UNIVERSIDAD NACIONAL AUTÓNOMA DE MEXICO

PROGRAMA DE MAESTRÍA Y DOCTORADO EN CIENCIAS QUÍMICAS

SUITE COMPUTACIONAL PARA LA NAVEGACIÓN DE ESPACIOS QUIMIOGENÓMICOS

TESIS

PARA OPTAR POR EL GRADO DE

DOCTOR EN CIENCIAS

PRESENTA

M. en C. Norberto Sánchez Cruz

Dr. José Luis Medina Franco
Departamento de Farmacia, Facultad de Química, UNAM

Ciudad de México, marzo de 2021

UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

PROGRAMA DE MAESTRÍA Y DOCTORADO EN CIENCIAS QUÍMICAS

SUITE COMPUTACIONAL PARA LA NAVEGACIÓN DE ESPACIOS QUIMIOGENÓMICOS

TESIS
PARA OPTAR POR EL GRADO DE

DOCTOR EN CIENCIAS

PRESENTA

M. en C. Norberto Sánchez Cruz

Dr. José Luis Medina Franco
Departamento de Farmacia, Facultad de Química, UNAM

Ciudad de México, marzo de 2021

JURADO ASIGNADO

Presidente	Dr. Francisco Hernández Luis	Facultad de Química, UNAM
Vocal	Dr. Fernando Cortés Guzmán	Instituto de Química, UNAM
Vocal	Dra. Laura Domínguez Dueñas	Facultad de Química, UNAM
Vocal	Dr. Marcelino Arciniega Castro	Instituto de Fisiología Celular, UNAM
Secretario	Dr. Mario Alberto Figueroa Saldívar	Facultad de Química, UNAM

COMITÉ TUTOR DE EVALUACIÓN

Semestral	Dr. José Luis Medina Franco	Facultad de Química, UNAM
Semestral	Dr. Ramón Garduño Juárez	Instituto de Ciencias Físicas, UNAM
Semestral	Dra. Laura Domínguez Dueñas	Facultad de Química, UNAM
Ampliado	Dr. Joaquín Barroso Flores	Instituto de Química, UNAM
Ampliado	Dr. José Correa Basurto	Escuela Superior de Medicina, IPN

LUGAR DONDE SE DESARROLLÓ EL TRABAJO

El presente trabajo de tesis fue desarrollado en el cubículo 117 del edificio F, en la Facultad de Química de la Universidad Nacional Autónoma de México, bajo la dirección del Dr. José Luis Medina Franco.

Los resultados de este trabajo fueron presentados en los siguientes eventos:

- **Sánchez-Cruz, N.** Combining Artificial Intelligence with Chemical Knowledge to Improve Binding Affinity Predictions. Presentación oral. Congreso Internacional de Biología Estructural. Estructura, Dinámica e Interacciones. Congreso virtual. Octubre de 2020.

- González-Medina, M.; Díaz-Eufracio, B. I.; **Sánchez-Cruz, N.**; Medina-Franco, J.L. D-Tools: Open Chemoinfomatic Web Servers for Drug Development. Poster. EFMC International Symposium on Medicinal Chemistry and the EFMC Young Medicinal Chemists' Symposium. Simposio virtual. Septiembre de 2020.

- **Sánchez-Cruz, N.**; Medina-Franco, J. L.; Mestres, J.; Barril, X. Improving Machine-Learning Based Protein-Ligand Binding Affinity Predictions Through Chemical Description. Poster. Mención honorífica en concurso de posters. VI Latin American Protein Society Meeting and VII Congreso de Fisicoquímica, Estructura y Diseño de Proteínas. Ciudad de México, México. Octubre de 2019.

- **Sánchez-Cruz, N.**; Mestres, J.; Medina-Franco, J. L. Epigenetic Target Fishing through Classification Models based on Statistical-Based Database Fingerprint Similarities. Poster. Eighth Joint Sheffield Conference on Chemoinformatics. Sheffield, UK. Junio de 2019.

- **Sánchez-Cruz, N.**; Medina-Franco, J. L. Statistical-based Database Fingerprint: Application in Ligand-based Virtual Screening. Presentación oral. 256th ACS National Meeting. Boston, MA, USA. Agosto de 2018.

- **Sánchez-Cruz, N.**; González-Medina, M; Naveja, J. J.; Medina- Franco, J. L. Análisis y Diseño de Herramientas en Línea para la Exploración del Espacio Quimiogenómico. Presentación oral. 52° Congreso Mexicano de Química y 36° Congreso Nacional de Educación Química. Jalisco, México. Septiembre de 2017.

Los resultados de esta tesis se reportan en las siguientes publicaciones revisadas por pares:

- **Sánchez-Cruz, N.***; Medina-Franco, J. L. Epigenetic Target Profiler: A Web Server to Predict Epigenetic Targets of Small Molecules. *J. Chem. Inf. Model.* **2021**. En prensa. https://doi.org/10.1021/acs.jcim.1c00045.

- **Sánchez-Cruz, N.***; Medina-Franco, J. L. Epigenetic Target Fishing with Accurate Machine Learning Models. *J. Med. Chem.* **2021**. En prensa. https://doi.org/10.1021/acs.jmedchem.1c00020.

- **Sánchez-Cruz, N.***; Medina-Franco, J. L.; Mestres, J.; Barril, X. Extended Connectivity Interaction Features: Improving Binding Affinity Prediction through Chemical Description. *Bioinformatics* **2020**. En prensa https://doi.org/10.1093/bioinformatics/btaa982.

- Sessions, Z.; **Sánchez-Cruz, N.**; Prieto-Martínez, F. D.; Alves, V. M.; Santos, H. P.; Muratov, E.; Tropsha, A.; Medina-Franco, J. L. Recent Progress on Cheminformatics Approaches to Epigenetic Drug Discovery. *Drug Discov. Today* **2020**, *25* (12), 2268–2276. https://doi.org/10.1016/j.drudis.2020.09.021.

- 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**, *39* (11), 2000050. https://doi.org/10.1002/minf.202000050.

- **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. https://doi.org/10.12688/f1000research.21540.2.

- **Sánchez-Cruz, N.***; Medina-Franco, J. L. Statistical-Based Database Fingerprint: Chemical Space Dependent Representation of Compound Databases. *J. Cheminf.* **2018**, *10* (1), 55. https://doi.org/10.1186/s13321-018-0311-x.

- González-Medina, M.; Naveja, J. J.; **Sánchez-Cruz, N.**; Medina-Franco, J. L. Open Chemoinformatic Resources to Explore the Structure, Properties and Chemical Space of Molecules. *RSC Adv.* **2017**, *7* (85), 54153–54163. https://doi.org/10.1039/C7RA11831G.

Además, se contribuyó en la elaboración de las siguientes publicaciones revisadas por pares empleando conceptos y métodos relacionados con los objetivos de la tesis:

- Juárez-Mercado, K. E.; Prieto-Martínez, F. D.; **Sánchez-Cruz, N.**; Peña-Castillo, A.; Prada-Gracia, D.; Medina-Franco, J. L. Expanding the Structural Diversity of DNA Methyltransferase Inhibitors. *Pharmaceuticals* **2020**, *14* (1), 17. https://doi.org/10.3390/ph14010017.

- 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* (11), 1518. https://doi.org/10.3390/biom10111518.

- 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* (42), 25089–25099. https://doi.org/10.1039/D0RA04922K.

- Tran, T. D.; Ogbourne, S. M.; Brooks, P. R.; **Sánchez-Cruz, N.**; Medina-Franco, J. L.; Quinn, R. J. Lessons from Exploring Chemical Space and Chemical Diversity of Propolis Components. *Int. J. Mol. Sci.* **2020**, *21* (14), 4988. https://doi.org/10.3390/ijms21144988.

- Saldívar-González, F. I.; Gómez-García, A.; Chávez-Ponce de León, D. E.; **Sánchez-Cruz, N.**; Ruiz-Rios, J.; Pilón-Jiménez, B. A.; Medina-Franco, J. L. Inhibitors of DNA Methyltransferases from Natural Sources: A Computational Perspective. *Front. Pharmacol.* **2018**, *9*. https://doi.org/10.3389/fphar.2018.01144.

- Naveja, J. J.; Saldívar-González, F. I.; **Sánchez-Cruz, N.**; Medina-Franco, J. L. Cheminformatics Approaches to Study Drug Polypharmacology. En Roy K. (Ed). Multi-Target Drug Design Using Chem-Bioinformatic Approaches. Methods in Pharmacology and Toxicology. Humana Press, New York, NY.; 2018; pp 3–25. https://doi.org/10.1007/7653_2018_6.

AGRADECIMIENTOS

A la Universidad Nacional Autónoma de México (UNAM), en especial a la Facultad de Química, por brindar todas las herramientas necesarias para mi formación tanto académica como personal.

Al Consejo Nacional de Ciencia y Tecnología (CONACyT), por las becas otorgadas tanto para realizar mis estudios de posgrado, como para hacer una estancia de investigación de siete meses en la Universidad Pompeu Fabra en Barcelona, Cataluña, España. Número de becario: 335997.

