] such as SPLIF [62] or ECIF [63], and scoring functions like GNINA [64]. Recently, the geometric DL approach was used to learn distance distributions and ligand-target interactions and to predict the binding conformation of bioactive compounds. This potential performs as well as or better than well-established scoring functions [27]. Geometry DL uses a mesh on the protein surface [65] as a molecular representation.

New approaches to CADD based on AI methodologies

Chemoinformatics helps transform data into information and subsequently into knowledge in support of decision making. New techniques and methodologies have contributed significantly to encoding and analyzing chemical, biological, and clinical data patterns. For example, different types of neural networks (e.g., neural, deep neural, Kohonen-Self Organizing Maps (SOM), and graph-based) [7] support multitask learning, which facilitates the exploration and exploitation of synergies between prediction tasks in complex systems. This potentially alleviates the need for system reduction or approximation, an attractive approach for holistic drug discovery and design. Furthermore, it is possible to use these new techniques and methodologies for improving graph-based pharmacophoric representations, fragment-based drug design, *de novo* drug design, binding energy predictions, and consensus classification models [18]. However, there are a number of caveats associated with these approaches that must be addressed in order for them to be fully mature.

De novo drug design and generative models

De novo drug design is one of the areas benefiting from DL. For example, DeepSARM is a deep learning extension of SARM for generative fragment-based analog design. DeepSARM [26] introduces chemical novelty into the design process based on recent developments in generative modeling adaptation and the further development of chemical language models. Iterative DeepSARM (iDeepSARM) [25] can rationally modify and extend sequence-to-sequence models and add iterative compound optimization and core-structure modifications.

Deep Graph Learning (DGL) which is based on ANNs, is capable of learning from graph-structured data [66]. It is included as part of the ProSurfScan platform developed by Chemotargets. This platform has been successfully applied to the identification of novel compounds for different targets. It yielded the first AI-designed drug for Huntington's disease, which is currently in clinical trials

[67]. ProSurfScan allows estimation of the compatibility and binding mode of fragments on different regions of a protein surface. Therefore, the protein surface is represented as a complete graph consisting of nodes with pharmacophoric features derived from the analysis of a triangulated mesh representation of the protein surface [68, 69]. Two complementary methods are employed to carry out the predictions. A clique detection algorithm is used to compare the protein surface with known surfaces associated with fragments from ligands present in structures from the Protein Data Bank (PDB) (*aka* fragment environments). This allows placement of the fragment based on the largest subgraph found between the fragment-environment and the protein surface. In addition, a series of DGL models is built using Graph Convolutional Neural Networks (GCNN) that estimate the compatibility of the fragments with respect to distinct regions of the protein surface.

Fernandez-de Gortari discussed the use of generators [16, 18] based on Variational Autoencoders (VAE), a deep neural network architecture. He discussed their advantage for constructing molecules with multi-target profiles and properties of pharmaceutical interest from lead molecule seeds. The methodology is based on using generators obtained from reasonable mutations of fragments [17], obtained by exchanging structurally similar fragments on the lead molecule seed based on a hypothetical continuous SAR for the development of a ML-based virtual screening classifier of Sarco(endo)plasmic reticulum Ca^{2+}-ATPase (SERCA) inhibitors.

Machine learning for the prediction of ADME-Tox properties

Low efficacy associated with bioavailability problems and adverse drug effects have been recognized as one of the main causes of attrition during clinical trials [70]. Thus, the number of possible causes for a compound to fail or to have barely tolerable adverse effects is quite large. Moreover, in vitro and in vivo characterization of a compound's properties can become very costly and time-consuming. For all of these reasons, considerable effort has been made to develop computational models for predicting ADME-Tox properties [70]. AI models have leveraged the information available in heterogeneous ADME-Tox data sets and helped to improve the accuracy of early drug efficacy and safety predictions. There is an increasing number of public and private sector initiatives aimed at the generation and evaluation of prospective models to assist decision-making processes and to generate future innovations for predicting ADME-Tox properties. Initiatives are also underway to permit public use and comparison of ML/DL models to increase confidence in and acceptance of these predictions. For example, Therapeutics Data commons (TDC) was introduced as a platform to systematically access and evaluate ML models across the entire range of therapeutics, accessible via an open python library [71, 72]. TDC encompasses AI-ready datasets and learning tasks for therapeutics; sets of tools to support data processing, model development, validation, and evaluation; and a collection of 'leaderboards' to support model comparison and benchmarking.

Other ML models derive hypothetical properties such as brain penetration (Kp) from limited experimental data or characterize in vivo properties from in vitro assay data. In a study conducted by Rodríguez-Pérez's group, multitask learning based on Graph Neural Networks (MT-GNN) showed superior performance to other ML approaches based solely on in vitro brain penetration data [20]. These promising models have considerable potential for practical applications in other property prediction tasks.

To provide a partial solution to the data issues and improve early drug safety assessment, an effort has been made to integrate preclinical and post-marketing drug safety data with other commonly used sources of information, such as chemical structure data and preclinical assays. Current trends focus on developing novel systems approaches to drug safety that offer a more mechanistic view of predictive safety based on similarity to drug classes, interaction with secondary targets, and interference with biological pathways beyond the traditional identification of chemical fragments associated with selected toxicity criteria [53]. An example of the integration of this information is $\text{CLARITY}_{\text{PV}}$ [73], a web platform for translational safety and pharmacovigilance studies that track side effects throughout all phases of the drug discovery and development process.

Importance of natural products in drug discovery

Natural products have historically contributed to drug discovery as a source of diverse, structurally complex bioactive molecules that have evolved to fulfill specific biological functions. However, drug development from NPs is more complex, costly, and inefficient than drug development from small molecules [74]. Similarly, the small amount of bioactivity data associated with NPs has limited potential applications of ML and DL in the study of naturally occurring compounds. Initiatives such as the NuBBE_{DB}, a virtual database of NPs and their derivatives from the Brazilian biodiversity [75, 76], have paved the way for developing new NP databases and projects like LOTUS [77] for NP storage, search, and analysis. A number of different chemoinformatics [78] and AI [32] applications have been proposed for analyzing the data collected to date. The main applications have focused on understanding the biological activity of NPs, carrying out the systematic search for bioactive NPs with respect to a molecular target of interest, and guiding the chemical synthesis of NP analogs with simplified structures and improved activity. The NuBBE_{DB} database has been expanded in collaboration with CAS (Chemical Abstracts Service). Currently, more than 54,000 substances are described with information on chemical, biological, and pharmacology data that can be explored in order to analyze their medicinal chemistry potential. Recent work on target predictions for compounds in the NuBBE_{DB} led to the identification of chalcones with potential application for the treatment of Chagas disease [79].

General opportunities

Access to AI technology and international networking can also accelerate the development of drugs for neglected diseases, Alzheimer's disease, and antibiotic resistance. The research group of Oprea developed ML models to identify a potential gene relevant to susceptibility to Alzheimer's disease [29]. This analysis also identified potential risk genes including FRRS1, CTRAM, SCGB3A1, FAM92B/CIBAR2, and TMEFF2.

Other chemoinformatics, ML, and DL models were proposed as a means of identifying compounds to combat antibiotic resistance, which is found in all parts of the world [80]. Peptides have been proposed as suitable alternatives since they display biological activity against bacteria, viruses, fungi, and parasites [81, 82]. Antimicrobial peptides (AMP) have a low propensity for bacteria resistance [83, 84]. The research group of Rondón-Villarreal [12] developed an AMP library using the CAMP_{R3} [85] database, and genetic algorithms. The peptide library was designed with specific physicochemical properties (charge, hydrophobicity, isoelectric point, and stability index) and tested against *Escherichia coli*, *Pseudomonas aeruginosa* and methicillin-resistant *Staphylococcus aureus*. This library could potentially lead to the discovery of potent antimicrobial peptides.

However, the challenges of peptide design might require addressing multiple parameters such as high toxicity, poor oral bioavailability, thermal and pH stability, and functional promiscuity in concert. In addition, costs associated with experimental time, human resources, and equipment involved [13], must also be accounted for. Chemoinformatics, ML, and DL approaches should provide a means for developing safe AMPs with reduced toxicity, predict their antibacterial activity and drug-likeness profile, and accelerate antibiotic discovery [86, 87]. Plisson et al. [13] proposed an ML-guided discovery and design project related to non-hemolytic peptides. The workflow is composed of collecting compounds for an AMP database, computing 56 physicochemical descriptors; developing binary-classifier models to predict hemolytic nature and activity; estimating the domain of applicability, and applying optimized models to the discovery of non-hemolytic AMPs from a known database (e.g., APD3) or design novel sequences. The models used in this study include support vector machines, decision trees, random forest, gradient boosting, and k-nearest neighbors. This research is part of a growing series of predictive and generative ML models applied to support the discovery and design of bioactive peptides, including antimicrobial peptides [56, 63]. The authors applied multivariate outlier detection to delineate the boundaries of their predictive models (i.e., applicability domain) leading to the identification of outlying sequences [9]. To date, little work is being carried out on estimating the domain(s) of applicability of peptide modeling, although it is necessary for the parallel application of multiple predictors on a given sequence space.

Recommendations for new generations of scientists

Some speakers shared their experiences as scientists. This section summarizes some general recommendations for future scientists. The early-career scientist should choose topics that open new possibilities and should not adhere to a single approach or technology. "If you have your data, run your own benchmarks tests, build your own models, and try to interpret them in context. Metrics are irrelevant. The only proof is unbiased predictivity".

One should always review the original publications to ensure integrity of information sources and avoid dilution or subjective bias. "Verify what you see, doubt what you find, and always obtain independent confirmation of your observations to validate your work".

Do not be afraid to say, "I do not know." Omniscient human beings are rare. Be ready to learn continuously. Focus on problem-solving skills; they are more important than static learning and memorization of facts. Always prize creativity and out-of-the-box thinking. As you progress in your career, you will learn that people are the most important asset. If someone "steals" your ideas, which does happen, remember that this is a form of flattery. It is not sufficient to only generate one great idea in your scientific life (the, indeed, it should be taken away …). Rather, one needs to generate new ideas continuously to cultivate individual creativity.

Discussion

Limited open-source data is a major bottleneck to AI approaches in many areas including drug discovery and design. It is hoped that synergy between academia, start-ups, and pharmaceutical companies will further increase available data for learning, accelerate the design of new drug candidates, and reduce the gap that often exists between academia and industry. This may, however, be a fond hope as the entities in the pharmaceutical industry typically have different research agendas from academic scientists, and there is, of course, the issue of proprietary data that is an important constraint on the sharing of data generated within pharmaceutical companies.

Chemoinformatic methods, including ML/DL approaches, offer significant benefits for the discovery and development of bioactive compounds. However, one of the major drawbacks of ML/DL methods discussed during the Colloquium was the lack of or limited interpretability of their predictions. This is more evident for DL approaches, in which the user has no knowledge about internal features (or priorities) of the model and their assignment.

Poorly curated databases and unbalanced datasets also complicate model assessment and interpretation. Better benchmarks and guidelines need to be established for the characterization and analysis of ML models, following the example of quantitative structure-activity relationship modeling.

It was also pointed out during the conference that regardless of the many statistics and metrics available to evaluate the performance of a predictive model, "true" validation requires prospective predictions and their experimental assessment. However, prospective predictions are not without pitfalls and thus require careful evaluation of the interdisciplinary context in which such predictions and associated experiments are conducted. Machine and deep learning models are only approximations to the underlying mechanistic components of the system under investigation. In this case, as Oprea pointed out we should ask ourselves: "Is what I am doing relevant to the problem I am trying to solve?"

Regardless of the speakers' diverse research environments and settings (Table 1), it was clear from the meeting that the number of opportunities in ML in career development is increasing. This is happening in academia, in research institutes, and in large and small pharmaceutical companies. This outcome from the meeting was valuable for the students, particularly those wondering about their professional future in this area and having to decide about their next career steps [88]. It was also valuable for students and early career investigators to become aware of the career paths of many speakers who have transitioned from different disciplines and have made significant scientific contributions in the exciting computer-aided drug design field. Several speakers with 20 to 30 or more years of experience, made the transition to computer-aided drug discovery from quantum mechanics, organic chemistry, biochemistry, computer engineering, medicine, and pharmacology. Their career paths are varied, and there is not a single straight path from one discipline to another. Research interests and opportunities evolve, and researchers adapt to the current needs, which can change.

During the meeting, some speakers shared their experiences in scientific publishing (which is crucial in science and has practical implications in academia). A highlight is that the speakers emphasize the need to be persistent while pursuing a research idea. For example, Prof. Gasteiger shared that his most cited paper was initially rejected for publication three times. This message is crucial for students and young scientists who often get discouraged by the rejection of a submitted manuscript. The message is that 'persistence pays off'.

Figure 1 shows the impact of chemoinformatics and AI approaches that have been around at all stages of the drug-discovery process, from target selection to the pharmacovigilance of approved drugs. The current technologies allow the use of a huge diversity of data (atomic, chemical, biological, clinical, and post market data) in combination with different approaches (e.g., data fusion, clustering, ML, DL, pairwise comparisons, dimensionality reduction, and networks) to classify, predict, or recognize patterns in order to explain or decode new knowledge, opening up a vast repertoire of possible combinations of methods that are applicable to the solution of drug-design problems.

Conclusion

The virtual Chemoinformatics and Artificial Intelligence Colloquium, Mexico City, June 15–17, 2022, provided an overview of the current developments, specific applications, and areas of opportunity in the application of AI, ML, and DL methods to the discovery and design of bioactive molecules. The perspective was provided by speakers at different career levels working in different research environments worldwide. During the colloquium, the role of chemists, chemoinformaticians, and data scientists in accelerating drug discovery and development, which regularly takes 10–15 years, was discussed.

The colloquium was the first open-access event hosted in a country in Latin America focused on chemoinformatics and AI and open to the scientific community, as it was accessible to registrants from more than 60 countries. It is expected that in the next few years, the Latin American community will be more integrated with chemoinformatics and AI methods being developed worldwide. Since it is known that scientific English can be a barrier for many that must be overcome, courses in English at the undergraduate level will be offered to promote practice among the students. Future editions of the meeting will include hands-on tutorials/workshops and poster/oral presentations by students. Also, it is expected that future meetings will be hybrid in order to benefit from one-on-one discussions and to facilitate the rapid

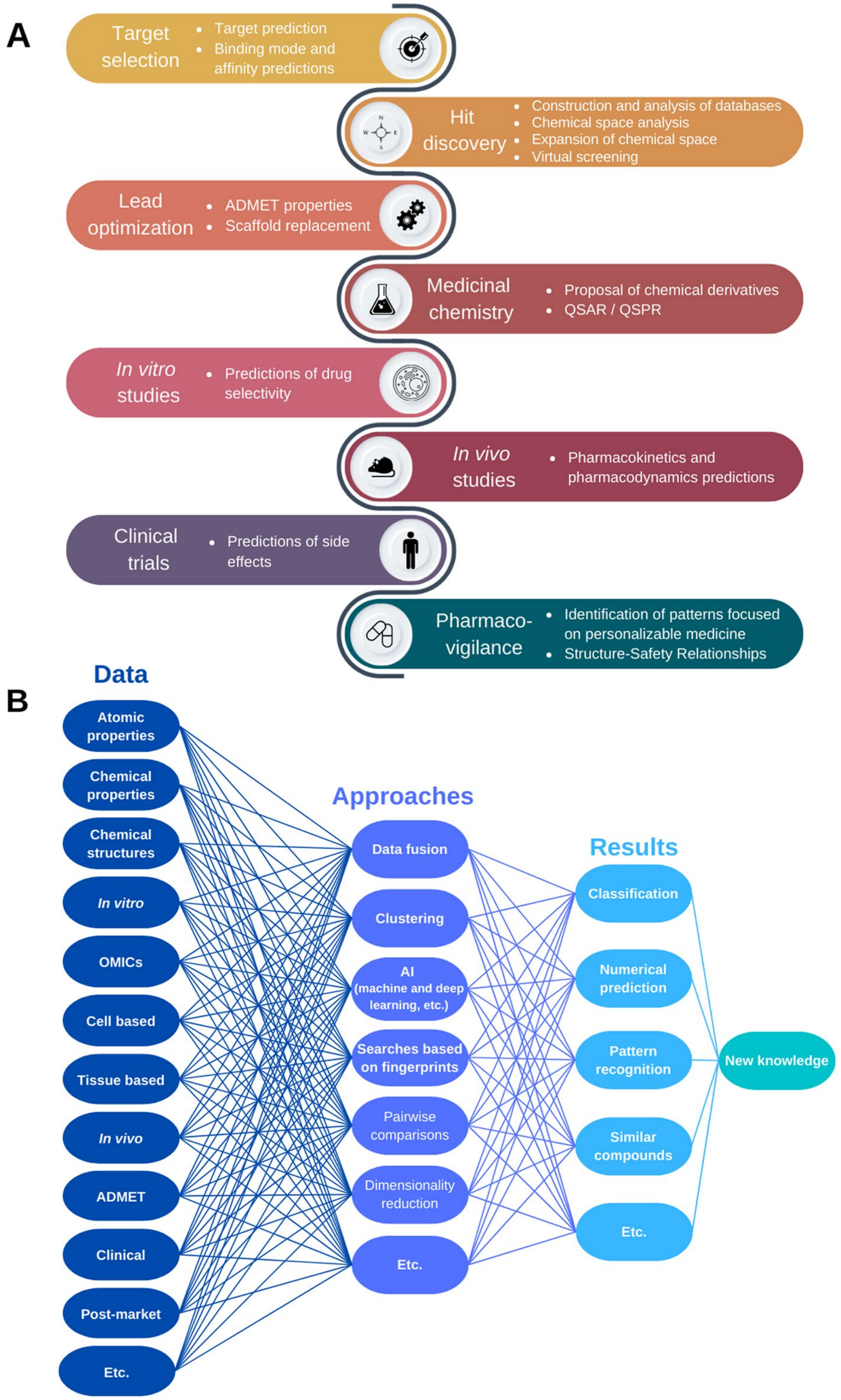


Fig. 1 Overview of applicability of chemoinformatics and AI technologies on drug design. **A** Main contributions of chemoinformatics and AI technologies on each step in the drug design process. **B** Combination of data, approaches, and type of results used in drug design

dissemination and contact with interested persons for which traveling is difficult.

The current colloquium is an early but hopefully continued effort to join other educational events on chemoinformatics that have a long tradition such as the chemoinformatics and pharmacy informatics schools that are periodically held at the University of Strasbourg in France, or the University of Vienna in Austria.

Acknowledgements
We thank all the speakers and audience on diverse platforms (Zoom, YouTube, and Facebook). Thanks a lot to UNAM-Directors: Dr. Carlos Amador, Dr. Costas Basin, Dr. Francisco Hernandez Luis. Thanks a lot to people at support teams, Informatics, Communication, and Outreach, Aida Hernández, Edgar López, Cedric Reyes, Margarita Macías, Verónica Ramón, Vianey Islas, José Adelfo Martínez. Thank you so much to Alejandro Ochoa Alvarez and Hassan Dilan Villegas for their valuable contributions to diffusion/announcements and all members of the DIFACQUIM research group for their support.

Author contributions
All authors have contributed equally to the present manuscript. All authors have read and approved the final manuscript.

Funding
We thank the support of the School of Chemistry of the National Autonomous University of Mexico.

Availability of data and materials
The summary of registrations is available at https://doi.org/10.6084/m9.figshare.20113169. Complete recordings of all three days (sessions) are freely available on YouTube, and the full program is freely accessible here https://www.difacquim.com/english/events/2022-colloquium/.

Declarations

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Department of Life Science Informatics, B-IT, LIMES Program Unit Chemical Biology and Medicinal Chemistry, Rheinische Friedrich-Wilhelms-Universität, Friedrich-Hirzebruch-Allee 5/6, 53113 Bonn, Germany. [2]DIFACQUIM Research Group, Department of Pharmacy, School of Chemistry, National Autonomous University of Mexico, 04510 Mexico City, Mexico. [3]Ersilia Open Source Initiative, Cambridge, UK. [4]Joint IRB-BSC-CRG Programme in Computational Biology, Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, Barcelona, Catalonia, Spain. [5]Nanosafety Laboratory, International Iberian Nanotechnology Laboratory, 4715-330 Braga, Portugal. [6]Computer-Chemie-Centrum, University of Erlangen-Nuremberg, Erlangen, Germany. [7]Department of Pharmacology, Center for Research and Advanced Studies of the National Polytechnic Institute (CINVESTAV), 07360 Mexico City, Mexico. [8]BIO5 Institute, University of Arizona, Tucson, AZ 85721, USA. [9]Recursion Pharmaceuticals, Salt Lake City, USA. [10]Chemotargets SL, Baldiri Reixac 4, Parc Cientific de Barcelona (PCB), 08028 Barcelona, Catalonia, Spain. [11]Research Group on Systems Pharmacology, Research Program on Biomedical Informatics (GRIB), IMIM Hospital del Mar Medical Research Institute and University Pompeu Fabra, Parc de Recerca Biomedica (PRBB), 08003 Barcelona, Catalonia, Spain. [12]Department of Chemistry, University of Florida, Gainesville, FL 32603, USA. [13]Department of Internal Medicine, University of New Mexico School of Medicine, Albuquerque, NM 87131, USA. [14]Department of Rheumatology and Inflammation Research, Institute of Medicine, Sahlgrenska Academy at Gothenburg University, 40530 Gothenburg, Sweden. [15]Novo Nordisk Foundation Center for Protein Research, Faculty of Health and Medical Sciences, University of Copenhagen, 2200 Copenhagen, Denmark. [16]Present Address: Roivant Discovery Sciences, Inc., 451 D Street, Boston, MA 02210, USA. [17]Department of Biotechnology and Biochemistry, Center for Research and Advanced Studies of the National Polytechnic Institute (CINVESTAV-IPN), Irapuato Unit, 36824 Irapuato, Gto, Mexico. [18]Chemistry Institute, National Autonomous University of Mexico, 04510 Mexico City, Mexico. [19]Novartis Institutes for Biomedical Research, 4002 Basel, Switzerland. [20]Universidad de Santander, Facultad de Ciencias Médicas y de la Salud, Instituto de Investigación Masira, Calle 70 No. 55-210, 680003 Santander, Bucaramanga, Colombia. [21]Instituto de Química, Unidad Mérida, Universidad Nacional Autónoma de México, Carretera Mérida-Tetiz Km. 4.5, Yucatán, 97357 Ucú, Mexico. [22]Nuclei of Bioassays, Biosynthesis and Ecophysiology of Natural Products (NuBBE), Department of Organic Chemistry, Institute of Chemistry, São Paulo State University-UNESP, Araraquara, Brazil.

Received: 9 August 2022 Accepted: 25 November 2022
Published online: 02 December 2022

References

1. Ntie-Kang F, Telukunta KK, Fobofou SAT et al (2021) Computational applications in secondary metabolite discovery (CAiSMD): an online workshop. J Cheminform 13:64
2. Wu J, Rajesh A, Huang Y-N et al (2021) Virtual meetings promise to eliminate geographical and administrative barriers and increase accessibility, diversity and inclusivity. Nat Biotechnol 40:133–137
3. Medina-Franco JL, López-López E (2022) The essence and transcendence of scientific publishing. Front Res Metr Anal 7:822453
4. Engel T, Gasteiger J (eds) (2018) Chemoinformatics—basic concepts and methods. Wiley, Hoboken.
5. Gasteiger J (2020) Chemistry in times of artificial intelligence. Chemphyschem 21:2233–2242
6. Engel T, Gasteiger J (eds) (2018) Applied chemoinformatics—achievements and future opportunities. Wiley, Hoboken
7. Zupan J, Gasteiger J (1999) Neural networks in chemistry and drug design, 2nd edn. Wiley, Hoboken
8. NuBBE database. http://nubbe.iq.unesp.br/portal/nubbe-search.html. Accessed 28 Jul 2022
9. Smith JS, Isayev O, Roitberg AE (2017) ANI-1: an extensible neural network potential with DFT accuracy at force field computational cost. Chem Sci 8:3192–3203
10. Jumper J, Evans R, Pritzel A et al (2021) Highly accurate protein structure prediction with AlphaFold. Nature 596:583–589
11. Chen M (2021) Collective variable-based enhanced sampling and machine learning. Eur Phys J B 94:211
12. Cruz J, Rondon-Villarreal P, Torres RG et al (2018) Design of bactericidal peptides against *Escherichia coli* O157:H7, *Pseudomonas aeruginosa* and methicillin-resistant *Staphylococcus aureus*. Med Chem 14:741–752
13. Plisson F, Ramírez-Sánchez O, Martínez-Hernández C (2020) Machine learning-guided discovery and design of non-hemolytic peptides. Sci Rep 10:16581
14. Ersilia (2022) https://www.ersilia.io/. Accessed 10
15. Duran-Frigola M, Pauls E, Guitart-Pla O, Bertoni M, Alcalde V, Amat D, Juan-Blanco T, Aloy P (2020) Extending the small-molecule similarity principle to all levels of biology with the Chemical checker. Nat Biotechnol 38:1087–1096
16. Prieto-Martínez FD, Fernández-de Gortari E, Medina-Franco JL, Espinoza-Fonseca LM (2021) An in silico pipeline for the discovery of multitarget ligands: a case study for epi-polypharmacology based on DNMT1/HDAC2 inhibition. Artif Intell Life Sci 1:100008
17. Polishchuk P (2020) CReM: chemically reasonable mutations framework for structure generation. J Cheminform 12:28
18. Winter R, Montanari F, Steffen A, Briem H, Noé F, Clevert D-A (2019) Efficient multi-objective molecular optimization in a continuous latent space. Chem Sci 10:8016–8024
19. Chemotargets (2022) https://chemotargets.com/services/. Accessed 10
20. Hamzic S, Lewis R, Desrayaud S, Soylu C, Fortunato M, Gerebtzoff G, Rodríguez-Pérez R (2022) Predicting in vivo compound brain penetration using multi-task graph neural networks. J Chem Inf Model 62:3180–3190
21. Rodríguez-Pérez R, Gerebtzoff G (2021) Identification of bile salt export pump inhibitors using machine learning: predictive safety from an industry perspective. Artif Intell Life Sci 1:100027
22. Petit J, Meurice N, Kaiser C, Maggiora G (2012) Softening the rule of five–where to draw the line? Bioorg Med Chem 20:5343–5351

23. Miranda-Quintana RA, Bajusz D, Rácz A, Héberger K (2021) Extended similarity indices: the benefits of comparing more than two objects simultaneously. Part 1: theory and characteristics†. J Cheminform 13:32
24. Miranda-Quintana RA, Rácz A, Bajusz D, Héberger K (2021) Extended similarity indices: the benefits of comparing more than two objects simultaneously. Part 2: speed, consistency, diversity selection. J Cheminform 13:33
25. Yoshimori A, Bajorath J (2021) Iterative DeepSARM modeling for compound optimization. Artif Intell Life Sci 1:100015
26. Yoshimori A, Bajorath J (2020) Deep SAR matrix: SAR matrix expansion for advanced analog design using deep learning architectures. Future Drug Discov 2:FDD36
27. Méndez-Lucio O, Ahmad M, del Rio-Chanona EA, Wegner JK (2021) A geometric deep learning approach to predict binding conformations of bioactive molecules. Nat Mach Intell 3:1033–1039
28. Oprea TI, Bologa CG, Brunak S et al (2018) Unexplored therapeutic opportunities in the human genome. Nat Rev Drug Discov 17:317–332
29. Binder J, Ursu O, Bologa C et al (2022) Machine learning prediction and tau-based screening identifies potential Alzheimer's disease genes relevant to immunity. Commun Biol 5:125
30. Medina-Franco JL, Martinez-Mayorga K, Fernández-de Gortari E, Kirchmair J, Bajorath J (2021) Rationality over fashion and hype in drug design. F1000 Research 10:397
31. Zupan J, Novič M, Li X, Gasteiger J (1994) Classification of multicomponent analytical data of olive oils using different neural networks. Anal Chim Acta 292:219–234
32. Saldívar-González FI, Aldas-Bulos VD, Medina-Franco JL, Plisson F (2022) Natural product drug discovery in the artificial intelligence era. Chem Sci 13:1526–1546
33. López-López E, Fernández-de Gortari E, Medina-Franco JL (2022) Yes SIR! On the structure-inactivity relationships in drug discovery. Drug Discov Today 27:2353–2362
34. Sánchez-Cruz N, Pilón-Jiménez BA, Medina-Franco JL (2019) Functional group and diversity analysis of BIOFACQUIM: a mexican natural product database. F1000 Research 8:2071
35. Pilon AC, Valli M, Dametto AC, Pinto MEF, Freire RT, Castro-Gamboa I, Andricopulo AD, Bolzani VS (2017) NuBBEDB: an updated database to uncover chemical and biological information from brazilian biodiversity. Sci Rep 7:7215
36. Medina-Franco JL, López-López E, Andrade E, Ruiz-Azuara L, Frei A, Guan D, Zuegg J, Blaskovich MAT (2022) Bridging informatics and medicinal inorganic chemistry: toward a database of metallodrugs and metallodrug candidates. Drug Discov Today 27:1420–1430
37. Rogers D, Hahn M (2010) Extended-connectivity fingerprints. J Chem Inf Model 50:742–754
38. Durant JL, Leland BA, Henry DR, Nourse JG (2002) Reoptimization of MDL keys for use in drug discovery. J Chem Inf Comput Sci 42:1273–1280
39. Johnson M, Maggiora GM (eds) (1990) Concepts and applications of molecular similarity. Wiley, New York
40. Tanimoto T(1958) An elementary mathematical theory of classification and prediction. Internal IBM Technical Report
41. Jaccard P (1912) The distribution of the flora in the alpine zone.1. New Phytol 11:37–50
42. Martin YC, Kofron JL, Traphagen LM (2002) Do structurally similar molecules have similar biological activity? J Med Chem 45:4350–4358
43. Bellman RE (2003) Dynamic programming. Courier Dover Publications, Inc, USA
44. Bellman R (1961) Adaptive control processes: a guided tour. Princeton University Press, USA
45. Maggiora GM, Bajorath J (2014) Chemical space networks: a powerful new paradigm for the description of chemical space. J Comput-Aided Mol Des 28:795–802
46. Maggiora GM (2014) Introduction to molecular similarity and chemical space. In: Martinez-Mayorga K, Medina-Franco JL (eds) Foodinformatics: applications of chemical information to food chemistry. Springer International Publishing, Cham, pp 1–81
47. Medina-Franco JL, Chávez-Hernández AL, López-López E, Saldívar-González FI (2022) Chemical multiverse: an expanded view of chemical space. Mol Inf 41:2200116
48. Gentile F, Agrawal V, Hsing M, Ton A-T, Ban F, Norinder U, Gleave ME, Cherkasov A (2020) Deep docking: a deep learning platform for augmentation of structure based drug discovery. ACS Cent Sci 6:939–949
49. Gentile F, Yaacoub JC, Gleave J, Fernandez M, Ton A-T, Ban F, Stern A, Cherkasov A (2022) Artificial intelligence-enabled virtual screening of ultra-large chemical libraries with deep docking. Nat Protoc 17:672–697
50. Baek M, DiMaio F, Anishchenko I et al (2021) Accurate prediction of protein structures and interactions using a three-track neural network. Science 373:871–876
51. Baek M, Baker D (2022) Deep learning and protein structure modeling. Nat Methods 19:13–14
52. Silverman RB (2008) From basic science to blockbuster drug: the discovery of Lyrica. Angew Chem Int Ed Engl 47:3500–3504
53. Garcia-Serna R, Vidal D, Remez N, Mestres J (2015) Large-scale predictive drug safety: from structural alerts to biological mechanisms. Chem Res Toxicol 28:1875–1887
54. Waszkowycz B, Clark DE, Gancia E (2011) Outstanding challenges in protein–ligand docking and structure-based virtual screening. Wiley Interdiscip Rev Comput Mol Sci 1:229–259
55. Ross GA, Morris GM, Biggin PC (2013) One size does not fit all: the limits of structure-based models in drug discovery. J Chem Theory Comput 9:4266–4274
56. Hameduh T, Haddad Y, Adam V, Heger Z (2020) Homology modeling in the time of collective and artificial intelligence. Comput Struct Biotechnol J 18:3494–3506
57. Valanciute A, Nygaard L, Zschach H, Jepsen MM, Lindorff-Larsen K, Stein A (2022) Accurate protein stability predictions from homology models. bioRxiv 2022.07.12.499700
58. Zanette C, Bannan CC, Bayly CI, Fass J, Gilson MK, Shirts MR, Chodera JD, Mobley DL (2019) Toward learned chemical perception of force field typing rules. J Chem Theory Comput 15:402–423
59. Bonati L, Rizzi V, Parrinello M (2020) Data-driven collective variables for enhanced sampling. J Phys Chem Lett 11:2998–3004
60. Noé F, Olsson S, Köhler J, Wu H (2019) Boltzmann generators: sampling equilibrium states of many-body systems with deep learning. Science 365:eaaw1147
61. Wang DD, Chan M-T, Yan H (2021) Structure-based protein–ligand interaction fingerprints for binding affinity prediction. Comput Struct Biotechnol J 19:6291–6300
62. Da C, Kireev D (2014) Structural protein-ligand interaction fingerprints (SPLIF) for structure-based virtual screening: method and benchmark study. J Chem Inf Model 54:2555–2561
63. Sánchez-Cruz N, Medina-Franco JL, Mestres J, Barril X (2021) Extended connectivity interaction features: improving binding affinity prediction through chemical description. Bioinformatics 37:1376–1382
64. McNutt AT, Francoeur P, Aggarwal R, Masuda T, Meli R, Ragoza M, Sunseri J, Koes DR (2021) GNINA 1.0: molecular docking with deep learning. J Cheminform 13:43
65. Gainza P, Sverrisson F, Monti F, Rodolà E, Boscaini D, Bronstein MM, Correia BE (2020) Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning. Nat Methods 17:184–192
66. Morris C, Ritzert M, Fey M, Hamilton WL, Lenssen JE, Rattan G, Grohe M (2018) Weisfeiler and Leman go neural: higher-order graph neural networks. arXiv:1810.02244.
67. Chemotargets announces first ai-designed drug for Huntington's disease to enter clinical trials. https://chemotargets.com/chemotargets-announces-first-ai-designed-drug-for-huntingtons-disease-to-enter-clinical-trials/. Accessed 27 Jun 2022
68. Jalencas X, Mestres J (2013) Chemoisosterism in the proteome. J Chem Inf Model 53:279–292
69. Xu D, Zhang Y (2009) Generating triangulated macromolecular surfaces by euclidean distance transform. PLoS ONE 4:e8140
70. Wang Y, Xing J, Xu Y et al (2015) In silico ADME/T modelling for rational drug design. Q Rev Biophys 48:488–515
71. Huang K, Fu T, Gao W, Zhao Y, Roohani Y, Leskovec J, Coley CW, Xiao C, Sun J, Zitnik M(2021) Therapeutics Data Commons: Machine learning datasets and tasks for drug discovery and development. arXiv:2102.09548v2
72. Therapeutics Data Commons. https://tdcommons.ai/. Accessed 20 Jul 2022

73. Chemotargets(2022) CLARITY PV. https://chemotargets.com/clarity-pv/. Accessed 11 Jul 2022
74. Pye CR, Bertin MJ, Lokey RS, Gerwick WH, Linington RG (2017) Retrospective analysis of natural products provides insights for future discovery trends. Proc Natl Acad Sci USA 114:5601–5606
75. Valli M, dos Santos RN, Figueira LD, Nakajima CH, Castro-Gamboa I, Andricopulo AD, Bolzani VS (2013) Development of a natural products database from the biodiversity of Brazil. J Nat Prod 76:439–444
76. Saldívar-González FI, Valli M, Andricopulo AD, da Silva Bolzani V, Medina-Franco JL (2019) Chemical space and diversity of the NuBBE database: a chemoinformatic characterization. J Chem Inf Model 59:74–85
77. Rutz A, Sorokina M, Galgonek J et al (2022) The LOTUS initiative for open knowledge management in natural products research. Elife 11:e70780
78. Chen Y, Kirchmair J (2020) Cheminformatics in natural product-based drug discovery. Mol Inf 39:e2000171
79. de Oliveira AS, Valli M, Ferreira LL et al (2022) Novel trypanocidal thiophen-chalcone cruzain inhibitors: structure- and ligand-based studies. Future Med Chem 14:795–808
80. Fjell CD, Hiss JA, Hancock REW, Schneider G (2011) Designing antimicrobial peptides: form follows function. Nat Rev Drug Discov 11:37–51
81. Cruz J, Suárez-Barrera MO, Rondón-Villarreal P, Olarte-Diaz A, Guzmán F, Visser L, Rueda-Forero NJ (2021) Computational study, synthesis and evaluation of active peptides derived from Parasporin-2 and spike protein from Alphacoronavirus against colorectal cancer cells. Biosci Rep 41:BSR20211964
82. Ropero-Vega JL, Redondo-Ortega JF, Rodríguez-Caicedo JP, Rondón-Villarreal P, Flórez-Castillo JM (2022) New PEPTIR-2.0 peptide designed for use as recognition element in electrochemical biosensors with improved specificity towards *E. coli* O157:H7. Molecules 27:2704
83. Huan Y, Kong Q, Mou H, Yi H (2020) Antimicrobial peptides: classification, design, application and research progress in multiple fields. Front Microbiol 11:582779
84. Nguyen LT, Haney EF, Vogel HJ (2011) The expanding scope of antimicrobial peptide structures and their modes of action. Trends Biotechnol 29:464–472
85. Waghu FH, Barai RS, Gurung P, Idicula-Thomas S (2016) CAMPR3: a database on sequences, structures and signatures of antimicrobial peptides. Nucleic Acids Res 44:D1094–D1097
86. Melo MCR, Maasch JRMA, de la Fuente-Nuñez C (2021) Accelerating antibiotic discovery through artificial intelligence. Commun Biol 4:1050
87. Das P, Sercu T, Wadhawan K et al (2021) Accelerated antimicrobial discovery via deep generative models and molecular dynamics simulations. Nat Biomed Eng 5:613–623
88. Medina-Franco JL(2021) DeLIRa: decisions-life impact relationships and decision cliffs in career development. Available at SSRN: https://doi.org/10.2139/ssrn.3973083

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Manual de Quimioinformática en español

JOURNAL OF CHEMICAL EDUCATION

pubs.acs.org/jchemeduc

Technology Report

A Spanish Chemoinformatics GitBook for Chemical Data Retrieval and Analysis Using Python Programming

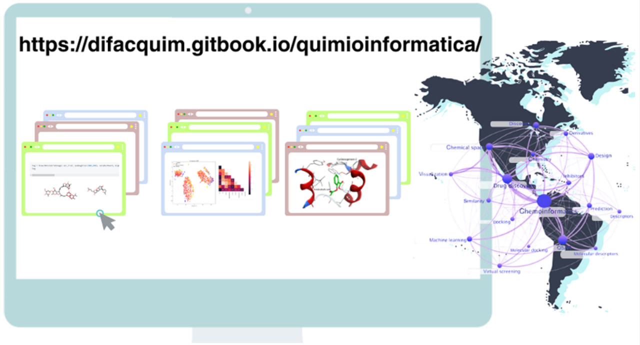
Fernanda I. Saldivar-González,* Diana L. Prado-Romero, Raziel Cedillo-González, Ana L. Chávez-Hernández, Juan F. Avellaneda-Tamayo, Alejandro Gómez-García, Luis Juárez-Rivera, and José L. Medina-Franco*

Cite This: https://doi.org/10.1021/acs.jchemed.4c00041

Read Online

ACCESS | Metrics & More | Article Recommendations

ABSTRACT: Searching, retrieving, and analyzing chemical information are among the main tasks faced by students and professionals in chemistry-related scientific disciplines. Currently, freely available modules developed in programming languages, such as Python, allow efficient data management and facilitate the obtaining of information and knowledge from the data. This article describes an electronic handbook generated on the GitBook platform to introduce the Python programming language and the analysis, computational representation, and visualization of chemical data. This manual explores the most common molecular representations of low molecular weight organic compounds and their applications in various contexts. It also illustrates the acquisition of chemical data from large public molecular databases such as ChEMBL and PubChem and the analysis and visualization of chemical information using concepts such as chemical space. The GitBook is freely available (https://difacquim.gitbook.io/quimioinformatica/) and is expected to foster open science and facilitate learning for chemistry students at the undergraduate and graduate levels, as well as professionals interested in chemical data analysis and visualization.