Al Banco Santander, por la beca otorgada para efectuar una estancia de investigación de cuatro meses en la Universidad de Barcelona en Barcelona, Cataluña, España.

Al Programa de Apoyo a los Estudios de Posgrado (PAEP) y al Posgrado en Ciencias Químicas de la UNAM, por el apoyo económico otorgado para la asistencia a un congreso internacional.

Al Programa de Apoyo a la Investigación y el Posgrado (PAIP 5000-9163), Programa de Apoyo a Proyectos de Investigación e Innovación Tecnológica (PAPIIT IA203718), Programa de Apoyo a Proyectos para la Innovación y Mejoramiento de la Enseñanza (PAPIME PE200118) y Programa de Nuevas Alternativas para el Tratamiento de Enfermedades Infecciosas de la UNAM, así como al Proyecto de Ciencia Básica de CONACyT (282785), por el financiamiento otorgado para la realización de este proyecto.

A la *Dirección General de Cómputo y de Tecnologías de Información y Comunicación* (DGTIC) de la UNAM, por los recursos y soporte brindados para el uso de la supercomputadora “Miztli” (LANCAD-UNAM-DGTIC-335), así como a la compañía ChemAxon y al Dr. Chao-Yie Yang, por las licencias académica para el uso de los programas JChem y X-Score, respectivamente.

Al Dr. José Luis Medina Franco, por haberme permitido formar parte de su grupo de investigación, por su apoyo como supervisor y amigo durante la realización de este proyecto y por todas las oportunidades brindadas para mi desarrollo académico.

Al Dr. Ramón Garduño Juárez y la Dra. Laura Domínguez Dueñas por sus constantes y valiosas aportaciones para el desarrollo de este proyecto.

Al Dr. Xavier Barril y el Dr. Jordi Mestres, por aceptarme como un miembro más de sus grupos durante mis estancias de investigación.

A mi familia, a quienes debo todo lo que soy, por su cariño y apoyo incondicional desde siempre.

A la familia Manríquez Tolsá, por tomarme como un miembro más de su familia y acompañarme a lo largo de estos cuatro años. En especial a Patricia Manríquez Tolsá, por la compañía, cariño y apoyo brindados durante los buenos y malos momentos que involucraron la realización de este trabajo.

A mis amigos, los de toda la vida y los que tuve la oportunidad de conocer en este camino, por permitirme compartir con ellos esta grata experiencia.

A mis compañeros del grupo de Diseño de Fármacos Asistido por Computadora (DIFACQUIM), por todas sus enseñanzas y los buenos momentos compartidos.

A los miembros del jurado por sus valiosos comentarios y correcciones al presente trabajo de tesis.

A mis queridos padres, David y Magdalena
Sin ustedes nada de esto habría sido posible

ÍNDICE

ABREVIATURAS

1-NN, Vecino más cercano, *1-Nearest-Neighbour*
2D: Bidimensional
3D: Tridimensional
ADMET: Absorción, Distribución, Metabolismo, Excreción y Toxicidad
ANN: Redes Neuronales Artificiales, *Artificial Neural Networks*
BA: Exactitud Balanceada, *Balanced Accuracy*
BET: Proteína que contiene Bromodominio, *Bromodomain and Extra-Terminal protein*
CHR: Remodelador de Cromatina, *Chromatin Remodeler*
CNN: Redes Neuronales Convolucionales, *Convolutional Neural Networks*
DIFACQUIM: Diseño de Fármacos Asistido por Computadora en la Facultad de Química
DM: Distancia al Modelo, *Distance-to-Model*
DNMT: DNA Metiltransferasa, *DNA Methyltransferase*
ECFP: Huellas Digitales de Conectividad Extendida, *Extended Connectivity Fingerprints*
EBI: Instituto Europeo de Bioinformática, *European Bioinformatics Institute*
EMBL: Laboratorio Europeo de Biología Molecular, *European Molecular Biology Laboratory*
FDA: Administración de Alimentos y Medicamentos, *Food and Drug Administration*
GBT: Árboles de Gradiente Potenciado, *Gradient Boosting Trees*
GPCR: Receptor acoplado a proteína G*, G protein-coupled receptors*
GUI: Interfaz Gráfica de Usuario, *Graphical User Interface*
HAT: Histona Acetiltransferasa, *Histone Acetyltransferase*
HDAC: Histona Deacetilasa, *Histone Deacetylase*
HDM: Histona Demetilasa, *Histone Demethylase*
HMR: Lector de Metilación de Histonas, *Histone Methyl Reader*
HMT: Histona Metiltransferasa, *Histone Methyltransferase*
k-NN: *k* Vecinos más cercanos, *k-Nearest Neighbors*
KIN: Cinasa, *Kinase*
MACCS: *Molecular ACCess System*
MRNA: Modificador de RNA, *RNA Modifier*
NCBI: Centro Nacional de Información Biotecnológica, *National Center for Biotechnology Information*
NIH: Instituto Nacional de Salud, *National Institute of Health*
PcG: Proteína del Grupo Polycomb, *Polycomb Group Protein*
PDB: Banco de Datos de Proteínas, *Protein Data Bank*
PPV: Precisión, *Positive Predictive Value*
PRMT: Arginina Metiltransferasa, *Protein Arginine Methyltransferase*
QSAR: Relaciones Cuantitativas Estructura-Actividad, *Quantitative Structure-Activity Relationships*
RF: Bosque Aleatorio, *Random Forest*
ROCS: *Rapid Overlay of Chemical Structures*
SAR: Relaciones Estructura-Actividad, *Structure-Activity Relationships*
SB-DFP*: Statistical-Based Database Fingerprint*
SDF: *Structure Data File*
SMARt*:* Relaciones Estructura Múltiple-Actividad, *Structure-Multiple Activity Relationships*
SMILES: *Simplified Molecular-Input Line-Entry System*
SVM: Máquinas de Soporte Vectorial, *Support Vector Machines*
T_c, Coeficiente de Tanimoto, *Tanimoto coefficient*
TPR: Sensibilidad, *True Positive Rate*
TUD: Proteína que contiene el dominio Tudor, Tudor Domain-Containing Protein
USR: Ultrafast Shape Recognition

RESUMEN

En este trabajo, se presenta el desarrollo, validación, implementación y uso de tres herramientas computacionales con aplicación en la predicción de interacciones entre moléculas pequeñas y dianas biológicas: 1) *Statistical-Based Database Fingerprint* (SB-DFP), una huella digital de bases de datos de compuestos con aplicación en el cribado virtual basado en el ligando, el estudio de relaciones entre dianas biológicas y la visualización del espacio químico, 2) *Epigenetic Target Profiler* (ETP), una aplicación web que implementa modelos de aprendizaje automático para el cribado virtual inverso basado en el ligando, con un enfoque en dianas epigenéticas, y 3) *Extended Connectivity Interaction Features* (ECIF), una representación de complejos proteína-ligando y su empleo en la construcción de una función de puntuación basada en aprendizaje automático para la predicción de su afinidad de unión. En beneficio de la ciencia abierta, el código generado para el desarrollo, uso e implementación de estos recursos se encuentra disponible en repositorios públicos. Los resultados de esta tesis doctoral contribuyen al desarrollo y optimización de métodos computacionales para la navegación de espacios quimiogenómicos. Las herramientas desarrolladas tienen un potencial considerable para su implementación en proyectos multidisciplinarios para la identificación de nuevas interacciones entre moléculas pequeñas y dianas biológicas. Se espera que esta tesis doctoral contribuya al desarrollo conceptual y aplicado de la quimiogenómica computacional.