KEYWORDS: *Chemoinformatics, Scientific Education, Latin America, Python, Spanish-Speaking Community, Open Science, Handbook*

■ INTRODUCTION

Chemoinformatics is one of the independent disciplines that has become a pillar during the development and design of new drugs and, therefore, is indispensable in pharmaceutical chemistry. This area of knowledge allows one to solve problems in the management and presentation of information in chemistry by integrating different computational techniques and methods.[1] Chemoinformatics merges chemistry and informatics to solve tasks in chemistry. Figure 1 schematically illustrates the broad applications of Chemoinformatics in drug discovery and many other chemistry areas.[2]

One of Chemoinformatics's most widely acknowledged accomplishments is its contribution to providing access to chemical information within databases.[3] The vast volume of data related to chemical compounds, encompassing their physical, chemical, and biological properties, has promoted the creation of databases designed for efficient storage and electronic dissemination of this information. Various computational methods have been devised to further enhance the utility of these databases. These methods facilitate more effective information retrieval by enabling comprehensive searches based on complete structures, substructures, and similarity. This approach streamlines data mining and enhances the overall efficiency of database searches. Examples of such computational methods encompass virtual screening campaigns involving molecular docking and similarity searching, pharmacophoric modeling, quantitative structure−activity relationship (QSAR) analyses in both 2D and 3D formats, ligand-based drug design, and fragment-based drug design, among others continuously evolving within the field.[4]

In the same context, different types of molecular representations have enabled improved searches and expanded applications in various areas of chemistry. Examples include the development of new chemical compounds (*de novo* design),[5,6] property predictions such as absorption, distribution, metabolism, and excretion (ADME),[7] structure−activity

Received: January 12, 2024
Revised: May 4, 2024
Accepted: May 20, 2024

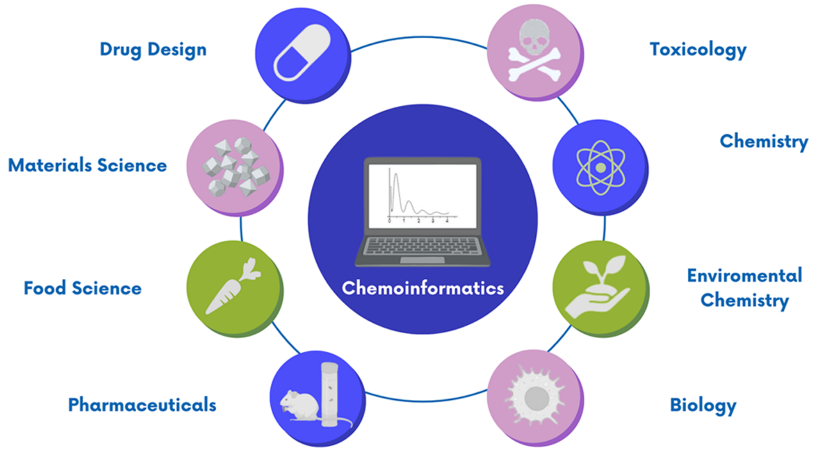


Figure 1. Overview of the applications of Chemoinformatics.

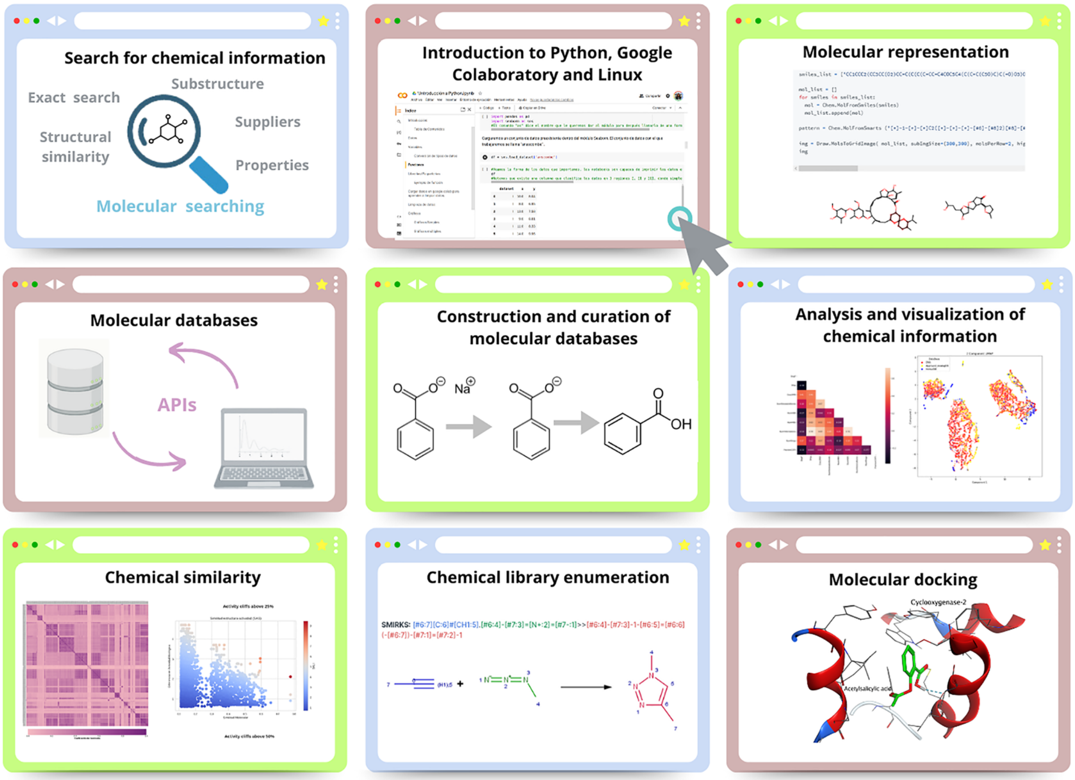


Figure 2. Snapshots of the nine main chapters of the GitBook.

relationships (SARs), structure−properties relationships (SPRs),[8] and development of new chemical descriptors.[9] Notable examples of molecular descriptors commonly employed in drug discovery applications include structural fingerprints and molecular properties. Fingerprints encode molecular fragments or functional groups of a molecule, such as Molecular ACCes System (MACCS) Keys[10] and extended connectivity fingerprints (ECFPs).[11] Typical calculated molecular properties are whole molecular properties of pharmaceutical interest that are part of empirical rules to assess drug-likeness.[12]

Chemical space analysis, molecular docking, and application of similarity concepts are prominent topics in computer-aided drug design (CADD). These topics are widely utilized in the pharmaceutical industry, universities, and research centers, demonstrating high-frequency drug discovery and development applications. The chemical space is the conceptual basis of Chemoinformatics,[13] several definitions are reviewed elsewhere.[14] Recently, the notion of the chemical multiverse has emerged, representing a collection of chemical spaces for a given set of compounds, each characterized by a distinct set of descriptors.[15] This concept finds applications in drug design, encompassing diversity analysis, SAR and SPR analysis, and the design of molecular libraries.[14]

Applications of the similarity concept, such as SAR/SPR analysis and activity landscapes, have been very useful in describing changes in biological activity associated with changes in chemical structures and the subsequent design of new compounds with improved activities.[16] Often, these differences can be confirmed with tools such as molecular

Table 1. Contents of the Chemoinformatics GitBook

Chapter	Content	Objectives
1. Search for chemical information.	• SciFinder-n	• Enhance the skills of students and professionals in the chemical field to seek scientific information proficiently.
	•Web of Science	• Become familiar with diverse scientific information types and various tools or engines employed to search for chemical information.
	•Scopus	• Evaluate and choose information based on distinct search criteria.
	•CAS Source Index (CASSI)	• Recognize the varied forms of scientific publications, understanding their structure and content.
	•Bibliometrics: tools and software	• Acquaint oneself with tools and software employed in bibliometric analysis and visualization.
2. Introduction to Python, Linux, and Google Colaboratory.	• Fundamentals of programming	• Introduce basic Python definitions and functions.
	• Data cleaning	• Introduce the concept of packages.
	• Installation of the environment in local and WSL-based	• Learn how to import, manage, and clean existing data sets.
	• Basic commands on Linux	• Introduce basic commands in Bash programming language for Linux interaction from the terminal.
		• Introduce Bash basic commands for managing, processing, and analyzing data from the Linux terminal.
3. Molecular representation.	• SMILES	• Introduce the most common molecular representations of low molecular weight organic compounds and their applications in various contexts.
	• SMARTS	• Learn how to convert compounds between the different molecular representations as appropriate.
	• InChI/InchI keys	• Introduce the use of RDKit, py3Dmol, and smilesDrawer packages to manage chemical structures.
		• Apply the knowledge learned to compare structures, filter databases, and visualize molecules with specific characteristics.
4. Molecular databases.	• PubChem	• Acquire proficiency in utilizing databases pertinent to drug research, including ChEMBL, PubChem, DrugBank, and ZINC
	• ChEMBL	• Identify the specific categories of information accessible within each resource, enabling streamlined and efficient information retrieval.
	• DrugBank	• Become familiar with using APIs to access information from public databases programmatically.
	• ZINC	
	• ChemSpider	
5. Construction and curation of molecular databases.	• Construction of compound databases	• Build compound databases annotated with biological activity.
		• Acquire knowledge about the molecular characteristics necessary for subsequent *in silico* studies.
		• Identify and eliminate molecules that could alter computational calculations.
		• Curate compound databases using RDKit and Molvs modules.
6. Analysis and visualization of chemical information.	• Calculation and analysis of molecular descriptors	• Introduce EDA for chemical data.
	• Visualization of chemical space	• Employ visual methodologies to examine physicochemical properties crucial to pharmaceutical applications, along with descriptors linked to molecular complexity.
		• Explore potential correlations among various variables within the data set.
		• Utilize chemical space visualization methods to generate comprehensive profiles of chemical databases.
		• Introduce the concept of a chemical multiverse and showcase it through a Chemical Art gallery.
7. Chemical similarity.	• Molecular representation (fingerprints)	• Introduce the concept of chemical similarity and its applications in drug design.
	• Similarity functions	• Acquire a fundamental understanding of the critical components used to assess the similarity between chemical compounds.
	• QSAR	• Study structure−activity relationships through QSAR and activity landscape modeling.
	• Activity landscapes	
8. Chemical library enumeration.	• Chemical reactions	• Illustrate examples of virtual chemical library enumeration.
	• Transformation rules	• Gain proficiency in employing SMARTS and SMIRKS for encoding chemical reactions and transformations.
9. Molecular docking.	• LeDock	• Give a general, nonexhaustive overview of what a molecular docking study is.
	• AutoDock Vina	• Explain the steps for a protein−ligand molecular docking study with two open-access programs.

docking, which consider the mechanism of action at a structural level. However, their application is limited to the availability of information related to the therapeutic target.

To explore the topics mentioned above further, the following is a Spanish handbook that addresses concepts and tools of Chemoinformatics with applications in drug design. In recent years, there have been several contributions to teaching Chemoinformatics[17−19] and machine learning for chemists in an organized and formal manner.[20,21] However, it is still necessary to improve the teaching and dissemination of the applications of Chemoinformatics in Latin America.[22]

Teaching the applications and basic concepts of Chemoinformatics equips professionals with advanced skills in using computational tools for chemical data analysis, which benefits both academic research and the chemical and pharmaceutical industries. It also contributes to fostering international collaboration and technological expertise.

In this context, the goal of developing an electronic handbook on Chemoinformatics in Spanish is to strengthen the understanding of its basic principles in chemistry students and professionals and to contribute to the user's ability to handle and interpret computational techniques associated with this scientific discipline in the context of bioactive compounds. This, in turn, will contribute to the formation of students and researchers who want to learn and benefit from the appropriate use and implementation of computational methods for their professional development.

GitBook Structure and Content

The Chemoinformatics handbook is implemented within the GitBook platform (https://difacquim.gitbook.io/quimioinformatica/). We opted for GitBook due to its suitability for our specific needs, providing accessibility and user-friendliness for a diverse range of contributors. Additionally, GitBook's support for various content formats and simple customization features enabled us to present our content in an appealing and organized manner. As depicted in Figure 2, the GitBook is structured into nine chapters that seek to promote the acquisition of basic Chemoinformatics concepts and develop competencies for the search, acquisition, and analysis of chemical information by employing programming and open-access computational tools. The handbook covers a broad and diverse range of topics, including a general introduction to Python concepts and packages and basic commands on Linux. It also covers the search and analysis of chemical information using different databases and software for bibliometric visualization. It guides users on the applications of different molecular representations of low molecular weight organic compounds, further provides proficiency in utilizing databases pertinent to drug research through Application Programming Interfaces (APIs), and instructs on building compound databases annotated with biological activity. Additionally, it introduces Exploratory Data Analysis (EDA) for examining physicochemical properties, exploring correlations within data sets, and visualizing the chemical space. The concept of a chemical multiverse is introduced and exemplified through a chemical art gallery.[23] Applications of the similarity concept to conduct QSARs and activity landscape modeling are also covered. The GitBook also exemplifies published examples of chemical library enumeration in more detail.[24,25] This can prove highly beneficial for investigating proposed methodologies in chemical synthesis, facilitating the exploration of an affordable chemical space through the utilization of open-access Chemoinformatics tools. Finally, a general overview of molecular docking is provided, including steps to conduct protein−ligand molecular docking studies using open-access programs LeDock[26] and AutoDock Vina.[27,28]

Table 1 summarizes the contents and main objectives of each handbook chapter. In each chapter, the objectives are mentioned, followed by an introduction of the most essential concepts for each topic and their relevance. Subsequently, procedures are developed with "applicable" examples in research. The chapters end with exercises to reinforce the topics learned. The related Notebooks to explain each topic were developed on Google Colaboratory and can be found at https://github.com/DIFACQUIM/Cursos. The DIFACQUIM/Cursos repository is under the MIT license. "A short and simple permissive license with conditions only requiring preservation of copyright and license notices." Licensed works, modifications, and larger works may be distributed under different terms and without source code.

Implementation and Discussion

An analysis of the current trends and challenges confronting Chemoinformatics in Latin America underscores the uneven trajectory of this discipline's growth within the region, in contrast to its counterparts in Europe, Asia, and North America.[22] In Latin America, the lack of sustained funding, adequate infrastructure, supportive policies and offering of Chemoinformatics-related subjects in academic curricula has resulted in a slower and fragmented progression of Chemoinformatics.[29]

As part of an effort to advance the field of Chemoinformatics in Latin America, the first school of Chemoinformatics in Latin America was launched in 2022.[30] Six lectures, one workshop, and one roundtable with four editors were presented during an online public event with speakers from academia, big pharma, and public research institutions. It is anticipated that this initiative will endure and that the presented material will facilitate the proposal of additional workshops in future editions.

To introduce and disseminate the GitBook among chemistry students and professionals, a free online workshop was recently conducted on the UAMedia platform (https://www.uamediadigital.com/cursos-online). The workshop lasted 20 h and covered the content summarized in Table 1, excluding molecular docking, which was subsequently included based on attendee requests. Additionally, the GitBook is being implemented as teaching material for the Chemoinformatics class at the School of Chemistry, UNAM, at the undergraduate and graduate levels. Moreover, this work has also been presented at the Mexican Chemical Society meeting and soon in the tenth edition of a Medicinal Chemistry symposium at UNAM. The aim is to continue disseminating this material among educators and professionals within the field, to foster the incorporation of this discipline into diverse educational institutions across Latin America, and to further collaboration among distinct research collectives.

It is noteworthy to emphasize that any undergraduate or graduate chemistry school can use the GitBook content because it is written at the introductory level. While most examples during the course focused on drug design, its application extends beyond and can be effectively utilized in diverse fields, such as materials science or food chemistry, among others.

■ CONCLUSIONS

Chemoinformatics is a scientific discipline that has emerged in response to the need to manage, classify, and efficiently interpret chemical information. One of the main applications of Chemoinformatics has been in drug development, since it has facilitated the integration of chemical and biological data to generate information and, ultimately, helpful knowledge (for instance, predictive models of biological activity). The Chemoinformatics handbook, constructed within the GitBook platform and organized into nine chapters, is a valuable educational resource for Spanish-speaking students and professionals entering the Chemoinformatics field. It facilitates learning in programming and highlights real and practical drug design applications. The fundamental concepts and applications can be adapted and extended to various chemical-related disciplines. The emphasis on teaching Chemoinformatics through open-access tools and the fact that the handbook is published in Spanish align with the broader goal of democratizing science and cultivating interest among students and professionals in the chemical field.

AUTHOR INFORMATION

Corresponding Authors

Fernanda I. Saldivar-González − *DIFACQUIM Research Group, Department of Pharmacy, School of Chemistry, Universidad Nacional Autónoma de México, Mexico City 04510, Mexico;* orcid.org/0000-0002-0435-8662; Email: fer.saldivarg@gmail.com

José L. Medina-Franco − *DIFACQUIM Research Group, Department of Pharmacy, School of Chemistry, Universidad Nacional Autónoma de México, Mexico City 04510, Mexico;* orcid.org/0000-0003-4940-1107; Email: medinajl@unam.mx

Authors

Diana L. Prado-Romero − *DIFACQUIM Research Group, Department of Pharmacy, School of Chemistry, Universidad Nacional Autónoma de México, Mexico City 04510, Mexico;* orcid.org/0000-0001-8918-6451

Raziel Cedillo-González − *DIFACQUIM Research Group, Department of Pharmacy, School of Chemistry, Universidad Nacional Autónoma de México, Mexico City 04510, Mexico;* orcid.org/0009-0009-9427-6959

Ana L. Chávez-Hernández − *DIFACQUIM Research Group, Department of Pharmacy, School of Chemistry, Universidad Nacional Autónoma de México, Mexico City 04510, Mexico;* orcid.org/0000-0002-6202-1769

Juan F. Avellaneda-Tamayo − *DIFACQUIM Research Group, Department of Pharmacy, School of Chemistry, Universidad Nacional Autónoma de México, Mexico City 04510, Mexico;* orcid.org/0009-0003-1819-6187

Alejandro Gómez-García − *DIFACQUIM Research Group, Department of Pharmacy, School of Chemistry, Universidad Nacional Autónoma de México, Mexico City 04510, Mexico;* orcid.org/0000-0003-4444-8221

Luis Juárez-Rivera − *DIFACQUIM Research Group, Department of Pharmacy, School of Chemistry, Universidad Nacional Autónoma de México, Mexico City 04510, Mexico*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jchemed.4c00041

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

We thank Norberto Sanchez-Cruz for his advice on the use of AutoDock Vina. The technical assistance and enthusiastic participation of Jesús Armando Rufino Valencia and Maria de la Luz Seseña Alcalde (School of Chemistry, UNAM) is greatly acknowledged. FISG, DLPR, RCL, ALCH, JFAT, and AGG thank the Consejo Nacional de Humanidades, Ciencias y Tecnologías (CONAHCyT), Mexico, for the Postgraduate grants 848061, 888207, 1099206, 847870, 1270553, and 912137, respectively.

REFERENCES

(1) Leach, A. R.; Gillet, V. J. *An Introduction to Chemoinformatics*; Springer, 2007.

(2) López-López, E.; Bajorath, J.; Medina-Franco, J. L. Informatics for Chemistry, Biology, and Biomedical Sciences. *J. Chem. Inf. Model.* **2021**, *61* (1), 26−35.

(3) Gasteiger, J. Chemoinformatics: Achievements and Challenges, a Personal View. *Molecules* **2016**, *21* (2), 151.

(4) Lin, X.; Li, X.; Lin, X. A Review on Applications of Computational Methods in Drug Screening and Design. *Molecules* **2020**, *25* (6), 1375.

(5) Mouchlis, V. D.; Afantitis, A.; Serra, A.; Fratello, M.; Papadiamantis, A. G.; Aidinis, V.; Lynch, I.; Greco, D.; Melagraki, G. Advances in de Novo Drug Design: From Conventional to Machine Learning Methods. *Int. J. Mol. Sci.* **2021**, *22* (4), 1676.

(6) Chávez-Hernández, A. L.; López-López, E.; Medina-Franco, J. L. Yin-Yang in Drug Discovery: Rethinking de Novo Design and Development of Predictive Models. *Front. Drug Des. Discovery* **2023**, 3, 1.

(7) Wang, Y.; Xing, J.; Xu, Y.; Zhou, N.; Peng, J.; Xiong, Z.; Liu, X.; Luo, X.; Luo, C.; Chen, K.; Zheng, M.; Jiang, H. In Silico ADME/T Modelling for Rational Drug Design. *Q. Rev. Biophys.* **2015**, *48* (4), 488−515.

(8) Ragno, R.; Esposito, V.; Di Mario, M.; Masiello, S.; Viscovo, M.; Cramer, R. D. Teaching and Learning Computational Drug Design: Student Investigations of 3D Quantitative Structure−Activity Relationships through Web Applications. *J. Chem. Educ.* **2020**, *97* (7), 1922−1930.

(9) Danishuddin; Khan, A. U. Descriptors and Their Selection Methods in QSAR Analysis: Paradigm for Drug Design. *Drug Discovery Today* **2016**, *21* (8), 1291−1302.

(10) Durant, J. L.; Leland, B. A.; Henry, D. R.; Nourse, J. G. Reoptimization of MDL Keys for Use in Drug Discovery. *J. Chem. Inf. Comput. Sci.* **2002**, *42* (6), 1273−1280.

(11) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50* (5), 742−754.

(12) Tian, S.; Wang, J.; Li, Y.; Li, D.; Xu, L.; Hou, T. The Application of in Silico Drug-Likeness Predictions in Pharmaceutical Research. *Adv. Drug Delivery Rev.* **2015**, *86*, 2−10.

(13) Varnek, A.; Baskin, I. I. Chemoinformatics as a Theoretical Chemistry Discipline. *Mol. Inform.* **2011**, *30* (1), 20−32.

(14) Saldívar-González, F. I.; Medina-Franco, J. L. Approaches for Enhancing the Analysis of Chemical Space for Drug Discovery. *Expert Opin. Drug Discovery* **2022**, *17* (7), 789−798.

(15) Medina-Franco, J. L.; Chávez-Hernández, A. L.; López-López, E.; Saldívar-González, F. I. Chemical Multiverse: An Expanded View of Chemical Space. *Mol. Inform.* **2022**, *41* (11), No. e2200116.

(16) Bajorath, J.; Peltason, L.; Wawer, M.; Guha, R.; Lajiness, M. S.; Van Drie, J. H. Navigating Structure−activity Landscapes. *Drug Discovery Today* **2009**, *14* (13), 698−705.

(17) Kim, S.; Bucholtz, E. C.; Briney, K.; Cornell, A. P.; Cuadros, J.; Fulfer, K. D.; Gupta, T.; Hepler-Smith, E.; Johnston, D. H.; Lang, A. S. I. D.; Larsen, D.; Li, Y.; McEwen, L. R.; Morsch, L. A.; Muzyka, J. L.; Belford, R. E. Teaching Cheminformatics through a Collaborative Intercollegiate Online Chemistry Course (OLCC). *J. Chem. Educ.* **2021**, *98* (2), 416−425.

(18) Sydow, D.; Rodríguez-Guerra, J.; Kimber, T. B.; Schaller, D.; Taylor, C. J.; Chen, Y.; Leja, M.; Misra, S.; Wichmann, M.; Ariamajd, A.; Volkamer, A. TeachOpenCADD 2022: Open Source and FAIR Python Pipelines to Assist in Structural Bioinformatics and Cheminformatics Research. *Nucleic Acids Res.* **2022**, *50* (W1), W753−W760.

(19) Walters, P. *Practical Cheminformatics*; https://practicalcheminformatics.blogspot.com/ (accessed 2024−01−10).

(20) Lafuente, D.; Cohen, B.; Fiorini, G.; García, A. A.; Bringas, M.; Morzan, E.; Onna, D. A Gentle Introduction to Machine Learning for Chemists: An Undergraduate Workshop Using Python Notebooks for Visualization, Data Processing, Analysis, and Modeling. *J. Chem. Educ.* **2021**, *98* (9), 2892−2898.

(21) Menke, J.; Homberg, S.; Koch, O. Introduction to Artificial Intelligence and Deep Learning Using Interactive Electronic Programming Notebooks. *Arch. Pharm.* **2023**, *356* (7), No. e2200628.

(22) Miranda-Salas, J.; Peña-Varas, C.; Valenzuela Martínez, I.; Olmedo, D. A.; Zamora, W. J.; Chávez-Fumagalli, M. A.; Azevedo, D. Q.; Castilho, R. O.; Maltarollo, V. G.; Ramírez, D.; Medina-Franco, J. L. Trends and Challenges in Chemoinformatics Research in Latin America. *Artif. Intell. Life Sci.* **2023**, *3*, No. 100077.

(23) Gaytán-Hernández, D.; Chávez-Hernández, A. L.; López-López, E.; Miranda-Salas, J.; Saldívar-González, F. I.; Medina-Franco, J. L. Art Driven by Visual Representations of Chemical Space. *J. Cheminform.* **2023**, *15* (1), 100.

(24) Saldívar-González, F. I.; Huerta-García, C. S.; Medina-Franco, J. L. Chemoinformatics-Based Enumeration of Chemical Libraries: A Tutorial. *J. Cheminform.* **2020**, *12* (1), 64.

(25) Saldívar-González, F. I.; Navarrete-Vázquez, G.; Medina-Franco, J. L. Design of a Multi-Target Focused Library for Antidiabetic Targets Using a Comprehensive Set of Chemical Transformation Rules. *Front. Pharmacol.* **2023**, *14*, No. 1276444.

(26) Liu, N.; Xu, Z. Using LeDock as a Docking Tool for Computational Drug Design. *IOP Conf. Ser.: Earth Environ. Sci.* **2019**, *218* (1), No. 012143.

(27) Trott, O.; Olson, A. J. AutoDock Vina: Improving the Speed and Accuracy of Docking with a New Scoring Function, Efficient Optimization, and Multithreading. *J. Comput. Chem.* **2010**, *31* (2), 455−461.

(28) Eberhardt, J.; Santos-Martins, D.; Tillack, A. F.; Forli, S. AutoDock Vina 1.2.0: New Docking Methods, Expanded Force Field, and Python Bindings. *J. Chem. Inf. Model.* **2021**, *61* (8), 3891−3898.

(29) 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* (1), 247−258.

(30) Gonzalez-Ponce, K.; Horta Andrade, C.; Hunter, F.; Kirchmair, J.; Martinez-Mayorga, K.; Medina-Franco, J. L.; Rarey, M.; Tropsha, A.; Varnek, A.; Zdrazil, B. School of Cheminformatics in Latin America. *J. Cheminform.* **2023**, *15* (1), 82.

Artículo de alumno de licenciatura

Gaytán-Hernández *et al.*
Journal of Cheminformatics (2023) 15:100
https://doi.org/10.1186/s13321-023-00770-4

Journal of Cheminformatics

EDUCATIONAL ARTICLE **Open Access**

Art driven by visual representations of chemical space

Daniela Gaytán-Hernández[1], Ana L. Chávez-Hernández[1], Edgar López-López[1,2], Jazmín Miranda-Salas[1], Fernanda I. Saldívar-González[1] and José L. Medina-Franco[1*]

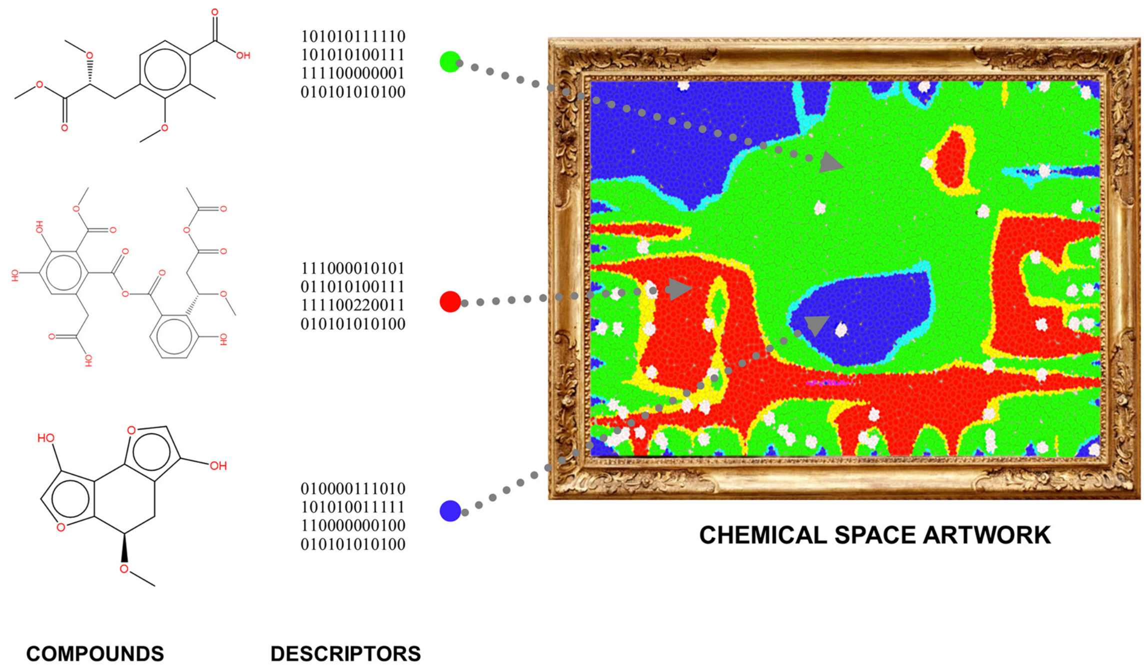
Abstract

Science and art have been connected for centuries. With the development of new computational methods, new scientific disciplines have emerged, such as computational chemistry, and related fields, such as cheminformatics. Chemoinformatics is grounded on the chemical space concept: a multi-descriptor space in which chemical structures are described. In several practical applications, visual representations of the chemical space of compound datasets are low-dimensional plots helpful in identifying patterns. However, the authors propose that the plots can also be used as artistic expressions. This manuscript introduces an approach to merging art with chemoinformatics through visual and artistic representations of chemical space. As case studies, we portray the chemical space of food chemicals and other compounds to generate visually appealing graphs with twofold benefits: sharing chemical knowledge and developing pieces of art driven by chemoinformatics. The art driven by chemical space visualization will help increase the application of chemistry and art and contribute to general education and dissemination of chemoinformatics and chemistry through artistic expressions. All the code and data sets to reproduce the visual representation of the chemical space presented in the manuscript are freely available at https://github.com/DIFACQUIM/Art-Driven-by-Visual-Representations-of-Chemical-Space-. *Scientific contribution*: Chemical space as a concept to create digital art and as a tool to train and introduce students to cheminformatics.

Keywords Artwork, Chemical space, Chemical multiverse, Chemoinformatics, Data visualization, Education, Food chemistry, Foodinformatics, Molecular representation, Open science

*Correspondence:
José L. Medina-Franco
medinajl@unam.mx
Full list of author information is available at the end of the article

Graphical Abstract



Introduction

Art can be considered as the set of activities and products of human beings with aesthetic, ethical, and communication objectives that impact individuals or societies [1]. Its impact may seek to transmit ideas, emotions, needs, concerns, or values [2]. Science can be considered an art tool that makes the materialization of ideas possible and delimits the ideas of artists. What is important about science is not only that it has served to enable the work to be executed. What is fundamental is that it has allowed it to be imagined. Furthermore, scientific knowledge allows for a more profound interpretation of art.

Historically, the relationship between science and art has existed since humans created art. One example is chemistry, a scientific discipline that historically has had a symbiotic relationship with art and has determined its respective evolutions. Among the many interactions of chemistry in art are the development of pigments and spectroscopic techniques, materials for conservation and restoration, to name just a few [3, 4].

The advent of computers gave rise first to computational chemistry and then chemoinformatics. Chemoinformatics, also frequently referred to in the literature as cheminformatics [5] aims to manage and organize information, visualize chemical space, perform data mining, and establish mathematical relationships between chemical structures and properties. While bioinformatics focuses on biologically relevant macromolecules, chemoinformatics is focused on small compounds [6]. As an independent theoretical discipline, chemoinformatics relies on the chemical space concept [7–10]. Understanding the concept of chemical space within and outside chemoinformatics can be complicated. Generally, this concept has been accompanied by various images that seek to represent characteristics that chemists have assigned according to the inherent purposes of their research, leaving aside the aesthetic composition that, in turn, can contribute to deepening and communicating beyond the common sense, which associates thinking to an operation that excludes its connections with the affections, sensitivity, and creation. In Chemoinformatics, chemical space has been defined as a chemical descriptor vector space (cf. Fig. 1A) set by the numerical vector X encoding property or molecular structure aspects as elements of the descriptor vector X [11]. As such, chemoinformatics methods strongly depend on molecular representation and numerical descriptors [12]. There are many descriptors whose selection will depend on the type of molecules studied, for example, organic, inorganic, small molecules, peptides (whose size can differ significantly), natural products, and food chemicals, to name a few. For small molecules (e.g., molecular weight < 1000 Da), it is common to use as descriptors molecular fingerprints [13, 14], whole molecule properties (e.g., properties of pharmaceutical relevance [15, 16]), and sub-structures such as molecular scaffolds [17]. Figure 1A shows a schematic representation of the concept of chemical space, e.g., a chemical space table as a matrix where compounds are the rows and the numerical descriptors are the columns. Graphical and reduction dimension techniques are used to map the usually large multi-dimensional spaces into two or three dimensions that can be plotted and easily visualized.

Since the chemical space of a set of compounds is not unique and will depend on the set of descriptors chosen to describe it, multiple chemical spaces are theoretically possible for the same data set. Continuing this line of thinking, a chemical multiverse was proposed recently and defined as "the group of numerical vectors that describe differently the same set of molecules." An alternative definition of the chemical multiverse is a "group of multiple chemical spaces, each defined by a given set of descriptors—a group of "descriptor universes" [7]. The chemical multiverse concept is represented in Fig. 1B.

Chemical spaces and chemical multiverses are, like many other types of analysis, frequently analyzed through data visualization techniques (Fig. 1). Indeed, data visualization is widely used in science and other areas to effectively summarize and communicate data to produce information and, ultimately, knowledge. Extensive reviews have been published concerning the visualization of chemical spaces [9, 10]. As reviewed, there are multiple methods of visualization, such as principal component analysis (PCA) [18], t-distributed stochastic neighbor embedding (t-SNE) [19], Tree MAP (TMAP) [20], self-organizing map (SOM) [21–23], and the generative topographic mapping (GTM) [24]. Each one will have advantages and disadvantages. As emphasized above, the visualization of a given data set will depend on the type of descriptors used.

The visual representation of chemical spaces can lead to visually appealing figures, particularly if appropriate color schemes are used. The visually attractive settings are used to emphasize patterns in the chemistry data to facilitate visual information extraction. For instance, to highlight grouping or clustering in the chemistry data or to rapidly identify patterns in the structure–property landscapes. At the same time, the visually attractive graphs can be for the chemistry expert and non-expert, a visually appealing graph, or a digital "painting" or work of art. In other words, the graph or digital painting is driven by chemical structures and descriptors. Therefore, the person generating the chemical space representation could be considered a chemical space artist who can communicate not only chemical data and information but even emotions if the chemical structures are associated with a personal, emotional, or another type of feeling the "artist" / author want to communicate through the visualization, e.g., an artistic expression.

In this sense, the concept of chemical space also opens up the possibility of searching for new representations that have to do with the need to configure another image of thought, and think in a novel fashion; it is a creative task and is similar to art.

This manuscript proposes the general notion of generating visual representations of chemical space and chemical multiverses as a means of chemical communication that produces new experiences and, in parallel, artistic expressions. To illustrate the proposal, we generated chemical space visualizations of four flavor categories from an extensive public database of food chemicals, FooDB [25], using different descriptors and molecular fingerprints. We considered four flavor categories, as detailed in the Methods section. The concept would further promote art driven by chemoinformatics and can be expanded to other information-related disciplines, such as bioinformatics. Using different descriptors and visualization methods, we show examples of chemical multiverse visualizations of four flavor categories from FooDB and other chemical compounds.

Methods

Data sets

Herein, we used food chemicals to generate visual representations of the chemical space as artworks. Food and its flavors, colors, textures, and aromas are generally associated with the great pleasures of life; for this reason, they have been a source of inspiration in art world. However, an approximation at the structural level of the molecules has yet to be addressed. Specifically, we used chemical structures from the public database

(See figure on next page.)