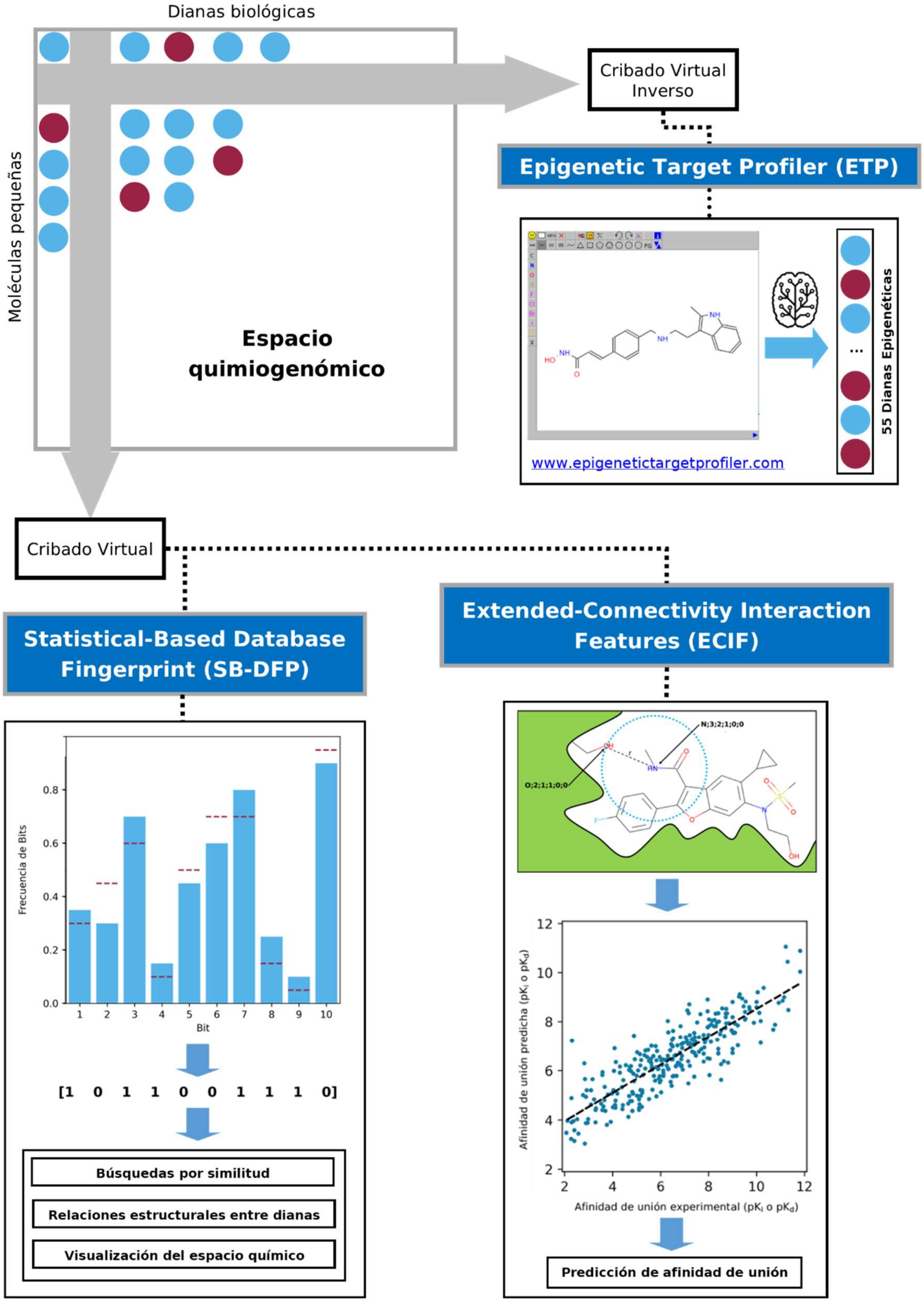
Palabras clave: Acoplamiento Molecular, Aplicaciones Web, Aprendizaje Automático, Bioinformática, Cribado Virtual, Epigenética, Inteligencia Artificial, Modelado molecular, Quimioinformática, Quimiogenómica Computacional.

ABSTRACT

In this work, the development, validation, implementation, and use of three computational tools with application in the prediction of interactions between small molecules and biological targets are presented: 1) *Statistical-Based Database Fingerprint* (SB-DFP), a fingerprint of compound databases with application in ligand-based virtual screening, the study of the relationships between biological targets and visualization of the chemical space, 2) *Epigenetic Target Profiler* (ETP), a web application that implements machine learning models for epigenetic target fishing, and 3) *Extended Connectivity Interaction Features* (ECIF), a representation of protein-ligand complexes and its use in the construction of a machine learning-based scoring function for predicting its binding affinity. For the benefit of open science, the code generated for the development, use and implementation of these resources is available at public repositories. This doctoral dissertation results contribute to the development and optimization of computational methods for navigating chemogenomic spaces. The tools herein presented have considerable potential for their implementation in multidisciplinary projects for identifying new interactions between small molecules and biological targets. It is expected that this thesis dissertation contributes to the further conceptual and practical development of computational chemogenomics.

Keywords: Artificial Intelligence, Bioinformatics, Cheminformatics, Computational Chemogenomics, Epigenetics, Machine Learning, Molecular Docking, Molecular Modeling, Virtual Screening, Web Applications.

RESUMEN GRÁFICO


Dianas biológicas
Moléculas pequeñas
Espacio quimiogenómico
Cribado Virtual Inverso
Epigenetic Target Profiler (ETP)
55 Dianas Epigenéticas
www.epigenetictargetprofiler.com
Cribado Virtual
Statistical-Based Database Fingerprint (SB-DFP)
Frecuencia de Bits
Bit
0.0
0.2
0.4
0.6
0.8
1
2
3
4
5
6
7
8
9
10
[1 0 1 1 0 0 1 1 1 0]
Búsquedas por similitud
Relaciones estructurales entre dianas
Visualización del espacio químico
Extended-Connectivity Interaction Features (ECIF)
N;3;2;1;0;0
O;2;1;1;0;0
Afinidad de unión predicha (pKi o pKd)
Afinidad de unión experimental (pKi o pKd)
2
4
6
8
10
12
Predicción de afinidad de unión

La suite computacional para la navegación de espacios quimiogenómicos desarrollada en esta tesis integra tres herramientas principales: SB-DFP, ETP y ECIF.

I. INTRODUCCIÓN

La quimiogenómica es la disciplina encargada de identificar las interacciones entre moléculas pequeñas y dianas biológicas, lo cual es un aspecto fundamental en el desarrollo de fármacos. La identificación de estas interacciones debe realizarse experimentalmente, sin embargo, al tener en cuenta la cantidad de compuestos químicos y dianas biológicas que existen, la exploración de todas las posibles interacciones entre estos es un escenario imposible de alcanzar en forma experimental.

Los métodos computacionales han mostrado ser una herramienta útil en la predicción de estas interacciones, reduciendo así tanto el tiempo como los recursos invertidos en la realización de experimentos. Con ello surge la quimiogenómica computacional, un área de investigación que forma parte del diseño de fármacos asistido por computadora mediante la integración de conceptos de la quimioinformática y la bioinformática. Actualmente, existe una gran variedad de métodos computacionales para predecir interacciones entre moléculas pequeñas y dianas biológicas. Estos métodos utilizan la información experimental conocida para construir modelos que permitan el planteamiento de nuevas hipótesis de interacción. Lo anterior representa un sesgo de los modelos actuales hacía los grupos de dianas más estudiadas, haciendo que las dianas poco exploradas, como las relacionadas con la regulación epigenética, estén representadas pobremente en ellos a pesar de su relevancia terapéutica. Además, la cantidad de información experimental disponible aumenta constantemente, por lo que deben desarrollarse nuevos modelos que utilicen la disponibilidad de datos creciente.

En esta tesis se desarrollaron, validaron, implementaron y aplicaron tres nuevas herramientas computacionales para predecir interacciones entre moléculas pequeñas y dianas biológicas. Estos recursos abarcan aspectos relacionados tanto a metodologías basadas en la estructura del ligando, como metodologías basadas en la estructura de la diana. Dos de estos recursos tienen una aplicación general, mientras que la tercera centra su aplicación en dianas relacionadas con la regulación epigenética. Los recursos computacionales que se desarrollaron son de libre acceso y asistirán la identificación de nuevas interacciones entre moléculas pequeñas y dianas biológicas al permitir la priorización de compuestos para su evaluación experimental.

II. ANTECEDENTES

II.1. Quimiogenómica en el descubrimiento de fármacos

El descubrimiento de fármacos es un proceso complejo y costoso. Sin embargo, a pesar de los avances tecnológicos y el constante incremento de la inversión en investigación y desarrollo por parte de las compañías farmacéuticas, el número de fármacos aprobados para uso clínico se ha frenado (Booth and Zemmel, 2004; Paul *et al.*, 2010; Smietana *et al.*, 2015). Por ejemplo, el número de fármacos aprobados por la Administración de Alimentos y Medicamentos de los Estados Unidos (FDA, *Food and Drug Administration*) mostró una disminución clara en la década de los 2000 y no fue sino hasta 2018 que superó su máximo histórico, alcanzado en 1996 (Mullard, 2021) (Figura 1). Este suceso fue posible en parte, gracias a la creciente tendencia por el reposicionamiento de fármacos, la búsqueda de nuevas aplicaciones para los fármacos existentes (Chong and Sullivan, 2007; Pantziarka *et al.*, 2018), que se ha mostrado en la última década.

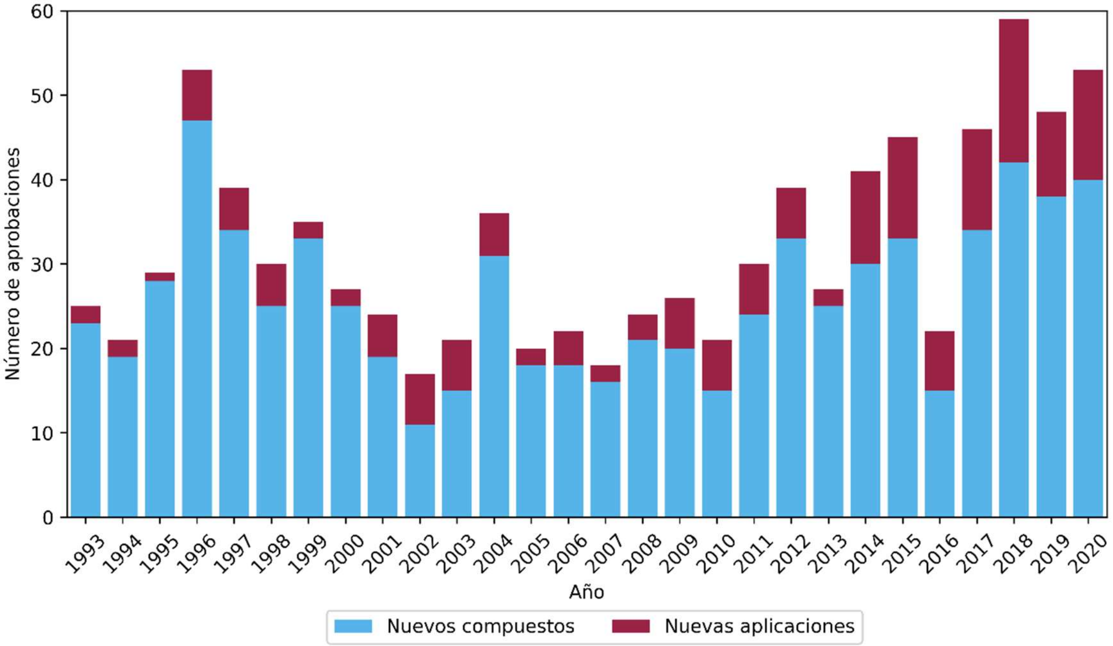


Figura 1. Fármacos aprobados por la FDA desde 1993. *Número de nuevos compuestos y nuevas aplicaciones aprobadas por la FDA. Las vacunas y terapias genéticas no se incluyen en esta gráfica. Adaptado de* (Mullard, 2021)*.*

Ya sea para el reposicionamiento de un fármaco, el desarrollo u optimización de una nueva molécula, la elucidación de los mecanismos de acción de compuestos bioactivos o la anticipación de potenciales efectos adversos de estas, la identificación de las interacciones de compuestos con dianas biológicas relevantes es uno de los pasos claves (Medina-Franco *et al.*, 2013). Determinar estas interacciones requiere un esfuerzo multidisciplinario, involucrando conceptos y técnicas de disciplinas como el cribado de alto rendimiento, la química combinatoria, la bioinformática, la quimioinformática y el modelado molecular, entre otras (López-López *et al.*, 2021). La búsqueda constante de nuevas moléculas con actividad biológica y nuevas dianas terapéuticas condujo a que todas estas disciplinas fueran integradas en una sola, dando origen a la quimiogenómica (Bleicher, 2002; Bredel and Jacoby, 2004).

La quimiogenómica busca, en principio, la identificación de todas las posibles interacciones entre compuestos químicos y dianas biológicas (interacciones diana-ligando). Esta meta es prácticamente imposible de lograr, debido al inmenso número de compuestos y dianas biológicas que existen; se estima que el número de compuestos con potencial actividad biológica supera los 10^{60} (Lipinski *et al.*, 1997; Reymond, 2015), mientras que tan solo el proteoma humano consta de más de 20 mil proteínas (Gaudet *et al.*, 2013). Por esta razón, la quimiogenómica se ha enfocado en el cribado experimental (y computacional, discutido abajo), de bibliotecas de compuestos (o quimiotecas) sobre familias de dianas biológicas específicas, que se pueden entender como espacios quimiogenómicos acotados (Müller, 2004).

De particular interés para el descubrimiento de fármacos, son los espacios quimiogenómicos que involucran dianas biológicas cuya modulación a través de moléculas pequeñas conlleva a un efecto terapéutico (dianas terapéuticas). Las dianas biológicas más estudiadas dentro de estos espacios son las proteínas, dentro de las cuales destacan algunas clases, como las cinasas, los receptores acoplados a proteínas G (GPCRs, *G protein-coupled receptors*) y los canales iónicos (Oprea et al., 2018; Zdrazil et al., 2020). Sin embargo, otras clases de proteínas, como las relacionadas con la regulación epigenética, han sido poco exploradas. Este es un hecho destacable, puesto que a la fecha, ocho fármacos para el tratamiento de distintos tipos de cáncer actúan sobre este tipo de proteínas (Ganesan *et al.*, 2019; de Lera and Ganesan, 2020) y que además, la desregulación de estas dianas biológicas ha sido vinculada con el desarrollo de otras enfermedades, como la esclerosis múltiple (Küçükali *et al.*, 2015), algunos trastornos depresivos (Januar *et al.*, 2015) y otras enfermedades catalogadas como raras (Brindisi *et al.*, 2020). Estas razones hacen que los reguladores epigenéticos sean un foco atractivo para la investigación en quimiogenómica.

II.2 Quimiogenómica computacional

Si bien la identificación de las interacciones diana-ligando debe realizarse experimentalmente, estas determinaciones son muy costosas, tanto en tiempo como en recursos. Por ello, los métodos computacionales han surgido como una herramienta útil en la predicción de potenciales interacciones, priorizando y reduciendo así el espacio de búsqueda a explorar mediante experimentos. Esta complementariedad entre estudios experimentales y computacionales ha sido ampliamente reconocida en el campo (Rognan, 2007; Bajorath, 2013).

La quimiogenómica computacional requiere de bases de datos que integren los datos conocidos acerca de las relaciones estructura-actividad (SAR, *structure-activity relationships*) existentes entre moléculas pequeñas y dianas biológicas. Debido a que muchos compuestos presentan afinidad a varios blancos moleculares a la vez (ya sea ejerciendo un efecto clínico benéfico o reacciones adversas), es de interés creciente el estudio de relaciones estructura múltiple-actividad (SMARt, *structure-multiple activity relationships*) (Saldívar-González *et al.*, 2017). Dicha información es analizada para la construcción y validación computacional de modelos predictivos que permitan establecer hipótesis de potenciales interacciones. La validación experimental de dichas predicciones se traduce en el conocimiento de un mayor número de SAR/SMARt, que a su vez incrementan el número de SAR/SMARt conocidas y permiten el desarrollo de nuevos modelos predictivos (Strausberg, 2003). Este proceso se esquematiza en la Figura 2.

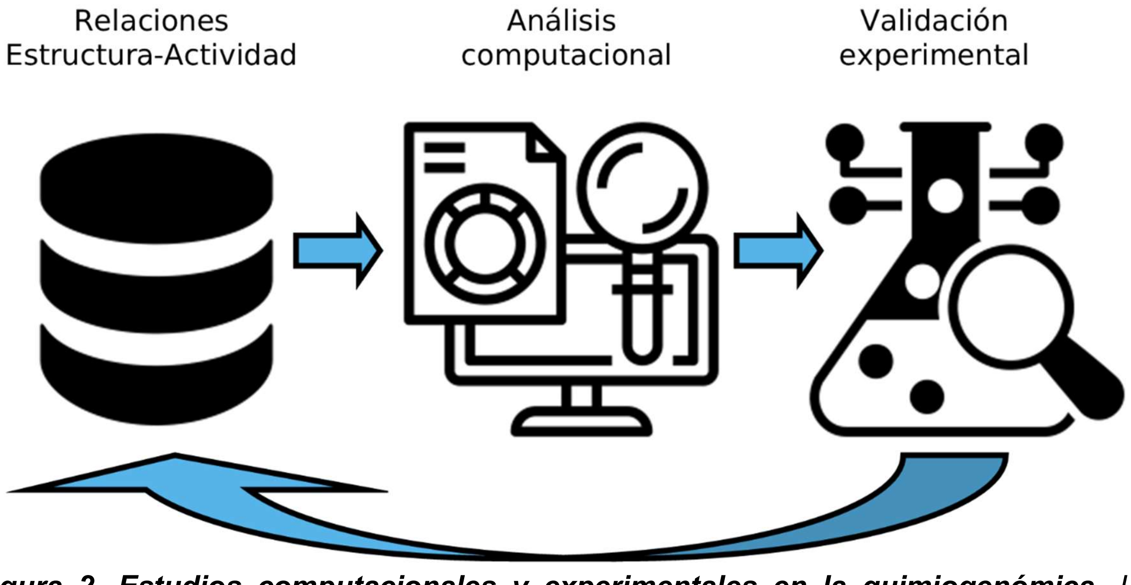


Figura 2. Estudios computacionales y experimentales en la quimiogenómica. *La combinación de ambos enfoques resulta en un ciclo iterativo y sinérgico que permite el descubrimiento de nuevas interacciones entre compuestos y dianas biológicas.*

La quimiogenómica computacional integra conceptos y herramientas relacionadas a la quimioinformática, bioinformática y modelado molecular para la generación de modelos predictivos clasificados en dos grandes categorías: métodos basados en la estructura del ligando y métodos basados en la estructura de la diana biológica. Las bases de datos con información de SAR/SMARt útiles para su desarrollo, así como las características de cada uno de los enfoques, se abordan en las secciones II.3 y II.4.

II.3 Bases de datos quimiogenómicas

Una parte esencial para el desarrollo de la quimiogenómica computacional es la disponibilidad de bases de datos que integren la información experimental existente de las SAR entre compuestos químicos y dianas biológicas. Actualmente existen múltiples bases de datos en el dominio público que abordan las interacciones diana-ligando desde distintos enfoques (González-Medina *et al.*, 2017; Al Mahmud *et al.*, 2018; Singh *et al.*, 2020). De especial interés para su análisis y posterior desarrollo de modelos predictivos, son aquellas que incluyen valores cuantitativos para las asociaciones diana-ligando. A continuación, se describen ejemplos representativos de estas bases de datos.