Fig. 1 Schematic concept of **A** chemical space and its visual representation in low-dimensions. **B** Schematic representation of a chemical multiverse for a hypothetical data set of n compounds: descriptors of different design (continuous properties, molecular fingerprints, constitutional descriptors, etc.) can lead to alternative chemical spaces for the same data set

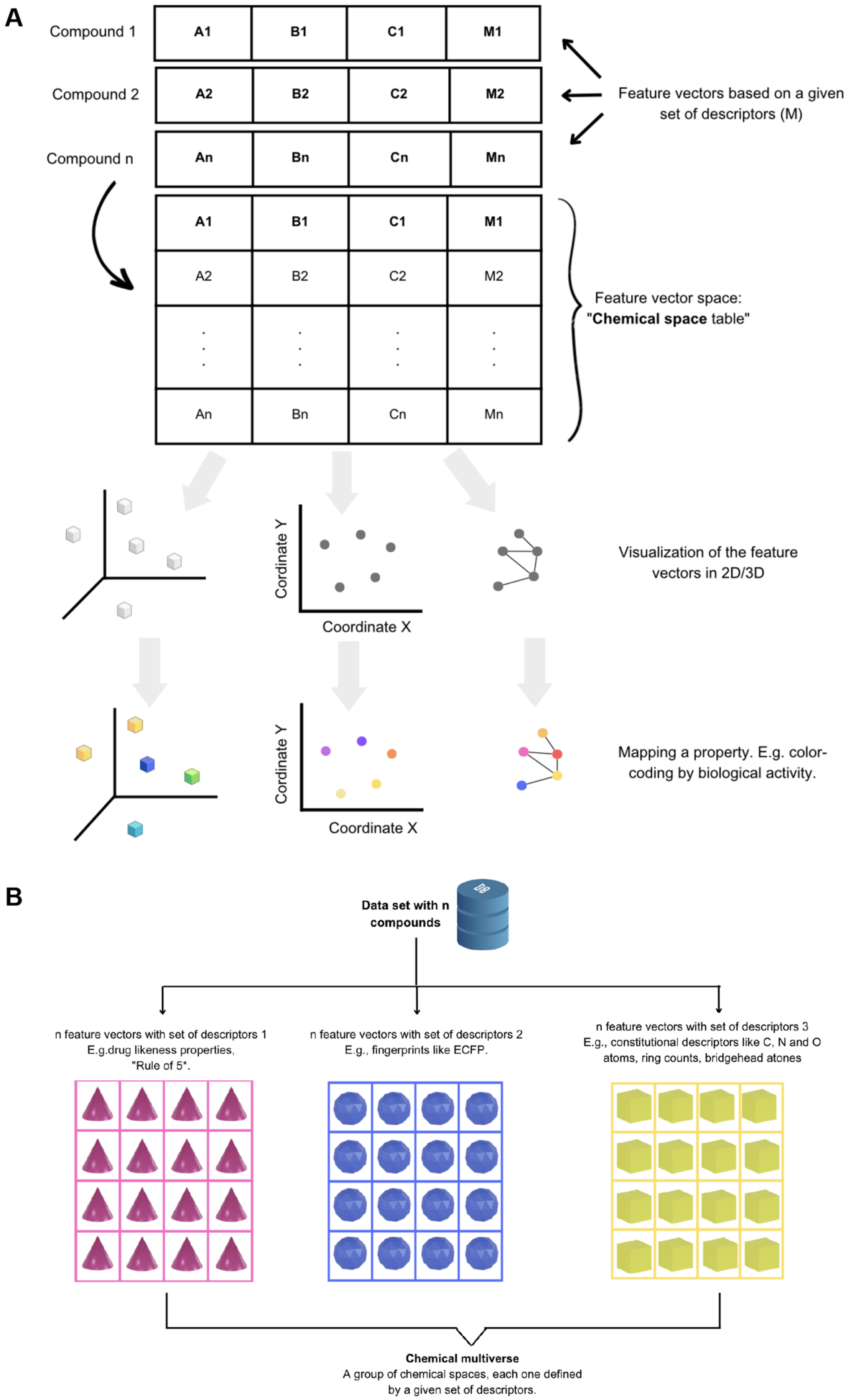

A
Compound 1
A1 B1 C1 M1
Compound 2
A2 B2 C2 M2
Compound n
An Bn Cn Mn
Feature vectors based on a given set of descriptors (M)
A1 B1 C1 M1
A2 B2 C2 M2
An Bn Cn Mn
Feature vector space: "Chemical space table"
Cordinate Y
Coordinate X
Visualization of the feature vectors in 2D/3D
Cordinate Y
Coordinate X
Mapping a property. E.g. color-coding by biological activity.
B
Data set with n compounds
DB
n feature vectors with set of descriptors 1 E.g.drug likeness properties, "Rule of 5".
n feature vectors with set of descriptors 2 E.g., fingerprints like ECFP.
n feature vectors with set of descriptors 3 E.g., constitutional descriptors like C, N and O atoms, ring counts, bridgehead atones
Chemical multiverse
A group of chemical spaces, each one defined by a given set of descriptors.

Fig. 1 (See legend on previous page.)

FooDB [25]. The current version of FooDB contains 70,477 compounds, and after data set standardization (described in detail in Sect. "Data set standardization") has 52,856 molecules. FooDB has information about macronutrients, micronutrients, and food chemicals that give food flavor, color, taste, texture, and aroma to foods. Each chemical item in FooDB contains more than 100 separate data fields providing detailed compositional, biochemical, and physiological information [25]. From FooDB, 4964 natural flavorings derived from food compounds were identified across twenty flavor categories. Figure 2 summarizes the frequency of the seven most populated categories.

From the twenty-seven flavor categories, we defined four new flavor categories: (1) ground flavors, (2) wine-tasting, (3) contrast between fatty and spicy, and (4) natural remedies. Additional file 1: Table S1 shows the number of compounds in each of the four categories considered in this work. Flavors of the ground/flavor similar to herbaceous are earthy, herbaceous, and green flavors. Wine tasting is composed of fruity and floral flavors. The contrast between fatty and spicy is composed of fatty and spicy flavors. Medicinal comprises balsamic, chemical, and medicinal, which are characteristic flavors found in ointments, alcohol, and syrups. Additional file 1: Fig. S1 shows the overlapping compounds between the selected flavor categories.

Data set standardization

Compounds in FooDB, encoded as SMILES strings [12], were standardized using the open-source cheminformatics toolkit RDKit [26] and Standardizer, LargestFragmentChoser, Uncharger, Reionizer y TautomerCanonicalizer functions implemented in MolVS [27]. Compounds with valence errors or any chemical element other than H, B, C, N, O, F, Si, P, S, Cl, Se, Br, and I were removed. Stereochemistry information, when available, was retained. Compounds with multiple components were split, and the largest component was retained. The remaining compounds were neutralized and reionized to generate the corresponding canonical tautomer.

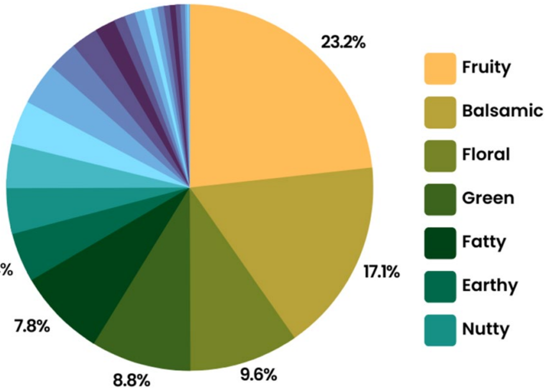


Fig. 2 The seven most frequent flavor categories identified in FooDB

Molecular descriptors

For each molecule, physicochemical properties and molecular fingerprints were calculated as descriptors using Python language and RDKit. The whole molecule descriptors computed were hydrogen bond donors (HBD), hydrogen bond acceptors (HBA), topological polar surface area (TPSA), number of rotatable bonds (RB), molecular weight (MW), and partition coefficient octanol/water (LogP). Molecular fingerprints computed were Molecular Access System (MACCS) Keys (166-bits) [13], extended connectivity fingerprint (ECFP) [14] of 1024-bits with diameter 4 (ECFP4). Of note, virtually any other descriptors can be used, as further commented in the Sect. "Discussion".

Visualization methods

In this study, we used three well-known dimensionality reduction methods: t-SNE, PCA, and TMAPs, although additional visualization methods can be used. Briefly, t-SNE generates plots that organize compounds. Similar compounds form clusters and dissimilar compounds are distant from each other. PCA is a linear dimensionality reduction technique that transforms data with many dimensions (i.e., descriptors) into a lower dimensional space and keeps the different relationships between the data points as much as possible [18]. PCA was generated from six whole molecule descriptors (MW, HB, HBA, SlogP, TPSA, and RB). TMAPs allow visualization of many chemical compounds through the distance between clusters and the detailed structure of these through branches and sub-branches. Local sensitive hashing allows each compound to be grouped hierarchically according to common substructures using molecular fingerprints. In this work, we use MACCS keys (166-bits) [13] fingerprints. Then, each chemical compound was encoded using the MinHash algorithm. The number of nearest neighbors, k = 50, and the factor used by the augmented query algorithm, kc = 10, were used to generate the TMAPs [20].

Results

Figures 3, 4, 5, 6 show examples of so-called "Art Galleries" composed by visualization of the chemical space of different food chemical categories. The visual representations of chemical space were generated with t-SNE (Figs. 3 and 4), PCA (Fig. 5), and TMAPs

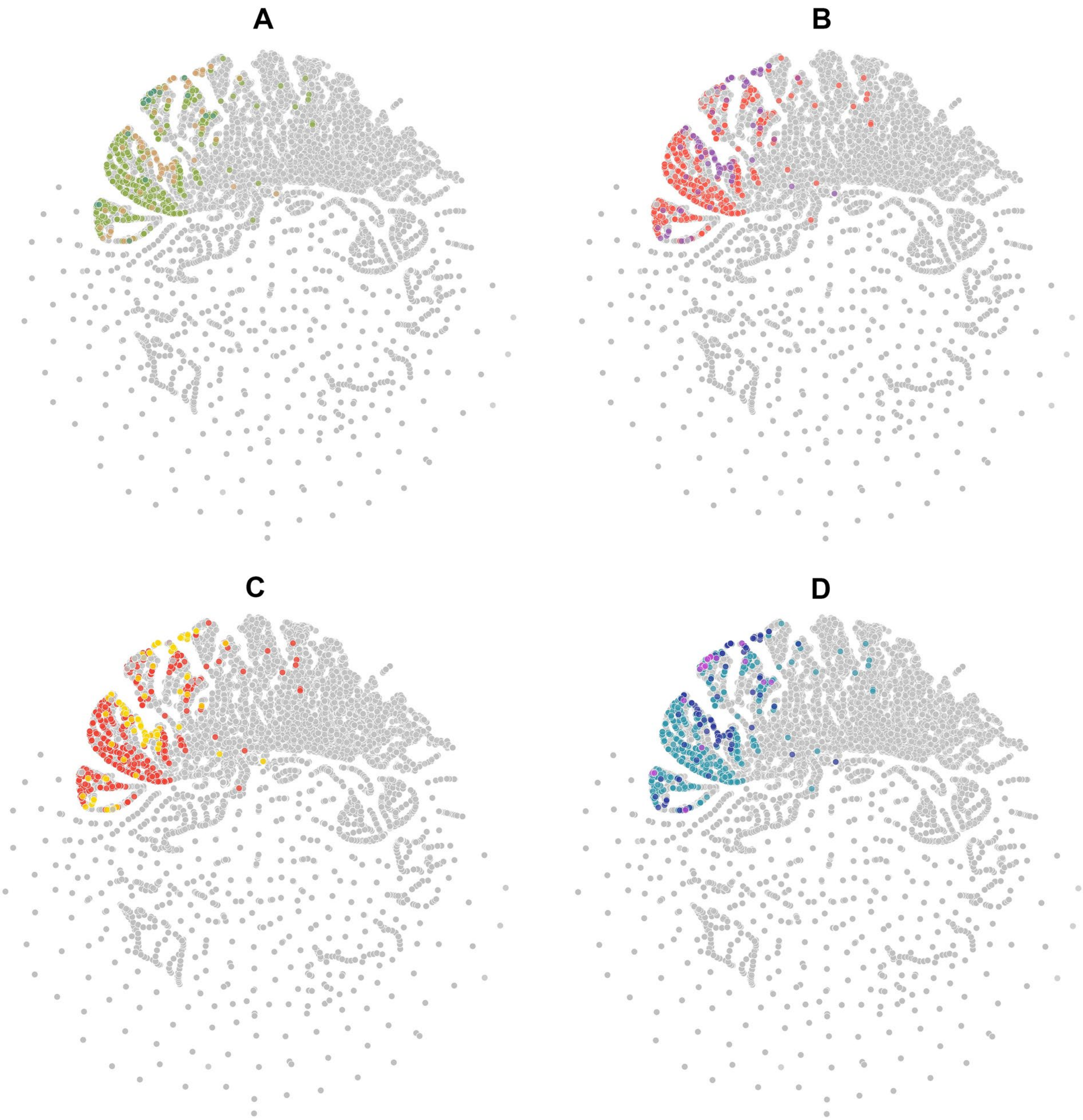


Fig. 3 Four flavor categories and full FooDB. The flavor categories are **A** Ground flavors (655 compounds), **B** Wine-tasting (1024 compounds), **C** Contrast between fatty and spicy (430 compounds), and **D** Natural remedies (762 compounds)

(Fig. 6). Below each image (i.e., "digital paintings") is presented basic information of the "technique" (visualization method, allusive to the techniques used in paintings), descriptors, and chemicals (that would be meaningful information for a chemistry-oriented person to understand the data presented). Each visual representation of the chemical space or Artwork includes a "Title" that is reminiscent of the name of the piece of art or digital painting.

Discussion

Chemoinformatics has been broadly used in drug discovery. Still, it has many more applications in chemistry, with increasing applications in food chemistry, as evidenced by the emergence of the research areas of food chemical informatics or food informatics [28, 29]. There are others, such as natural products [30, 31], polymers, and materials, to name a few [6]. Herein, we propose expanding the realm of chemoinformatics´ applications through the visual

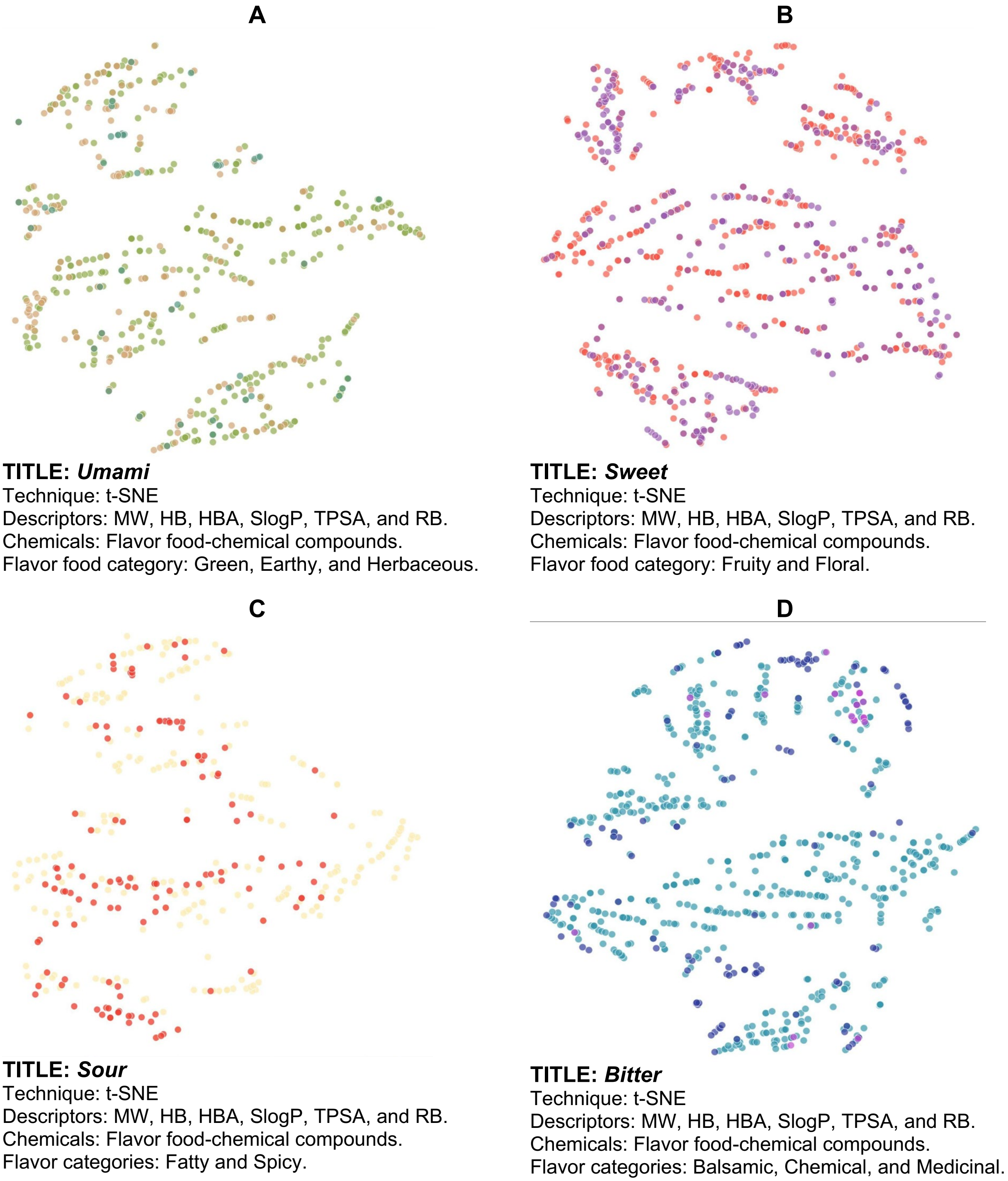


Fig. 4 Four flavor categories: **A** Ground flavors (655 compounds), **B** Wine-tasting (1024 compounds), **C** Contrast between fatty and spicy (430 compounds), and **D** Natural remedies (762 compounds)

representation of the chemical space of compound data sets—herein illustrated with food chemicals—to yield exemplary “art pieces.” The connection or synergy between chemoinformatics and art has a strong potential to bring together at least two sectors of the population that might be otherwise disconnected. From an educational point of view, which is a central need in