PubChem es una base de datos creada y mantenida por el Centro Nacional de Información Biotecnológica (NCBI, *National Center for Biotechnology Information*) del Instituto Nacional de Salud (NIH, *National Institutes of Health*) de los Estados Unidos de América. Incluye datos de actividad biológica depositados por la academia, la industria, institutos privados y agencias gubernamentales alrededor de todo el mundo. Su contenido es accesible a través de tres recursos principales: PubChem BioAssay Database, PubChem Compound Database y PubChem Substance Database. La primera contiene los resultados del cribado de moléculas pequeñas en múltiples ensayos biológicos. Las últimas dos incluyen la estructura química e información adicional de las moléculas pequeñas involucradas, abarcando su estructura, algunas propiedades fisicoquímicas, e incluso información de proveedores comerciales. Su versión actual (revisada en enero de 2021) incluye información de más de 109 millones de compuestos, asociados a más de 99 mil dianas biológicas a través de más de 273 millones de datos de actividad biológica. Al ser los cribados de alto rendimiento su principal fuente de información, la mayoría de estas asociaciones no son cuantitativas.

ChEMBL (Gaulton *et al.*, 2012; Bento *et al.*, 2014; Gaulton *et al.*, 2017; Mendez *et al.*, 2019) es una base de datos mantenida por el Instituto Europeo de Bioinformática (EBI, *European Bioinformatics Institute*) del Laboratorio Europeo de Biología Molecular (EMBL, *European Molecular Biology Laboratory*). Incluye datos de actividad biológica cuantitativos, extraídos tanto de la literatura científica como de PubChem. En lo que se refiere a la información de moléculas pequeñas, incluye su estructura, propiedades fisicoquímicas calculadas, así como datos relacionados a su perfil de Absorción, Distribución, Metabolismo, Excreción y Toxicidad (ADMET). En cuanto a las dianas biológicas, incluye información extensa sobre los ensayos biológicos que las involucran, abarcando el tipo de diana, la línea celular y/o tejido donde fue expresada, así como información del organismo al que pertenece. Las relaciones diana-ligando son establecidas mediante datos cuantitativos de actividad biológica, como porcentajes de inhibición, constantes de afinidad, etc. Su versión más reciente (ChEMBL 28, actualizada en febrero de 2021) incluye información referente a más de 14 mil dianas biológicas y más de 2 millones de compuestos, relacionados entre sí mediante más de 17 millones de datos de actividad biológica.

Binding Database (Liu *et al.*, 2007; Gilson *et al.*, 2016) es una base de datos desarrollada por la Universidad de California en San Diego, Estados Unidos. Incluye datos de afinidad de unión de moléculas pequeñas a proteínas consideradas como potenciales dianas terapéuticas. Estos datos son extraídos de PubChem, ChEMBL, literatura científica adicional e incluso patentes. Su versión actual (revisada en enero 2021) cuenta con cerca de 2 millones de datos de afinidad que relacionan a más de 8 mil proteínas con más de 920 mil compuestos diferentes.

Las bases de datos anteriores son probablemente los recursos públicos más grandes disponibles de datos de actividad biológica, por lo que han servido como punto de partida para el desarrollo de bases de datos enfocadas. Por ejemplo, ExCAPE-DB (Sun *et al.*, 2017) que integra los datos de actividad biológica de PubChem y ChEMBL únicamente para proteínas, PDBBind (Liu *et al.*, 2015) que integra estos datos para complejos diana-ligando con una estructura tridimensional conocida, DrugBank (Wishart, 2006; Wishart *et al.*, 2018) con un enfoque en fármacos, entre otras (Mathias *et al.*, 2013; Chang *et al.*, 2015; Skuta *et al.*, 2017).

A pesar de que las bases de datos que se mencionan representan una fuente de información pública muy valiosa, debe tenerse en cuenta que los datos reportados en ellas no están exentos de errores. Por tal razón, la utilización de sus datos para el desarrollo de modelos predictivos requiere de un análisis previo que permita la identificación y eliminación de los datos erróneos. Este tipo de análisis se conocen como curado de los datos y es un paso fundamental en el desarrollo de modelos predictivos a partir de información experimental en el dominio público (Fourches *et al.*, 2015; Bender and Cortes-Ciriano, 2021).

II.4 Desarrollo de modelos predictivos

Los modelos computacionales empleados para la generación de hipótesis de interacciones diana-ligando pueden ser clasificados en dos grandes grupos: basados en la estructura del ligando y basados en la estructura de la diana (Lill, 2013; Li, 2019). Cada uno de estos métodos involucra el conocimiento de distintos tipos de información y el uso de conceptos y herramientas diferentes, aspectos que se abordan en las siguientes secciones.

II.4.1 Métodos basados en la estructura del ligando

Los métodos basados en la estructura del ligando tienen su fundamento en el principio de similitud, el cual establece que moléculas similares tendrán propiedades similares y se unirán a la misma diana biológica (Klopmand, 1992). Estos modelos predicen la interacción de un compuesto con una diana biológica siempre que existan compuestos similares que se unan a dicha diana. Este tipo de aproximaciones requieren del conocimiento de al menos un ligando para la diana biológica de interés, aunque el conocimiento de múltiples ligandos es preferido, pues esto incrementa la confiabilidad de las predicciones. Las principales diferencias entre los distintos métodos que pueden ser agrupados dentro de esta categoría reside en la forma de representar a las moléculas y la métrica de similitud utilizada para su comparación (Jacob and Vert, 2008; Kristensen *et al.*, 2013). Desde luego se sabe que el principio de similitud no siempre se cumple y que existen los llamados “acantilados de actividad”, o en forma más general, “acantilados de propiedad”: pares de moléculas con estructuras químicas muy parecidas, pero propiedades muy diferentes, por ejemplo, actividades biológicas (Maggiora *et al.*, 2020). Una forma de disminuir y minimizar el impacto de estos posibles acantilados, es mediante el desarrollo de distintos modelos basados en la estructura de ligando usando representaciones moleculares de diferente diseño, pues estos acantilados son susceptibles a la forma de representación de las moléculas. También es fundamental verificar la confiabilidad de los datos experimentales de la propiedad de interés, para evitar la presencia de acantilados “fantasmas” o artificiales (Medina-Franco, 2013).

II.4.1.1 Representaciones moleculares

La representación molecular es uno de los conceptos centrales que distingue diversas disciplinas y subdisciplinas informáticas, incluyendo a la bioinformática y quimioinformática (López-López *et al.*, 2021). En quimioinformática, las representaciones de moléculas pequeñas usadas más ampliamente se pueden agrupar en dos grandes categorías: bidimensionales (2D) cuando se toma en cuenta solo la conectividad de las moléculas, y tridimensionales (3D) cuando se considera el volumen y la disposición espacial de las moléculas.

Dentro de las representaciones 2D destacan las huellas digitales moleculares, las cuales son una cadena de bits de un tamaño definido, en donde cada posición representa la presencia (1) o ausencia (0) de alguna subestructura. En general, existen tres tipos de huellas digitales moleculares: basadas en diccionario, circulares y topológicas. En las huellas digitales basadas en diccionario, los bits representan subestructuras previamente definidas en una lista acotada de posibilidades y se consideran huellas digitales independientes de la molécula. Las huellas digitales MACCS (*Molecular ACCes System*) Keys (166 bits en su versión más empleada) (Durant *et al.*, 2002) y PubChem (881 bits) son ejemplos representativos de este grupo. En las huellas digitales circulares, los bits representan subestructuras particulares de la molécula considerando la conectividad de cada átomo dentro de un diámetro (número de enlaces) determinado. Un ejemplo representativo de este grupo son las huellas digitales de conectividad extendida (ECFP, *Extended Connectivity Fingerprints*) (Rogers and Hahn, 2010). En las huellas digitales topológicas, los bits también representan subestructuras particulares de la molécula, pero en este caso se trata de subestructuras lineales de un número determinado de enlaces. Las huellas digitales Daylight son el ejemplo más representativo de este grupo. La Figura 3 muestra una representación esquemática de estos tres tipos de huellas digitales moleculares. Dado que tanto las huellas digitales circulares como las huellas digitales topológicas son generadas a partir de las subestructuras particulares de la molécula en cuestión, se les considera como huellas digitales dependientes de la molécula y son usualmente mapeadas a un número de bits predefinido, por ejemplo 1024.