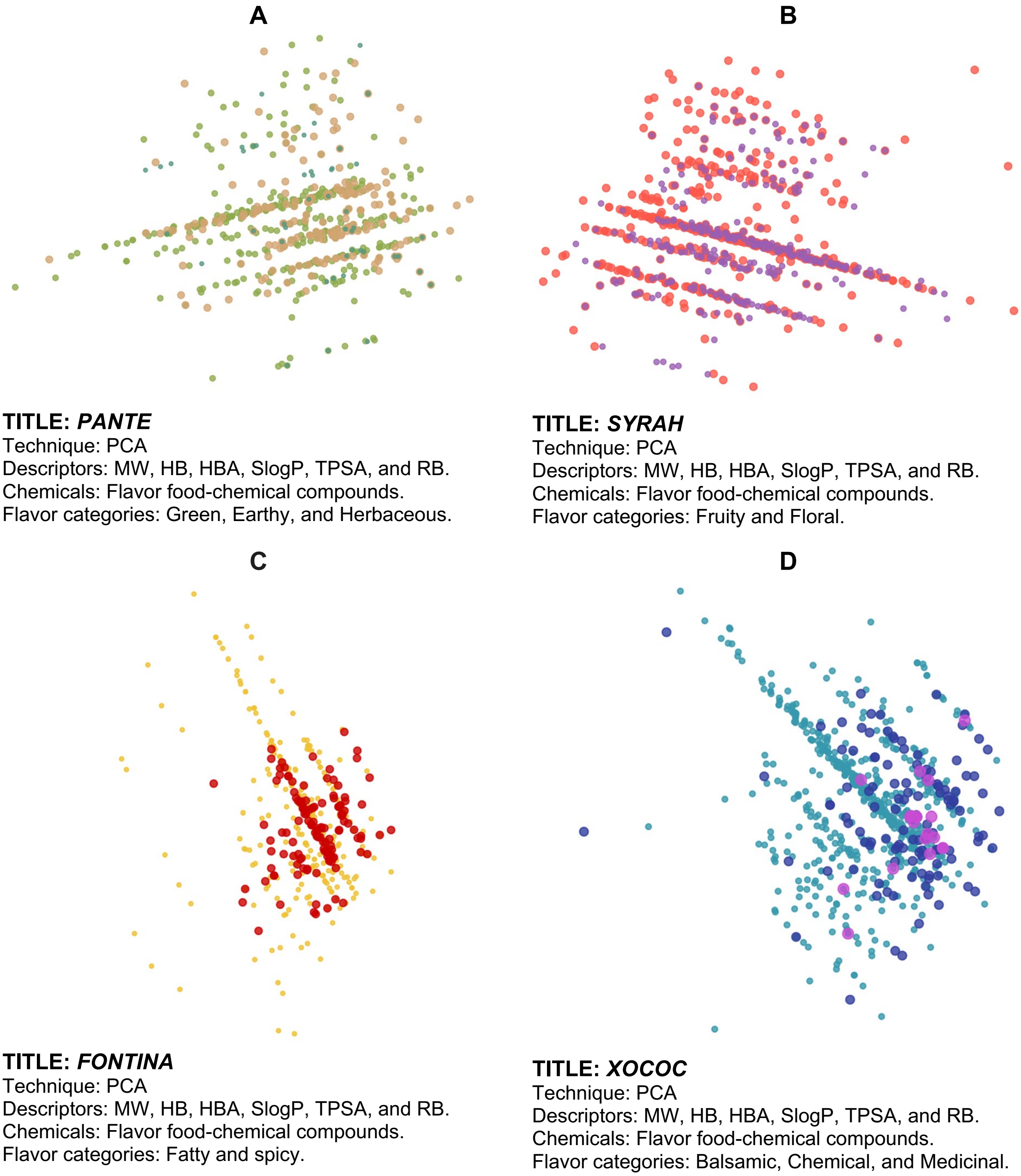


Fig. 5 Four flavor categories: **A** Ground flavors (655 compounds), **B** Wine-tasting (1024 compounds), **C** Contrast between fatty and spicy (430 compounds), and **D** Natural remedies (762 compounds)

chemoinformatics—the synergy might attract young students and kids to chemistry through art.

The subdiscipline of food informatics was proposed in 2014 as a specific application of chemoinformatics to food chemistry [28]. Since then, numerous applications of chemoinformatics to different aspects of food chemistry have been published, including analysis of the chemical space of food chemicals to characterize the structural diversity [32]. In Sect. "Results" we showed examples of visual representations of the chemical space of food chemicals as an artistic expression and scientific dissemination through art. There are many

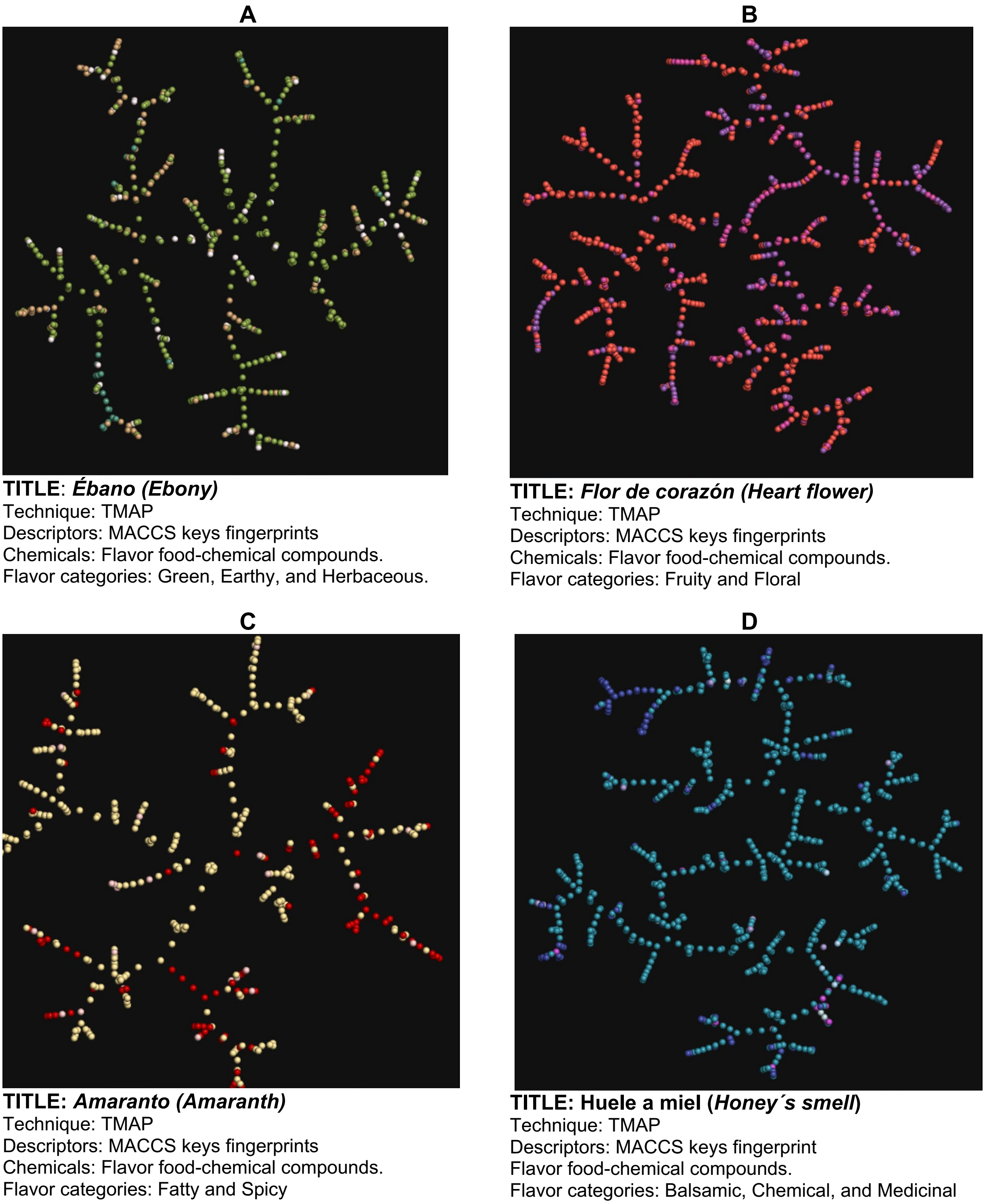


Fig. 6 Four flavor categories: **A** Ground flavors (655 compounds), **B** wine-tasting (1024 compounds), **C** Contrast between fatty and spicy (430 compounds), and **D** Natural remedies (762 compounds)

Table 1 Exemplary potential paintings based on the visualization of the chemical space of compound data sets

Data set	Artistic meaning	Artwork name
Random compounds	Aleatory molecules represent the vastness of our universe and daily life. We are in contact with many chemicals every time, but we don't look at their complexity and intrinsic disorder in our universe and daily life	"Chaos"
Diverse data set	The diversity offers many colors, flavors, tastes, and experiences. In nature, diversity (in all senses) is a constant feature	"Diversity"
Marine natural products	We don't understand the sea; It has life, death, color, and darkness. It's constantly changing	"The Ocean" "Immensity"
Drugs approved for the treatment of HIV	Everything happens in a positive HIV human; Fear, memories, happiness, and normality. The drugs help… but are not a complete answer	"Living with AIDS"
Hormones—neurotransmitters	Love = hormones + neurotransmitters + special persons	"The chemistry of love"
Chemicals associated with depression	Depression = hormones + neurotransmitters—purpose	"Darkness"
Food chemicals	The great pleasures of life are often accompanied by flavors, colors, textures, and aromas	"Bellyful" "Flavor trip"
ZINC database vs. drug-like compounds	We know a lot about our nature and composition, but we don't know much more. Our knowledge is a mere stain on an entire canvas that we do not yet understand	"Our knowledge"

possibilities to expand the genesis of the proposed "art-cheminformatics," as further elaborated in Sect. "Conclusions and outlook".

Exemplary art-related chemical spaces and multiverses

The examples of visual representation of chemical space as artistic representations presented in Sect. "Results"

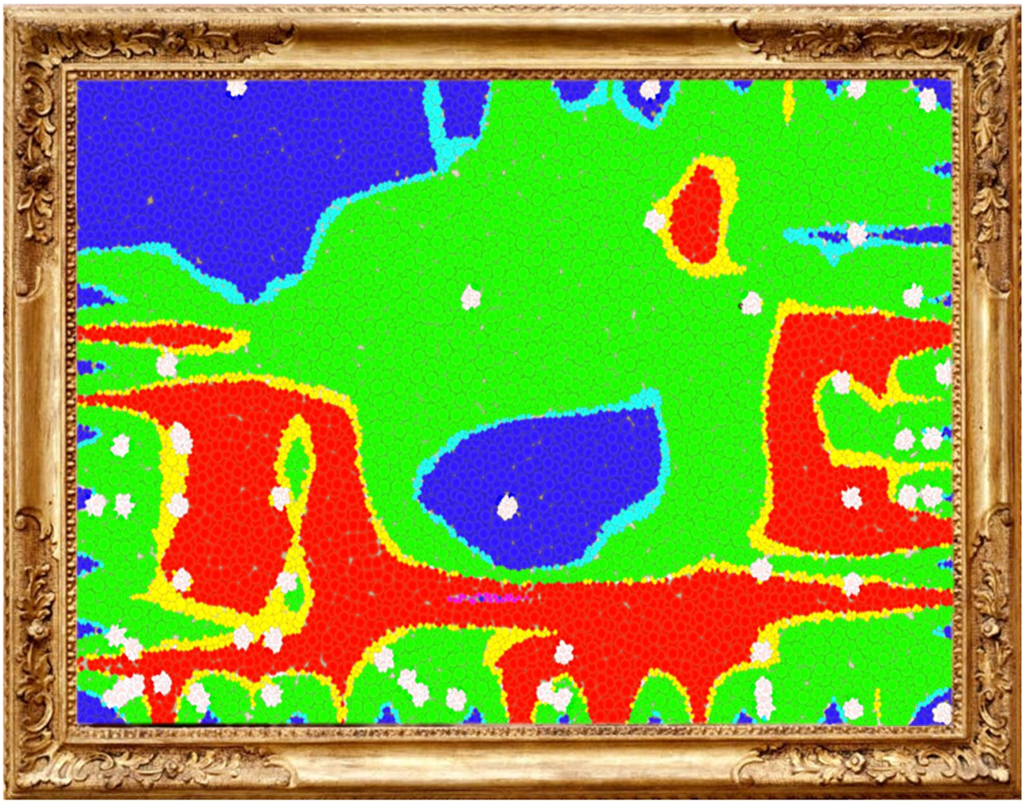
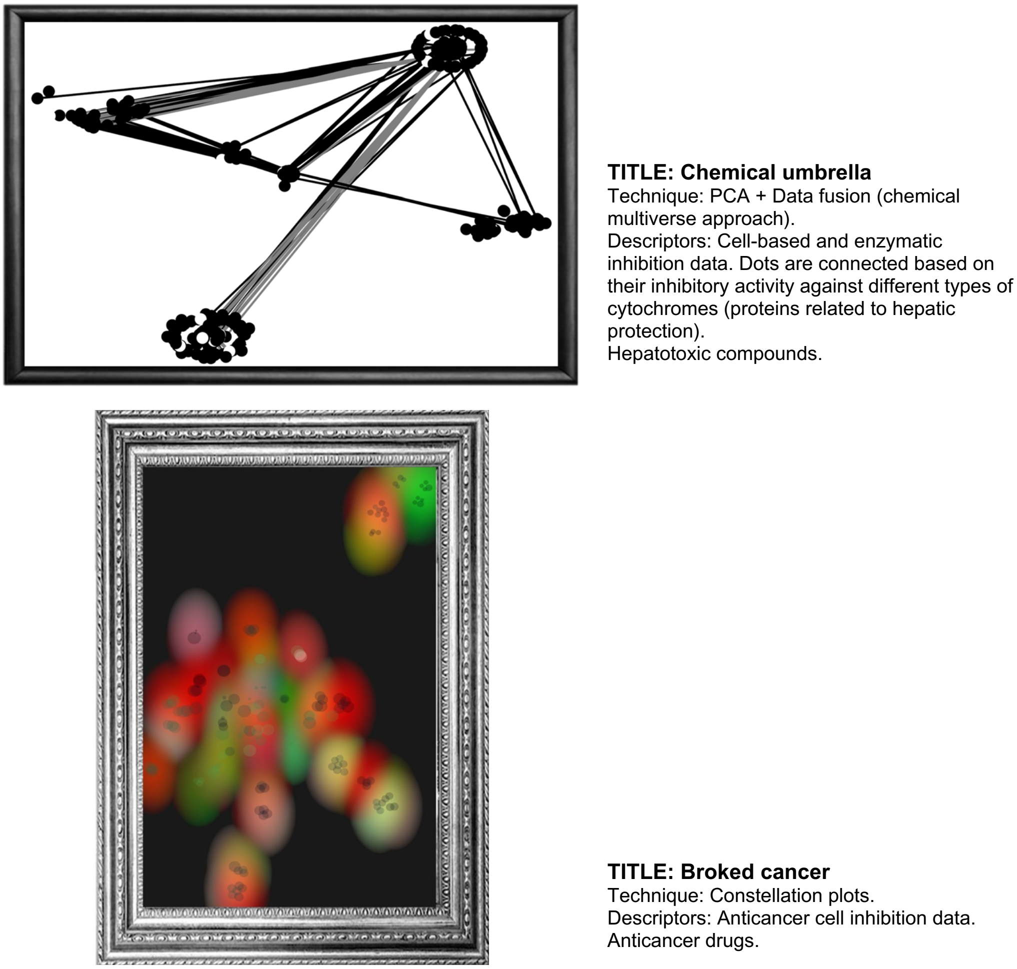


Fig. 7 Chemical space art example. Title: "Wise nature"; Autor: Edgar López-López; Technique: SOM—using DataWarrior software [33, 34]; Dataset: Random natural products (1000 compounds); Descriptors: predicted mutagenic, tumorogenic, Reproductive effective, and Irritant; Technical description: Each white point is a natural product, the regions colored in red represent the chemical space with a high predicted probability of containing compounds witch side effects, the opposite for the blue color; Artistic interpretation: The "nature" is not always healthy, in nature, there has always been a duality between what fills us with life and what takes it away



TITLE: Chemical umbrella
Technique: PCA + Data fusion (chemical multiverse approach).
Descriptors: Cell-based and enzymatic inhibition data. Dots are connected based on their inhibitory activity against different types of cytochromes (proteins related to hepatic protection).
Hepatotoxic compounds.

TITLE: Broked cancer
Technique: Constellation plots.
Descriptors: Anticancer cell inhibition data.
Anticancer drugs.

Fig. 8 Chemical space art examples. Chemical artworks were generated with public data [35–37]

are focused on food chemicals and molecular descriptors suitable to represent such chemical compounds. Also, examples of visualization methods used in the previous section are t-SNE, PCA, and TMAPs. However, as commented in the Introduction, the number of established visualization techniques, molecular descriptors, and, perhaps most importantly, the number of chemical structures are immense. Therefore, there are thousands or millions of ways to generate chemical space-driven works of art. To glimpse the artistic possibilities, Table 1 summarizes examples of the cheminformatics-driven visualization of chemical space and multiverses. The table summarizes examples of compound data sets with chemicals of different types that could be used to represent their vastness, complexity, diversity, and chaotic intrinsic features from an artistic perspective. Many more compound data sets and multiple combinations of descriptors and visualization techniques could be used. However, as with any other artistic vehicle, the real importance of any type of art is its capacity to tell histories or convey a message that sometimes is hidden.

To illustrate further the potential of generating artistic representations through visualization of chemical space, Fig. 7 shows an example of chemical space artwork from a random natural products dataset, decoding by their side effects descriptors (e.g., mutagenesis, tumorogenesis, and negative reproductive effects, etc.). Their color palette, from red to blue, represents the probability of each natural product generating side effects. The “canvas” was “painted” with a dotted technique, reflecting another possible set of textures that can be developed with this technique. Like in Fig. 7, we intrinsically know that "nature" is not always healthy and that within us, there is a delicate balance that is very easy to break.