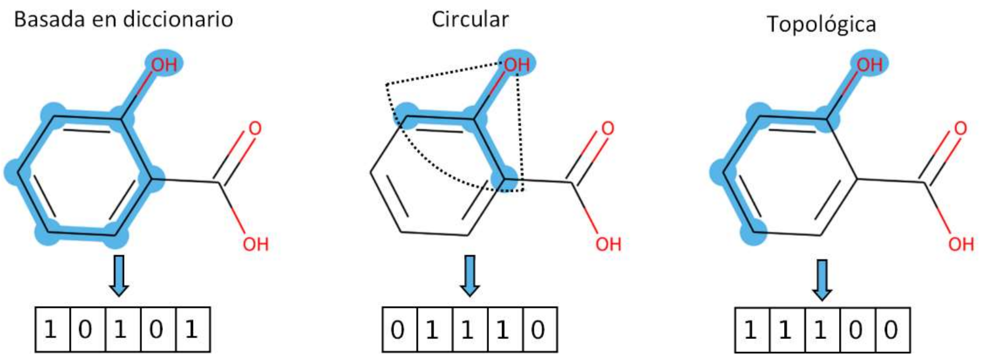


Figura 3. Huellas digitales moleculares. *Representación esquemática de distintos tipos de huellas digitales moleculares para el ácido salicílico. Por diseño, las representaciones son distintas entre sí, por lo que un bit en la misma posición representa subestructuras diferentes. En la basada en diccionario (izquierda) se representa un grupo fenol, en la circular (centro) se representa un fragmento centrado en el oxígeno con un diámetro de cuatro enlaces (o radio de dos enlaces) y en la topológica (derecha) se representa un fragmento lineal de cuatro enlaces.*

Dentro de las representaciones 3D, destacan las que se basan en el volumen de las moléculas, que representan a cada átomo de la molécula como una esfera de radio definido (típicamente su radio de van der Waals) o como una función Gaussiana centrada en las coordenadas del átomo (Grant and Pickup, 1995). Un ejemplo de este tipo de representaciones es la implementada en ROCS (*Rapid Overlay of Chemical Structures*) (Hawkins *et al.*, 2007), en la cual los átomos son además etiquetados con distintas características (donador/aceptor de puente de hidrógeno, anión, catión, anillo, etc.) para obtener una descripción más completa. Un tipo diferente de descripción 3D es la empleada por el método USR (*Ultrafast Shape Recognition*), que calcula las distribuciones de distancia de las coordenadas atómicas con respecto a cuatro puntos de referencia y emplea tres descriptores estadísticos de cada una de estas distribuciones para generar un vector de doce elementos que captura la forma de la molécula (Ballester, 2011).

Cada una de las representaciones moleculares descritas anteriormente ofrece distintas ventajas en la búsqueda de nuevas interacciones diana-ligando. Las representaciones 2D son más simples y a la vez más rápidas y fáciles de calcular, ya que solo requieren el conocimiento de la conectividad de la molécula. Por su parte, las representaciones 3D requieren, además de esto, el conocimiento o suposición de la conformación activa de la molécula de referencia y la exploración exhaustiva de las posibles conformaciones para las moléculas a comparar, lo cual es particularmente difícil cuando las moléculas son muy flexibles (por ejemplo, con más de 12 enlaces rotables). Esto hace que las comparaciones entre representaciones 3D sean, en general, más demandantes computacionalmente que las comparaciones en representaciones 2D, por lo que las representaciones 2D suelen ser más utilizadas, especialmente cuando se analizan centenas o millares de moléculas (Willett, 2006; Duan *et al.*, 2010). Además, se ha mostrado en varios estudios comparativos que su desempeño, en términos de la cantidad de nuevas interacciones diana-ligando identificadas, es similar al de las representaciones 3D. Sin embargo, al estar su aplicación limitada a la búsqueda de ligandos similares a los conocidos, las búsquedas 2D tienden a sugerir como nuevos ligandos a aquellos con conectividades muy parecidas a las de los ligandos de referencia, mientras que las búsquedas 3D, al enfocarse únicamente en la forma de las moléculas, suelen ser más útiles para sugerir ligandos con estructuras más diversas (Hu *et al.*, 2012; Finn and Morris, 2013).

II.4.1.2 Métricas de similitud

En los métodos basados en la estructura del ligando, el cálculo de la similitud entre la molécula de referencia y las moléculas a comparar es otro aspecto crucial. Se han propuesto distintas métricas para realizar dichas comparaciones, de las cuales el coeficiente de Tanimoto (T_c, *Tanimoto coefficient*) o similitud de Jaccard, es el más ampliamente usado (Jaccard, 1901; Syuib *et al.*, 2014; Bajusz *et al.*, 2015). El T_c se define de la siguiente manera:

$$T_c = \frac{C}{A + B - C}$$

En el caso de las huellas digitales moleculares, A y B representan el número de bits presentes (1) en las moléculas A y B, respectivamente, mientras que C representa el número de bits presentes en ambas moléculas. Por otro lado, en las representaciones basadas en el volumen de las moléculas, A y B representan el volumen ocupado por estas, mientras que C representa el traslape entre ellas.

II.4.2 Métodos basados en la estructura de la diana

Los métodos basados en la estructura de la diana biológica utilizan la información estructural de la diana biológica de interés para el cribado de quimiotecas. A diferencia de los métodos basados en la estructura del ligando, estos métodos no requieren del conocimiento de ningún ligando de referencia; en su lugar, solo requieren de la estructura 3D de la diana biológica, ya sea que esta haya sido determinada experimentalmente o por modelado computacional. La herramienta central empleada por los métodos basados en estructura es el acoplamiento molecular, la cual está diseñada para construir modos de interacción de los complejos diana-ligando, basándose en su geometría y propiedades fisicoquímicas (Lyne, 2002; Lionta *et al.*, 2014; Prieto-Martínez *et al.*, 2018). Un algoritmo de acoplamiento molecular consta esencialmente de dos componentes: una estrategia de búsqueda para identificar las posibles disposiciones en que el ligando y la diana se unen, y una función de puntuación que permita determinar la factibilidad de estos modos de unión y guiar así la búsqueda (Meng *et al.*, 2011). Ambos componentes se abordan en las siguientes secciones.

II.4.2.1 Estrategias de búsqueda

El acoplamiento molecular intenta modelar el proceso de reconocimiento molecular entre un ligando y su diana biológica, un proceso en el que ambas partes experimentan cambios conformacionales. De acuerdo con la forma de abordar la flexibilidad del sistema, el acoplamiento molecular contempla dos aproximaciones: acoplamiento con ligando flexible y acoplamiento con diana flexible o semi-flexible.

En acoplamiento con ligando flexible, los cambios conformacionales en el ligando son permitidos mediante la rotación de sus enlaces, mientras que la estructura de la diana permanece estática. Esta aproximación es posible debido a que la estructura de la diana es obtenida usualmente de complejos diana-ligando en los que el ligando es removido y el sitio de unión se encuentra en una conformación accesible. Aun considerando solo la flexibilidad del ligando, el espacio de búsqueda de los posibles patrones de interacción es tan grande que una búsqueda exhaustiva es prácticamente imposible, especialmente considerando que esta técnica se usa con frecuencia para analizar quimiotecas de miles o hasta millones de compuestos. Por esta razón, las herramientas disponibles incorporan diferentes algoritmos de optimización para realizar estas búsquedas en un tiempo razonable (de preferencia segundos o fracciones de segundo por molécula, especialmente cuando se analizan miles o millones de moléculas) (Guo *et al.*, 2014). Aunque la mayoría de los programas de acoplamiento molecular manejan bien la flexibilidad del ligando, modelar la flexibilidad de la diana en presencia de un ligando en tiempos cortos de cálculo es un problema abierto en este campo. Se han reportado múltiples aproximaciones para tratar este problema, que van desde la incorporación de flexibilidad únicamente en el sitio de unión (flexibilidad local) (Leach, 1994), hasta el uso de múltiples estructuras de la diana con diferentes conformaciones (flexibilidad global) (Knegtel *et al.*, 1997; Amaro *et al.*, 2018).

II.4.2.2 Funciones de puntuación

Las funciones de puntuación son el segundo elemento clave en las herramientas de acoplamiento molecular. Estas son usadas para estimar la afinidad de unión de los complejos diana-ligando encontrados en el proceso de búsqueda. Se espera que una función de puntuación robusta se desempeñe bien en cuatro tareas distintas y complementarías entre sí: obtener puntuaciones correlacionadas linealmente con datos experimentales (puntaje o *scoring*), ordenar los ligandos conocidos para una misma diana en orden de afinidad (jerarquización o *ranking*), identificar el modo de unión diana-ligando correcto (acoplamiento o *docking*) y discriminar los verdaderos ligandos de los que no lo son (cribado o *screening*) (Su *et al.*, 2019; Li *et al.*, 2014). Las funciones de puntuación más empleadas establecen la relación entre los complejos diana-ligando y su afinidad de unión mediante una forma funcional definida. Estas pueden ser clasificadas en tres grandes grupos: basadas en campos de fuerza, empíricas y basadas en el conocimiento (Liu and Wang, 2015).