Figure 8 shows additional examples of chemical space artwork that combine different reduction data

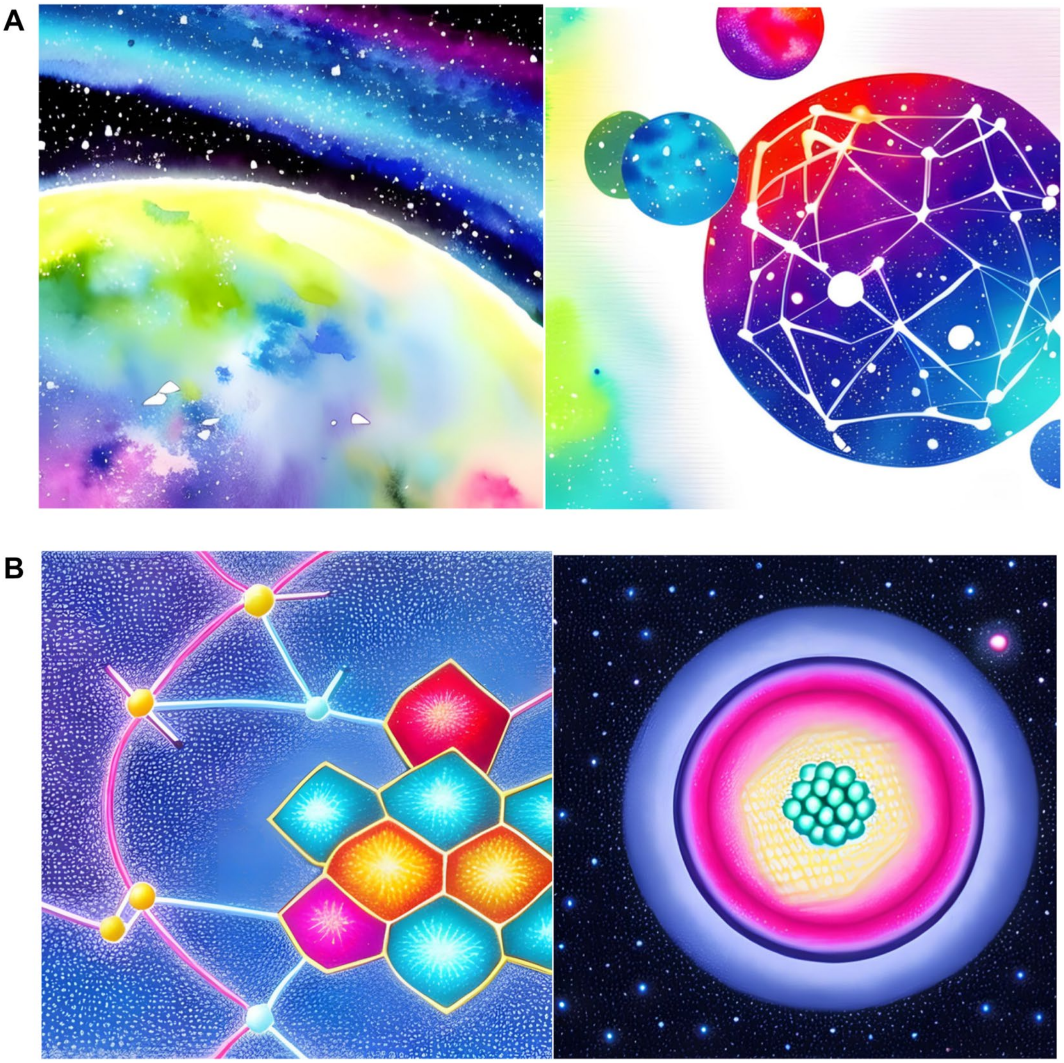


Fig. 9 Example(s) of artificial intelligence-driven art with the free application Canva (https://www.canva.com/) using the keyword chemical space and **A** Watercolor and **B** color pencil

Table 2 Representative developments of combining art with chemoinformatics through artistic visualizations of chemical space

Development	Putative outcome or application
Continue developing a digital collection focused on the artistic representation of the chemical space	*The Chemical Space Art Museum*
Generate automated workflows using open software or informatic tools to improve the accessibility of this kind of art to people with different academic/artistic backgrounds	*ChemArt Generators*
Establish a free, open-access, and permanent repository of art pieces. This encourages open science and open art. The scientific and artistic community could support the repository	*ChemART Gallery*. An example is at https://www.difacquim.com/chemical-art-gallery/
Set up a sustained educational or cultural program as a continued open and permanent exposition	*Art Driven by Chemical Space Visualization* program

methods and descriptors to generate an artistic visual representation of the chemical data. We encourage the readers to reflect and find other artistic interpretations that these figures could have. The examples of chemical space visualization as work art have been included in a Chemical Space Art Gallery freely available at https://www.difacquim.com/chemical-art-gallery/

Artificial intelligence and digital art

Artificial intelligence (AI) is used to generate artistic representations [38, 39]. Although it is not the central point of this manuscript, Fig. 9 illustrates images generated with free resources using keywords associated with "chemical space." Specifically, the figure shows an example of a chemical multiverse/chemical space driven by an AI-web server training on words. Although the images are attractive, a striking difference with the chemical space artworks presented in previous sections (Figs. 3, 4, 5, 6, 7, 8) is that the images in Fig. 9 are based on keyword training. The former are derived directly from chemical structures encoded with molecular descriptors. Another important aspect is a greater understanding and human intervention in the former representations, something questionable in AI-guided pictures.

Conclusions and outlook

Science and art have long been intimately related. A typical example is summarized by the phrase, "Drug discovery is as much an art as it is a science." Certainly, chemistry is substantially used in art, such as in art restoration and preservation. However, an emerging trend exists to apply chemistry and its concepts to generate artwork. Herein, we discuss an approach to combining art with chemoinformatics through the visual representations of chemical space. We presented a few examples of chemical space artworks that can be "digital paintings." The author of the low-dimensional graphs can use the plots with dual general purposes: communicate data and generate chemical information (as generally done with the visualizations of chemical space) and convey an emotional or personal meaning to the graph (driven by chemistry and informatics principles).

We also conclude that chemical space-driven works of art can be tools to promote science in general and chemistry in particular for the broad audience. Thus, chemistry informatic-driven artistic expressions can be an approach to disseminating science. Such an approach aligns with the graphical abstracts frequently used in peer-reviewed journals. The "chemical art" could be useful to represent complex data by using an artistic and attractive perspective. The person generating the chemical space representation could be considered a "chemical space artist."

We envision several further developments and areas of opportunity for art driven by visual representations of chemical space. Table 2 summarizes ongoing chemical arts projects, from the generation of "easy to use" tools, the first chemical art gallery, and the implementation of this artistic mode to introduce the new generation of chemoinformaticians to the chemical space concept. In parallel, AI methods will continue expanding and exploring the chemical space, offering new types of molecules and descriptors that could be used to increase the possibilities of representing chemical space from an artist's perspective.

Abbreviations

AI	Artificial intelligence
ECPF	Extended connectivity fingerprint
HBD	Hydrogen bond donors
HBA	Hydrogen bond acceptors
GTM	Generative topographic mapping
LogP	Partition coefficient octanol/water
MACCS	Molecular ACCes System
MW	Molecular weight
PCA	Principal component analysis
TMAP	TreeMap
t-SNE	T-Distributed stochastic neighbor embedding
TPSA	Topological polar surface area
SOM	Self-organizing map

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13321-023-00770-4.

Additional file 1: Table S1. The number of flavor compounds, flavor notes, and flavor categories. **Figure S1.** Unique and overlapping structures of four flavor categories from FooDB. All the code and data sets to reproduce the visual representation of the chemical space presented in the manuscript are freely available at https://github.com/DIFACQUIM/Art-Driven-by-Visual-Representations-of-Chemical-Space-.

Acknowledgements

We are grateful for the rich and useful discussions with Dra. Karina Martínez-Mayorga. D.G.-H. thanks the support from CONAHCYT. A.L.-C.-H., E.L.-L., and F.I.S.-G thank CONAHCYT, Mexico, for the Ph.D. scholarships number 847870, 894234, and 848061, respectively.

Author contributions

D. G-H and E. L-L. generated visual representations of the chemical space and interpreted the data. AL C-H supervised the work of D. G-H, participated in the conceptualization of the study, and helped to draft the first version of the draft. FI. S-G and J. M-S contributed to the discussion of the concepts. JL M-F conceptualized the project, supervised the project and drafted the first version of the manuscript. All authors contributed to edit and prepare the final version of the manuscript.

Funding

No funding was received to perform this research.

Availability of data and materials

All data related to this manuscript can be accessed in the Supplementary material.

Declarations

Ethics approval and consent to participate

Not applicable.

Competing interests

Authors declare that have no competing interests.

Author details

[1] DIFACQUIM Research Group, Department of Pharmacy, School of Chemistry, Universidad Nacional Autónoma de México, Avenida Universidad 3000,

04510 Mexico City, Mexico. [2]Department of Chemistry and Graduate Program in Pharmacology, Center for Research and Advanced Studies of the National Polytechnic Institute, 07000 Mexico City, Mexico.

Received: 13 July 2023 Accepted: 13 October 2023
Published online: 21 October 2023

References

1. La Galván-Madrid JL (2011) Química y el Arte: ¿Cómo mantener el vínculo? Educ Quím 22(3):207–211. https://doi.org/10.1016/S0187-893X(18)30136-8
2. Bello DG (2023) La química de lo bello, 2nd edn. Ediciones Paidós, Barcelona
3. Orna MV (2001) Chemistry, color, and art. J Chem Educ 78(10):1305. https://doi.org/10.1021/ed078p1305
4. Kafetzopoulos C, Spyrellis N, Lymperopoulou-Karaliota A (2006) The chemistry of art and the art of chemistry. J Chem Educ 83(10):1484. https://doi.org/10.1021/ed083p1484
5. Miranda-Salas J, Peña-Varas C, Valenzuela Martínez I, Olmedo DA, Zamora WJ, Chávez-Fumagalli MA, Azevedo DQ, Castilho RO, Maltarollo VG, Ramírez D, Medina-Franco JL (2023) Trends and challenges in chemoinformatics research in Latin America. Artif Intell Life Sci 3(1):100077. https://doi.org/10.1016/j.ailsci.2023.100077
6. López-López E, Bajorath J, Medina-Franco JL (2020) Informatics for chemistry, biology, and biomedical sciences. J Chem Inf Model 61(1):26–35. https://doi.org/10.1021/acs.jcim.0c01301
7. Medina-Franco JL, Chávez-Hernández AL, López-López E, Saldívar-González FI (2022) Chemical multiverse: an expanded view of chemical space. Mol Inf 41(11):2200116. https://doi.org/10.1002/minf.202200116
8. Medina-Franco JL, Sánchez-Cruz N, López-López E, Díaz-Eufracio BI (2022) Progress on open chemoinformatic tools for expanding and exploring the chemical space. J Comput Aided Mol Des 36(5):341–354. https://doi.org/10.1007/s10822-021-00399-1
9. Osolodkin DI, Radchenko EV, Orlov AA, Voronkov AE, Palyulin VA, Zefirov NS (2015) Progress in visual representations of chemical space. Expert Opin Drug Discov 10(9):959–973. https://doi.org/10.1517/17460441.2015.1060216
10. Medina-Franco J, Martinez-Mayorga K, Giulianotti M, Houghten R, Pinilla C (2008) Visualization of the chemical space in drug discovery. Curr Comput Aided Drug Des 4(4):322–333. https://doi.org/10.2174/157340908786786010
11. Saldívar-González FI, Medina-Franco JL (2022) Approaches for enhancing the analysis of chemical space for drug discovery. Expert Opin Drug Discov 17(7):789–798. https://doi.org/10.1080/17460441.2022.2084608
12. Weininger D (1988) SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. J Chem Inf Comput Sci 28(1):31–36. https://doi.org/10.1021/ci00057a005
13. Durant JL, Leland BA, Henry DR, Nourse JG (2002) Reoptimization of MDL keys for use in drug discovery. J Chem Inf Comput Sci 42(6):1273–1280. https://doi.org/10.1021/ci010132r
14. Rogers D, Hahn M (2010) Extended-connectivity fingerprints. J Chem Inf Model 50(5):742–754. https://doi.org/10.1021/ci100050t
15. Lipinski CA, Lombardo F, Dominy BW, Feeney PJ (2001) Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Adv Drug Deliv Rev 46(1–3):3–26. https://doi.org/10.1016/s0169-409x(00)00129-0
16. Veber DF, Johnson SR, Cheng H-Y, Smith BR, Ward KW, Kopple KD (2002) Molecular properties that influence the oral bioavailability of drug candidates. J Med Chem 45(12):2615–2623. https://doi.org/10.1021/jm020017n
17. Singh N, Guha R, Giulianotti MA, Pinilla C, Houghten RA, Medina-Franco JL (2009) Chemoinformatic analysis of combinatorial libraries, drugs, natural products, and molecular libraries small molecule repository. J Chem Inf Model 49(4):1010–1024. https://doi.org/10.1021/ci800426u
18. Greener JG, Kandathil SM, Moffat L, Jones DT (2021) A guide to machine learning for biologists. Nat Rev Mol Cell Biol 23(1):40–55. https://doi.org/10.1038/s41580-021-00407-0
19. van der Maaten L, Hinton G (2023) Visualizing data using t-SNE. https://www.jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf?fbcl. Accessed 1 Jun 2023
20. Probst D, Reymond J-L (2020) Visualization of very large high-dimensional data sets as minimum spanning trees. J Cheminf 12(1):1–13. https://doi.org/10.1186/s13321-020-0416-x
21. Kohonen T (2001) Self-organizing maps. Springer, Berlin Heidelberg, pp 105–176
22. Schneider P, Tanrikulu Y, Schneider G (2009) Self-organizing maps in drug discovery: compound library design, Scaffold-Hopping. Repurpos Curr Med Chem 16(3):258–266. https://doi.org/10.2174/092986709787002655
23. Virshup AM, Contreras-García J, Wipf P, Yang W, Beratan DN (2013) Stochastic voyages into uncharted chemical space produce a representative library of all possible drug-like compounds. J Am Chem Soc 135(19):7296–7303. https://doi.org/10.1021/ja401184g
24. Bishop CM, Svensén M, Williams CKI (1998) Developments of the generative topographic mapping. Neurocomputing 21(1):203–224. https://doi.org/10.1016/S0925-2312(98)00043-5
25. FooDB https://foodb.ca/. Accessed 20 Apr 2023
26. RDKit https://www.rdkit.org. Accessed 8 Jan 2022
27. MolVS https://molvs.readthedocs.io/en/latest/. Accessed 8 Jan 2022
28. Martinez-Mayorga K, Medina-Franco JL, Eds (2014)Foodinformatics: applications of chemical information to food chemistry. Springer International Publishing: Cham
29. Peña-Castillo A, Méndez-Lucio O, Owen JR, Martínez-Mayorga K, Medina-Franco JL (2018) Chemoinformatics in food science. In Applied chemoinformatics, Wiley-VCH Verlag GmbH & Co. KGaA: Weinheim, Germany, pp 501–525
30. Kirchmair J (2020) Molecular informatics in natural products research. Mol Inf 39(11):2000206. https://doi.org/10.1002/minf.202000206
31. Medina-Franco JL, Saldívar-González FI (2020) Cheminformatics to characterize pharmacologically active natural products. Biomolecules 10(11):1566. https://doi.org/10.3390/biom10111566
32. Naveja JJ, Rico-Hidalgo MP, Medina-Franco JL (2018) Analysis of a large food chemical database: chemical space, diversity, and complexity. F1000Res. https://doi.org/10.12688/f1000research.15440.2
33. Sander T, Freyss J, von Korff M, Rufener C (2015) DataWarrior: an open-source program for chemistry aware data visualization and analysis. J Chem Inf Model 55(2):460–473. https://doi.org/10.1021/ci500588j
34. López-López E, Naveja JJ, Medina-Franco JL (2019) DataWarrior: an evaluation of the open-source drug discovery tool. Expert Opin Drug Discov 14(4):335–341. https://doi.org/10.1080/17460441.2019.1581170
35. López-López E, Medina-Franco JL (2023) Towards decoding hepatotoxicity of approved drugs through navigation of multiverse and consensus chemical spaces. Biomolecules 13(1):176. https://doi.org/10.3390/biom13010176
36. Medina-Franco JL, Naveja JJ, López-López E (2019) Reaching for the bright StARs in chemical space. Drug Discov Today 24(11):2162–2169. https://doi.org/10.1016/j.drudis.2019.09.013
37. López-López E, Cerda-García-Rojas CM, Medina-Franco JL (2021) Tubulin inhibitors: a chemoinformatic analysis using cell-based data. Molecules 26(9):2483. https://doi.org/10.3390/molecules26092483
38. DALL·E 2 https://openai.com/dall-e-2/. Accessed 20 Jun 2023
39. Ramesh A, Dhariwal P, Nichol A, Chu C, Chen M (2022) Hierarchical text-conditional image generation with CLIP latents. arXiv:2204.06125

Publisher's Note

Congresos

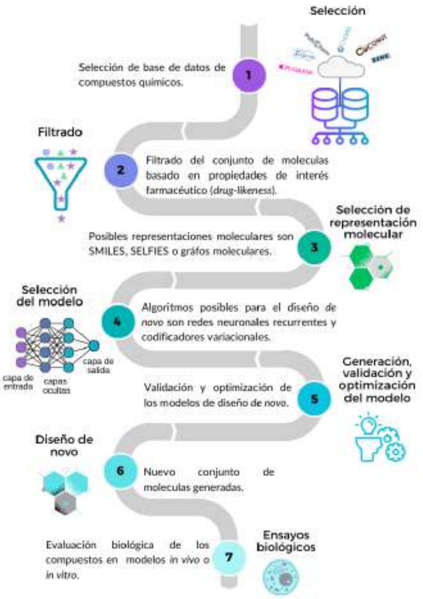
Subconjuntos de bases de datos de productos naturales: Generación y caracterización

Ana L. Chávez-Hernández*, José L. Medina-Franco*
DIFACQUIM, Departamento de Farmacia, Facultad de Química, Universidad Nacional Autónoma de México
Avenida Universidad 3000, Ciudad de México 04510, México. ***Correo:** anachavez3026@gmail.com

Palabras clave: Inteligencia artificial, diseño *de novo*, espacio químico, productos naturales.

Resumen

Los productos naturales son estructuras químicas importantes en el diseño *de novo* de fármacos debido a sus características estructurales como la fracción de átomos de carbono con hibridación sp3, centros quirales y diversidad de grupos funcionales. El diseño *de novo* basado en algoritmos de inteligencia artificial requiere del manejo de bases de datos muy grandes y uso de supercómputo [1]. Sin embargo, muchos grupos de investigación difícilmente pueden acceder a recursos de supercómputo.

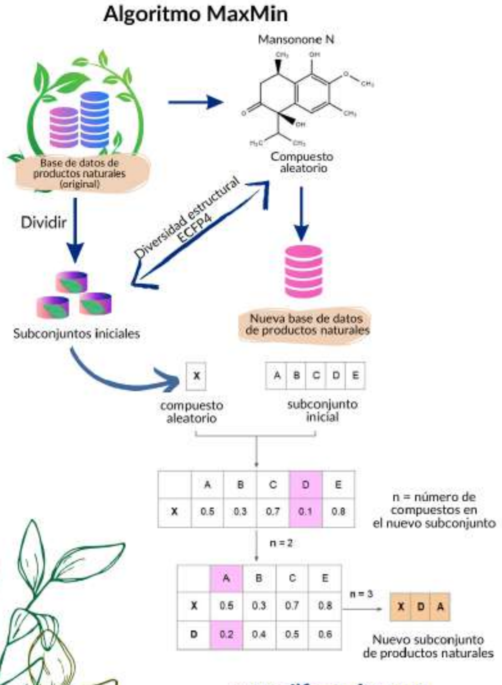

Selección
1
Selección de base de datos de compuestos químicos.
Filtrado
2
Filtrado del conjunto de moleculas basado en propiedades de interés farmacéutico (drug-likeness).
Selección de representación molecular
3
Posibles representaciones moleculares son SMILES, SELFIES o gráfos moleculares.
Selección del modelo
capa de entrada
capas ocultas
capa de salida
4
Algoritmos posibles para el diseño de novo son redes neuronales recurrentes y codificadores variacionales.
Generación, validación y optimización del modelo
5
Validación y optimización de los modelos de diseño de novo.
Diseño de novo
6
Nuevo conjunto de moleculas generadas.
Ensayos biológicos
7
Evaluación biológica de los compuestos en modelos in vivo o in vitro.

Objetivo

Generar y caracterizar subconjuntos de productos naturales de compuestos más diversos del *Universal Natural Product Database* (UNPD) [2] utilizando el algoritmo MaxMin [3].

Metodología

153,375 compuestos del UNPD [2] codificados como SMILES fueron estandarizados utilizando el lenguaje de programación Python. Se eliminaron los compuestos con errores de valencia y átomos diferentes a los elementos H, B, C, N, O, F, Si, P, S, Cl, Se, Br y I. Los compuestos con múltiples componentes fueron separados y el fragmento más grande fue retenido. De los compuestos remanentes, se conservó la información estereoquímica, se neutralizaron las cargas, y se eliminaron los compuestos repetidos. Después, se generaron subconjuntos de compuestos usando el algoritmo MaxMin [3]. Se dividió la base de datos de UNPD en subconjuntos, se seleccionó un compuesto aleatorio (X). Se calculó la similitud molecular entre X y los compuestos remanentes usando el coeficiente de Tanimoto y la huella digital molecular de conectividad extendida de diámetro 4 (ECFP4). Los compuestos con la similitud molecular más pequeña fueron seleccionados hasta obtener el número de compuestos deseados.


Algoritmo MaxMin
Mansonone N
Base de datos de productos naturales (original)
Compuesto aleatorio
Dividir
Diversidad estructural ECFP4
Subconjuntos iniciales
Nueva base de datos de productos naturales
X
compuesto aleatorio
A B C D E
subconjunto inicial
n = número de compuestos en el nuevo subconjunto
n = 2
n = 3
X D A
Nuevo subconjunto de productos naturales

	A	B	C	D	E
X	0.5	0.3	0.7	0.1	0.8

	A	B	C	E
X	0.5	0.3	0.7	0.8
D	0.2	0.4	0.5	0.6

www.difacquim.com

Resultados

Se generaron tres subconjuntos de productos naturales y calcularon sus propiedades de interés farmacéutico (donadores de puente de hidrógeno, aceptores de puente de hidrógeno, área topológica superficial, número de enlaces rotables, peso molecular y el coeficiente de partición octanol/agua). Se realizó una visualización de espacio químico usando el algoritmo t-SNE (en inglés *T-distributed Stochastic Neighbor Embedding*). La **figura 1** y **figura 2** muestran que los subconjuntos de compuestos derivados del UNPD cubren regiones similares del espacio químico cubierto por la base de datos original (UNPD) y son representativos del UNPD con respecto a sus propiedades de interés farmacéutico. La **figura 3** muestra que los compuestos más diversos usando la función de distribución comulativa son aquellos cuya pendiente es más pronunciada, por esta razón, los subconjuntos UNPD-A, UNPD-B y UNPD-C tuvieron las estructuras químicas más diversas seguido del UNPD, BIOFACQUIM y el subconjunto de DNMT1.

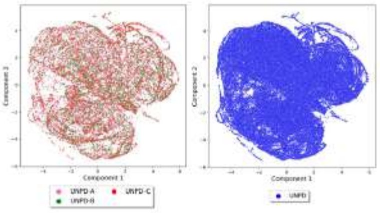


Figura 1. Visualización del espacio químico del UNPD y sus subconjuntos usando el algoritmo t-SNE (del inglés *T-distributed Stochastic Neighbor Embedding*) basado en sus propiedades de interés farmacéutico.

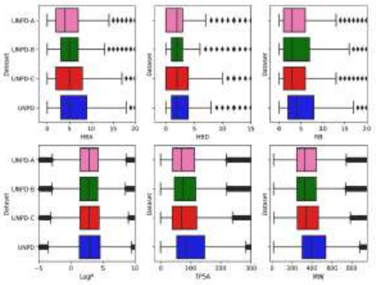


Figura 2. Gráfico de cajas de los subconjuntos del UNPD y UNPD usando seis propiedades de interés farmacéutico: donadores de puente de hidrógeno (HBD), aceptores de puente de hidrógeno (HBA), área topologica superficial (TPSA), número de enlaces rotables (RB), peso molecular (MW) y coeficiente de partición octanol/agua (log P). Las bases de datos son representadas en diferentes colores, UNPD-A (rosa), UNPD-B (verde), UNPD-C (rojo), y todo el UNPD (azul). Los diamantes negros muestran los puntos atípicos.

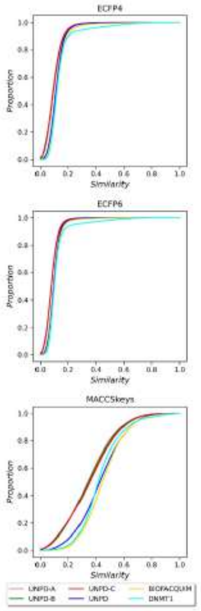


Figura 3. Función de distribución comulativa de similitud molecular de pares de moleculas utilizando el coeficiente de Tanimoto y como representación molecular las huellas digitales moleculares MACCS keys (166-bits), ECFP4 y ECFP6. Las bases de datos son representadas en una línea continúa y usando diferentes colores: UNPD-A (rosa), UNPD-B (verde), UNPD-C (rojo), UNPD (azul), BIOFACQUIM (amarillo) y DNMT1 (cyan).

Conclusiones y perspectivas

Se generaron y caracterizaron tres subconjuntos de productos naturales [4] con 14,994, 7,497 y 4,998 compuestos que tienen estructuras químicas y propiedades fisicoquímicas de interés farmacéutico muy similares a los compuestos en UNPD. Los subconjuntos se pueden utilizar para entrenar redes neuronales de aprendizaje profundo en grupos de investigación con recursos computacionales limitados.

Los subconjuntos del UNPD y el script de Python utilizado para generarlos están disponibles públicamente en repositorio de GitHub (código QR).

Agradecimientos

- A CONACyT por la beca de doctorado 847870.
- DGAPA, UNAM, proyecto No. IN201321.
- Al espacio de inovación UNAM-HUAWEI, proyecto No.7.
- DGTIC, UNAM, LANCAD-UNAM-DGTIC-335.

Referencias

1. Chávez-Hernández, A. L.; López-López, E.; Medina-Franco, J. L. *Front Drug Discov* **2023** (manuscrito en revisión).
2. Gu, J.; Gui, Y.; Chen, L.; et. al. *PLoS One* **2013**, 8, e62839.
3. Selecting Diverse Sets Of Compounds. In An Introduction To Chemoinformatics; Leach, A. R., Gillet, V. J., Eds.; Springer Netherlands: Dordrecht, 2007; pp 119–139.
4. Chávez-Hernández, A. L.; Medina-Franco, J. L. *Artif Intell Life Sci* **2023**, 3, 100066.

A Fragment Library of Natural Products and its Comparative Chemoinformatics Characterization

Ana L. Chávez-Hernández*, Norberto Sánchez-Cruz, and José L. Medina-Franco*
Department of Pharmacy, School of Chemistry, Universidad Nacional
Autónoma de México/ México
E-mail address: anachavez3026@gmail.com, medinajl@unam.mx

Abstract

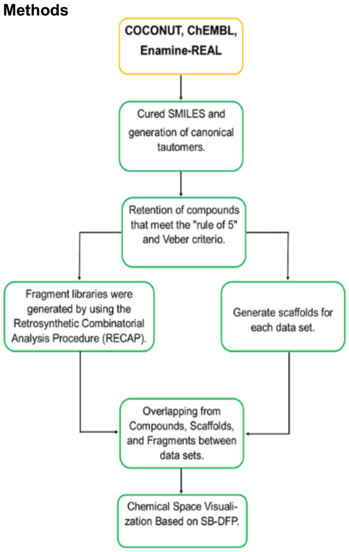
In this work, we discuss a compressive fragment library with 205, 903 fragments derived from recently published Collection of Open Natural Products (COCONUT) data set with more than 400,000 non-redundant natural products. The natural products-based fragment library was compared with other two fragment libraries herein generated from ChEMB (to represent biologically relevant compounds) and Enamine-REAL(a large on-demand or virtual collection of synthetic compounds), both used as reference data sets with relevance in drug discovery. It was found that there is a large diversity of unique fragments derived from natural products. It was also concluded that the entire chemical structures and fragments derived from natural products are more diverse and structurally complex than the two reference compound collections.

Introduction

Natural products (NPs) have been relevant in drug discovery pipelines since the beginning of the pharmaceutical era. They have gone through an adaptation process therefore they represent attractive ligands for several biological targets. NPs possess unique functional groups, unique scaffolds, and unique characteristic structural fragments that could provide important information related to biological activity. Thus, fragments obtained from NPs can be further used in traditional fragment-based or de novo drug design. Therefore, it is desirable to generate fragment libraries from NPs.

In this work, we report a novel and comprehensive database of fragments derived from NPs based on the COlleCtion of Open NatUral producTs (COCONUT), a recently published database with more than 400,000 nonredundant compounds. The fragment library was characterized and compared with fragment libraries herein generated from two large reference compound data sets with relevance in drug discovery: ChEMBL as a source of biologically relevant compounds, and Enamine-REAL, a large on-demand collection of synthetic compounds.

Methods



COCONUT, ChEMBL, Enamine-REAL

Cured SMILES and generation of canonical tautomers.

Retention of compounds that meet the "rule of 5" and Veber criterio.

Fragment libraries were generated by using the Retrosynthetic Combinatorial Analysis Procedure (RECAP).

Generate scaffolds for each data set.

Overlapping from Compounds, Scaffolds, and Fragments between data sets.

Chemical Space Visualization Based on SB-DFP.

Result

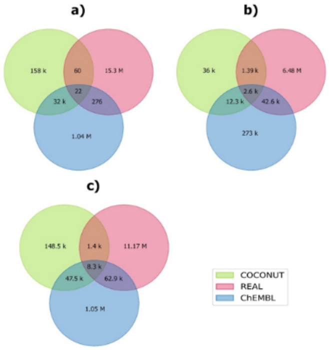


Figure 1. Unique and overlapping structures between COCONUT, ChEMBL and REAL data sets analyzed. Structural content was analyzed in terms of a) Compounds, b) Molecular scaffolds, and c) Fragments. The letter k represents thousands and the letter M represents millions.

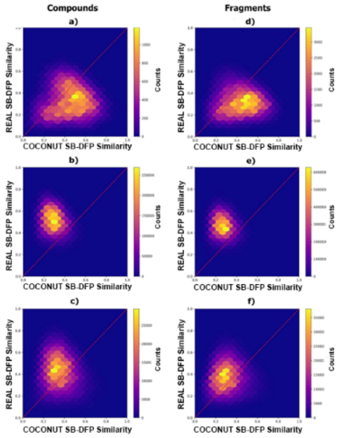


Figure 2. Visual representation of the chemical space for compounds and fragments of natural products, synthetics compounds, and biologically relevant compounds. The number of compounds is represented with a continuous color scale. Compounds data sets used: a) COCONUT, b) Enamine-REAL, c) ChEMBL. Fragment data sets used: d) COCONUT, e) Enamine-REAL, and f) ChEMBL.

Conclusions

The comparison of the unique and overlapping fragment of NPs with other reference collections revealed that there is a large diversity of unique fragments derived from NPs that could be used as building blocks for the de novo design and synthesis of novel compounds. It was also concluded that both the entire structures and fragments derived from NPs are more diverse and structurally complex than the two reference compound collections.

ChEMBL compounds had higher similarity to the REAL-SB-DFP further emphasizing the opportunity to increase the number of NPs tested for biological activity (e. g., enrich ChEMBL with drug-like compounds available in COCONUT).

References

Chávez-Hernández AL., Sánchez-Cruz, N. and Medina-Franco JL. (2020). A Fragment Library of Natural Products and its Comparative Chemoinformatic Characterization. *Mol. Inf.*

ASOCIACIÓN PERUANA DE ESTUDIANTES DE FARMACIA Y BIOQUÍMICA - UPNW

APEFYB

UNIVERSIDAD NORBERT
WIENER

CERTIFICADO DE RECONOCIMIENTO

SE OTORGA LA PRESENTE CONSTANCIA A

M.Sc. Chávez Hernández, Ana Luisa

por su participación como PONENTE en el evento "III CURSO TALLER DE BIOQUIMIOINFORMATICA - EL MUNDO DEL DISEÑO Y DESARROLLO DE FÁRMACOS" llevado a cabo los días 10 y 11 de Setiembre del 2022.
Duración de 2 horas académicas

David Rivera

Jesús David Rivera Salazar
PRESIDENTE APEFYB UPNW

Yeltsin Ramos Flores
VICEPRESIDENTE APEFYB UPNW

IBMC
Institute of Biomedical Chemistry

IYBSSD2022

2022 / May 24 - 26

SPECIAL DIPLOMA

XXVIII Symposium on Bioinformatics and Computer-Aided Drug Discovery

Ana Luisa Chavez-Hernandez

National Autonomous University of Mexico

«TOWARDS THE DE NOVO DESIGN OF HIV-1 PROTEASE INHIBITORS BASED ON NATURAL PRODUCTS»

Symposium Co-Chairs

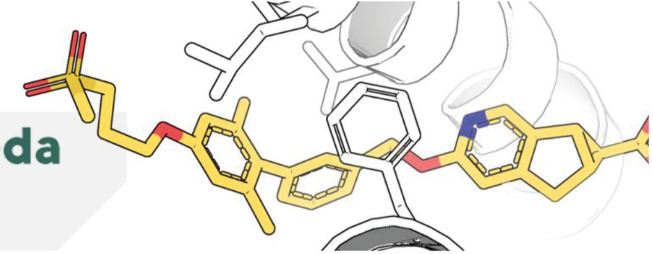
Corr. Member of Rus. Acad. Sci.
Vladimir V. Poroikov

Professor
Roman G. Efremov

Cursos

IX SIMPOSIO

Tendencias actuales en la búsqueda y desarrollo de fármacos



La Facultad de Química de la Universidad Nacional Autónoma de México

y el Colegio de Química Farmacéutica

otorgan la presente **Constancia** a:

M. en C. Ana Luisa Chávez Hernández

por su participación como ponente en el **Curso-Taller I (Quimioinformática)**, con una duración de 8 horas, impartido en el marco del IX Simposio *Tendencias actuales en la búsqueda y desarrollo de fármacos*, realizado los días 12 y 13 de junio de 2023.

"Por mi raza hablará el espíritu"

Ciudad Universitaria, Cd. Mx., a 29 de junio de 2023

Dr. Rodrigo Aguayo Ortiz	Dr. Alfonso S. Lira Rocha	Dr. Francisco Hernández Luis
Comité Organizador	Comité Organizador	Jefe del Departamento de Farmacia

Casa abierta al tiempo
UNIVERSIDAD AUTÓNOMA METROPOLITANA
Unidad Azcapotzalco

Coordinación de Docencia

UAMedia

Otorga la presente constancia a:

M. en C. Ana L. Chávez Hernández

Por impartir el curso **“Búsqueda, análisis, representación y visualización de información química contenida en bases de datos moleculares"** los días 9 al 13 de octubre y 16 al 19 de octubre del 2023, con una duración de 18 horas.

Mtra. Merary Denny Puga García
Coordinadora de docencia de la
Unidad Azcapotzalco

Distinciones científicas y reconocimientos

WILEY

Top Cited Article 2020-2021

Congratulations to:

Ana

whose paper has been recognized as a **top cited** paper in:

MOLECULAR INFORMATICS

A Fragment Library of Natural Products and its Comparative Chemoinformatic Characterization

PECI
Programa Estancias
Cortas de Investigación
FQ·UNAM

UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

La Facultad de Química,
a través de la Secretaría de Apoyo Académico
y la Coordinación de Atención a Alumnos,
otorga el presente

RECONOCIMIENTO

a la M en C Ana Luisa Chávez Hernández

por su asesoría como Profesora Adjunta del proyecto:
Desarrollo de flujos de trabajo aplicados al diseño de fármacos, ganador del
TERCER LUGAR
en la categoría de Química Farmacéutico Biológica,
del Programa Estancias Cortas de Investigación,
correspondiente al intersemestre 2023-1.

"Por mi raza hablará el espíritu"
Ciudad Universitaria, Cd. Mx., 25 de abril de 2023

Dr. Carlos Amador Bedolla
Director

FQ-DIR-ABRIL-2023-45

DGTIC HUAWEI SRE SECRETARÍA DE RELACIONES EXTERIORES

TIC/DG/DDTIC/0740/2022

DR. JOSÉ LUIS MEDINA FRANCO
PRESENTE

Con respecto a su propuesta de proyecto 7: **“Desarrollo y aplicación de algoritmos de inteligencia artificial para el diseño de fármacos aplicables al tratamiento de diabetes mellitus y cáncer,”** presentada para la Convocatoria 2022-Proyectos de investigación en Inteligencia Artificial en el Espacio de Innovación UNAM-HUAWEI, tenemos el agrado de comunicarle que después de realizar un análisis académico profundo por parte del Comité Evaluador del Grupo Especial de innovación, su proyecto ha sido **Aceptado.**

Por lo que, el acceso a los recursos de alto desempeño será de manera remota y proporcional de acuerdo con el número de proyectos aceptados, la persona responsable del proyecto aprobado deberá verificar que cuenta con los recursos locales de cómputo y conectividad que permitan el uso en línea de la infraestructura de alto desempeño que se pone a disposición mediante la presente convocatoria.

Con respecto a la asignación de becas, le informo que se encuentra en proceso de selección por parte del Comité de Evaluación, en cuanto se determine quienes cumplen con los lineamientos establecidos, lo haremos de su conocimiento.

Agradecemos su valiosa participación en este importante proyecto para promover el desarrollo de capacidades digitales en México.

Sin otro particular por el momento, reciba un afectuoso saludo.

Atentamente
“POR MI RAZA HABLARÁ EL ESPÍRITU”
Ciudad Universitaria, Cd. Mx., 23 de noviembre de 2022.

DR. HÉCTOR BENÍTEZ PÉREZ
Presidente del Grupo Especial
Espacio de Innovación UNAM-Huawei

CMF/NCC/lsc

WILEY

Top Downloaded Article

Congratulations to:

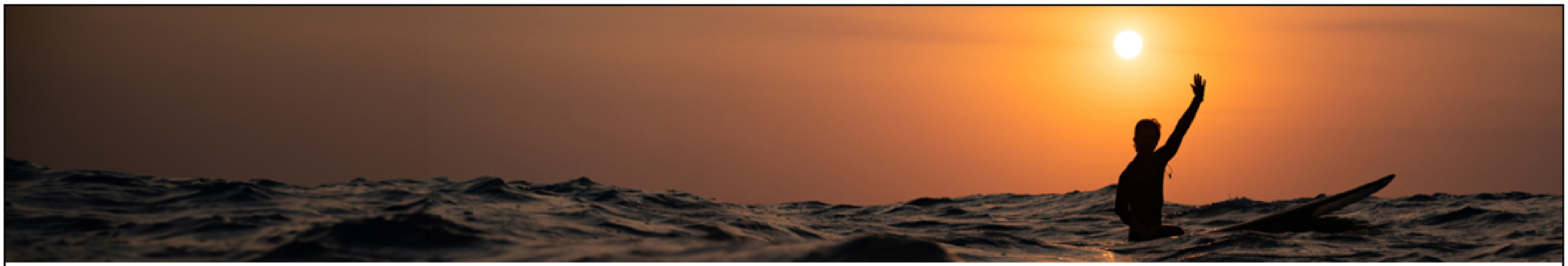
Ana Luisa Chávez-Hernández

Whose paper was one of the most downloaded* during its first 12 months of publication in:

MOLECULAR INFORMATICS

Chemical Multiverse: An Expanded View of Chemical Space

**Among work published in an issue between 1 January 2022 – 31 December 2022.*



WILEY

Top Cited Article 2022-2023

Congratulations to:

Ana Luisa Chávez Hernández

whose paper has been recognized as a top cited paper* in:

MOLECULAR INFORMATICS

Chemical Multiverse: An Expanded View of Chemical Space

**Among work published between 1 January 2022 – 31 December 2023.*