Las funciones de puntuación basadas en campos de fuerza emplean la suma de los términos energéticos correspondientes a los campos de fuerza usados en dinámica molecular, centrándose usualmente en aquellos términos que describen interacciones no covalentes. Las funciones empíricas utilizan una suma de contribuciones asociadas a distintos factores relacionados con la afinidad de unión, tales como los puentes de hidrógeno y los contactos lipofílicos, donde los pesos de estas contribuciones son ajustados empíricamente para reproducir resultados experimentales conocidos. Por otro lado, las funciones basadas en el conocimiento constan de una suma de términos derivados del análisis estadístico de complejos diana-ligando conocidos, bajo la suposición de que los estados más frecuentes en dichos complejos (entendidos como contactos, distancias de interacción, etc.), representan estados de baja energía en los mismos y que estos siguen la distribución de Boltzmann. A pesar de la gran cantidad de funciones de puntuación existentes, ninguna se desempeña mejor que las demás en las cuatro tareas descritas previamente, por lo que el desarrollo de nuevas funciones es un área de investigación activa.

II.4.3 Impacto de la inteligencia artificial

La inteligencia artificial es un conjunto de herramientas diseñadas para imitar, en aspectos particulares, la inteligencia humana. Estás herramientas involucran el uso de sistemas computacionales que puedan manejar grandes cantidades de datos de entrada, y aprender de ellos para posteriormente aplicar ese conocimiento en el procesamiento de nuevos datos. Conceptualmente, la inteligencia artificial se ha estado usando en química desde la década de los 60 (Gasteiger, 2020). Sin embargo, gracias a los avances tecnológicos en los últimos años, su uso ha incrementado en múltiples sectores, donde la quimiogenómica y el diseño de fármacos no son la excepción (Fleming, 2018; Paul *et al.*, 2020).

El constante incremento en el tamaño de las bases de datos quimiogenómicas (Mendez *et al.*, 2019; Kim *et al.*, 2019) hace que su manejo y análisis para la extracción de conocimiento sean cada vez más complejos. Esto ha llevado a la implementación de conceptos y metodologías de la inteligencia artificial para su procesamiento, pues estas permiten el análisis eficiente de estos datos para la generación de modelos predictivos precisos en tiempos relativamente cortos.

Dentro de las herramientas de la inteligencia artificial utilizadas en la quimiogenómica, destaca el aprendizaje automático (Vamathevan *et al.*, 2019; Patel *et al.*, 2020). Este puede ser entendido como el uso de algoritmos capaces de reconocer patrones en datos que han sido previamente clasificados (numérica o categóricamente), para la posterior clasificación de nuevos datos. Dentro de este tipo de algoritmos se encuentran algunos como los bosques aleatorios (RF, *Random Forest*) (Tin Kam Ho, 1998), las máquinas de soporte vectorial (SVM, S*upport Vector Machines*) (Cortes and Vapnik, 1995) y las redes neuronales artificiales (ANN, A*rtificial Neural Networks*) (Hopfield, 1982), los cuales han sido utilizados ampliamente para construir modelos de relaciones cuantitativas estructura-actividad (QSAR, *Quantitative Structure-Activity Relationships*). Con una implementación más reciente se encuentra el aprendizaje profundo, un campo particular del aprendizaje automático basado en las ANN, cuyas aplicaciones han sido exploradas en múltiples áreas del descubrimiento de fármacos (Chen *et al.*, 2018).

El aprendizaje automático ha impactado en el desarrollo de los modelos predictivos empleados en la quimiogenómica computacional (Zhang *et al.*, 2017; Thafar *et al.*, 2019). Sus aportaciones abarcan tanto los métodos basados en la estructura del ligando, como los métodos basados en la estructura de la diana.

Dentro del primer grupo, destaca el entrenamiento de este tipo de algoritmos a partir de huellas digitales moleculares para la generación de modelos predictivos de interacciones diana-ligando, los cuales han representado una mejora a los métodos clásicos (Mayr *et al.*, 2018). En el segundo grupo, sobresale el entrenamiento de estos algoritmos en representaciones de complejos proteína-ligando para la obtención de nuevas funciones de puntuación para su uso en el acoplamiento molecular. Los resultados obtenidos por estos modelos han mostrado una mejor correlación con los datos experimentales que las obtenidas por funciones de puntuación clásicas (Li *et al.*, 2020).

II.5 Herramientas web para la navegación de espacios quimiogenómicos

Cualquiera de los métodos descritos en la sección II.4 pueden ser utilizados con dos enfoques distintos, pero complementarios entre sí. A la búsqueda de compuestos que presenten interacciones favorables con una diana biológica en particular se le denomina frecuentemente cribado virtual, mientras que a la búsqueda de las dianas biológicas con las que interactúa un compuesto dado, se le conoce como cribado virtual inverso (Sydow *et al.*, 2019; Yang *et al.*, 2020). El cribado virtual puede entenderse como un proceso de filtrado o selección, ya sea de moléculas pequeñas (cribado directo) o de macromoléculas (cribado inverso) y son una etapa previa a las pruebas biológicas correspondientes. Estos enfoques se representan esquemáticamente en la Figura 4. Actualmente existe una gran cantidad de herramientas que permiten realizar estos estudios, desde los programas computacionales de acceso libre como RDKit, AutoDock (Morris *et al.*, 2009), AutoDock VINA (Trott and Olson, 2010) y rDock (Ruiz-Carmona *et al.*, 2014), hasta los disponibles comercialmente como MOE (Chemical Computing Group, 2018) y Glide (Friesner *et al.*, 2006), por mencionar algunos ejemplos. Cabe resaltar que, en la práctica, el proceso de filtrado (de compuestos o dianas) se hace para reducir tiempos y costos de las pruebas experimentales, pero no buscan reemplazarlas. De tal forma que ambas aproximaciones se acoplan en un proceso iterativo y sinérgico que involucra típicamente la realización de más de un ciclo, como se ilustra en la Figura 2.

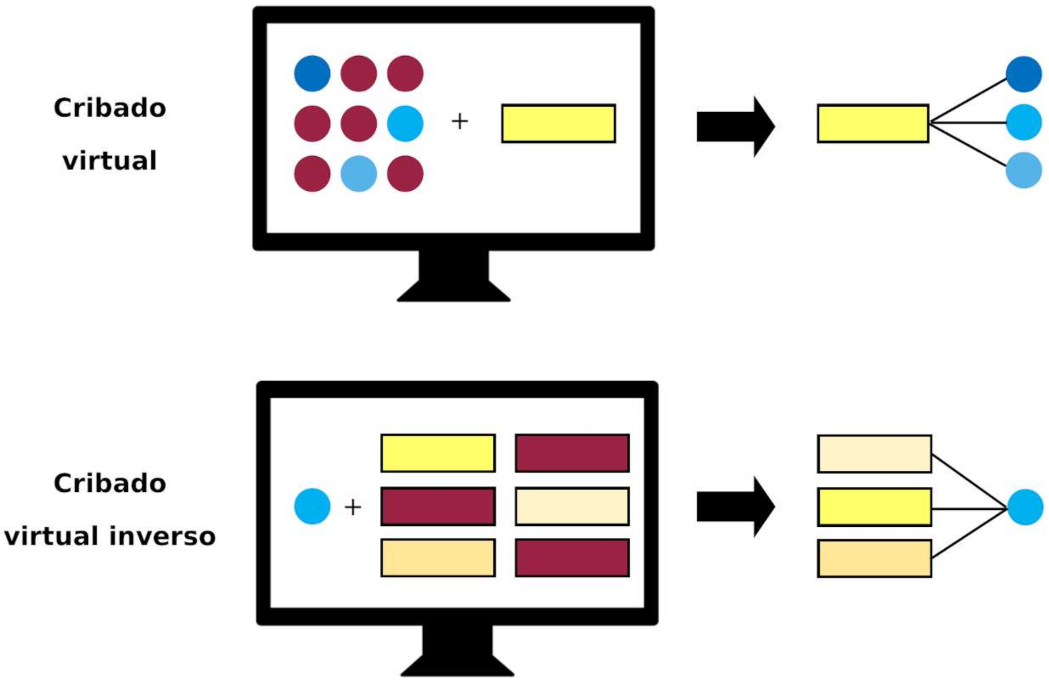


Figura 4. Representación esquemática de los enfoques principales usados en la quimiogenómica computacional. *Los círculos representan compuestos y los rectángulos representan dianas biológicas. El cribado virtual (directo e inverso) está acoplado a pruebas experimentales en un ciclo iterativo, esquematizado en la Figura 2.*

Las herramientas de acceso libre son las más utilizadas por el sector académico y algunos institutos de investigación tanto públicos como privados. Dentro de estos recursos computacionales, aquellos accesibles a través de la web resultan particularmente útiles, pues no requieren de su instalación en un equipo de cómputo por parte del usuario, son accesibles desde cualquier dispositivo con conexión a internet y generalmente están diseñadas para ser fáciles de usar. Esto se ve reflejado en la gran cantidad de aplicaciones web que han sido desarrolladas en los últimos años (Singh *et al.*, 2020), las cuales se abordan en las siguientes secciones. A pesar de su amplio uso, este tipo de implementaciones presentan algunas características que podrían ser consideradas una desventaja. Uno de los principales problemas es el hecho de que su mantenimiento y actualización depende de la disponibilidad de recursos por parte del grupo que las desarrolla, por lo que su presencia en línea difícilmente puede ser asegurada a largo plazo. Otro aspecto importante para tener en cuenta por parte del usuario es la privacidad de sus datos, pues al ser servicios públicos, la información introducida puede ser almacenada por el servidor, aun si se trata de datos sujetos a propiedad intelectual. Se debe considerar también que los recursos computacionales disponibles en este tipo de servicios son limitados, por lo que generalmente permiten el procesamiento de pocos datos. Finalmente, su facilidad de uso puede convertirlas en “cajas negras” para el usuario, por lo que es recomendable una adecuada documentación acerca del desarrollo de estas herramientas previo a su uso.

II.5.1 Cribado virtual

Las herramientas web disponibles para realizar cribado virtual permiten la búsqueda de nuevos compuestos con potencial interacción sobre una diana biológica de interés mediante la exploración de quimiotecas. La Tabla 1 integra los recursos disponibles para realizar cribado virtual basado en la estructura del ligando. Las principales diferencias entre estas implementaciones radican en el tipo de representación utilizado para describir a los compuestos (2D o 3D) y las quimiotecas que pueden explorar, las cuales van desde quimiotecas de relevancia quimiogenómica como DrugBank, ChEMBL y PubChem, hasta quimiotecas comerciales como ZINC (Irwin *et al.*, 2020).

Por otro lado, las aproximaciones basadas en la estructura de la diana permiten la exploración de quimiotecas mediante el acoplamiento molecular de sus compuestos en la estructura 3D de la diana biológica de interés. Al tener que procesar múltiples moléculas, la estrategia de elección es el acoplamiento molecular con diana rígida y ligando flexible, descrito en la sección II.4.2. Las diferencias principales entre las implementaciones disponibles radican en el programa computacional que utilizan para realizar el acoplamiento molecular, así como en las quimiotecas disponibles para su exploración, donde a diferencia de las implementaciones basadas en el ligando, algunas permiten el uso de quimiotecas del usuario para efectuar la búsqueda. La Tabla 2 resume recursos web disponibles para realizar cribado virtual basado en la estructura de la diana.

Tabla 1. Herramientas web para cribado virtual basado en la estructura del ligando.

Herramienta	URL*	Representación	Bases de datos	Referencia
ChemMine Tools	http://chemmine.ucr.edu/	2D	ChEMBL PubChem	(Backman *et al.*, 2011)
ZincPharmer	http://zincpharmer.csb.pitt.edu/	3D	ZINC	(Koes and Camacho, 2012)
Rchempp	http://shiny.bioinf.jku.at/Analoging/	2D	ChEMBL DrugBank	(Klambauer *et al.*, 2015)
Pharmit	http://pharmit.csb.pitt.edu	3D	ChEMBL PubChem Comerciales	(Sunseri and Koes, 2016)
SwissSimilarity	http://www.swisssimilarity.ch/	2D + 3D	ChEMBL DrugBank Comerciales	(Zoete *et al.*, 2016)
BRUSELAS	http://bio-hpc.eu/software/Bruselas/	3D	ChEMBL DrugBank Propias	(Banegas-Luna *et al.*, 2019)

*Disponibles en enero 2021

Tabla 2. Herramientas web para cribado virtual basado en la estructura de la diana.

Herramienta	URL*	Programa	Bases de datos	Referencia
DOCK Blaster	http://blaster.docking.org/	DOCK	ZINC	(Irwin et al., 2009)
e-LEA3D	https://chemoinfo.ipmc.cnrs.fr/LEA3D/index.html	PLANTS	FDA Personalizada	(Douguet, 2010)
MTiOpenScreen	https://bioserv.rpbs.univ-parisdiderot.fr/services/MTiOpenScreen/	AutoDock VINA	Propias Personalizada	(Labbé et al., 2015)
DockThor	https://dockthor.lncc.br/v2/	Propio	e-Drug3D Personalizada	(da Silveira et al., 2019)
EasyVS	https://easyvs.unifei.edu.br/	AutoDock VINA	ChEMBL DrugBank Personalizada	(Pires et al., 2020)

*Disponibles en enero 2021

II.5.2 Cribado virtual inverso

Los recursos web disponibles para realizar cribado virtual inverso permiten evaluar a un compuesto de interés por múltiples modelos asociados a distintas dianas biológicas. Dentro de las herramientas basadas en la estructura del ligando, además de la forma de describir a los compuestos para la construcción de los modelos, su diferencia principal reside en el número y tipo de dianas biológicas que se consideran. La Tabla 3 resume recursos web disponibles actualmente para llevar a cabo cribado virtual inverso basado en la estructura del ligando. En ella se puede apreciar que TargetNet (Yao *et al.*, 2016) incluye únicamente 623 proteínas, que corresponden a las presentes en Binding Database, mientras que el resto de las implementaciones incluyen las proteínas humanas presentes en distintas versiones de ChEMBL, siendo SEA (Keiser *et al.*, 2007) el recurso que considera un mayor número de proteínas, al no limitarse a proteínas humanas. Sorprendentemente, ninguno de estos recursos incluye los datos de la versión más reciente de ChEMBL (la versión 27), lo cual hace enfatizar nuevamente, que uno de los principales problemas de este tipo de herramientas es la falta de actualización de los modelos ante el constante incremento del tamaño de las bases de datos quimiogenómicas. Cabe mencionar que, además de las implementaciones presentadas en la Tabla 3, existen algunas enfocadas en grupos particulares de dianas biológicas, como DIA-DB (Pérez-Sánchez *et al.*, 2020), enfocada en dianas relacionadas con la diabetes y WDL-RF (Wu *et al.*, 2018), enfocada en GPCRs.

Tabla 3. Herramientas web para cribado virtual inverso basado en la estructura del ligando.

Herramienta	URL*	Representación / Algoritmo	Dianas	Referencia
SEA	http://sea.bkslab.org/	2D / similitud	13377 (ChEMBL 26)	(Keiser *et al.*, 2007)
TargetNet	http://targetnet.scbdd.com/calcnet/index/	2D / aprendizaje automático	623 (Binding Database)	(Yao *et al.*, 2016)
RFQSAR	http://rfqsar.kaist.ac.kr	2D / aprendizaje automático	1288 (ChEMBL 23)	(Lee *et al.*, 2017)
HitPickV2	http://mips.helmholtzmuenchen.de/HitPickV2/target_prediction.jsp	2D / similitud	2739 (ChEMBL 22 + Binding Database)	(Hamad *et al.*, 2019)
MuSSel	http://mussel.uniba.it:5000/	2D / similitud	3357 (ChEMBL 25)	(Montaruli *et al.*, 2019)
PPB2	http://gdbtools.unibe.ch:8080/PPB/	2D / aprendizaje automático	1720 proteínas (ChEMBL 22)	(Awale and Reymond, 2019)
SwissTargetPrediction	http://www.swisstargetprediction.ch/	2D + 3D / similitud	3068 (ChEMBL 23)	(Daina *et al.*, 2019)

*Disponibles en enero 2021

En el caso de las herramientas basadas en la estructura de la diana, sus diferencias radican tanto en el programa computacional de acoplamiento molecular utilizado, como en el número de estructuras de dianas sobre las cuáles realizan este acoplamiento. El número de recursos disponibles para efectuar este tipo de cálculos es más reducido en comparación con el grupo anterior, en parte debido a que los cálculos de acoplamiento molecular son computacionalmente más demandantes que los cálculos basados en la estructura del ligando. Este tipo de recursos se resumen en la Tabla 4.

Las estructuras de las dianas biológicas utilizadas en estas implementaciones son normalmente extraídas del Banco de Datos de Proteínas (PDB, *Protein Data Bank*), por lo que el número de dianas biológicas disponibles depende de los criterios seleccionados para dicha selección. idTarget (Wang *et al.*, 2012), por ejemplo, emplea 3046 sitios de unión en lugar de proteínas completas. Por su parte, ACID (Wang *et al.*, 2019) utiliza únicamente 831 estructuras de proteínas consideradas de relevancia terapéutica, mientras que CRDS (Lee and Kim, 2019) no emplea este criterio de exclusión. Al igual que en el grupo anterior, existen implementaciones de métodos basados en la estructura de la diana enfocadas en un número más reducido de dianas biológicas, como GOMoDO (Sandal *et al.*, 2013) y GUT-DOCK (Pasznik *et al.*, 2019), enfocadas en GPCRs, y EDMON (Schneider *et al.*, 2020), enfocada en receptores nucleares. Otro ejemplo es COVID-19 Docking Server, que permite el acoplamiento de moléculas pequeñas, péptidos y anticuerpos sobre 27 potenciales dianas del SARS-CoV2 (Kong *et al.*, 2020).

Tabla 4. Herramientas web para cribado virtual inverso basado en la estructura de la diana.

Herramienta	**URL***	**Software**	**Dianas**	**Referencia**
idTarget	http://idtarget.rcas.sinica.edu.tw	MEDock	3046	(Wang *et al.*, 2012)
ACID	http://chemyang.ccnu.edu.cn/ccb/server/ACID/	AutodDock VINA LeDock PLANTS PSOVINA	831	(Wang *et al.*, 2019)
CRDS	http://pbil.kaist.ac.kr/CRDS	Gold AutoDock VINA LeDock	5254	(Lee and Kim, 2019)

*Disponibles en enero 2021

III. PLANTEAMIENTO DEL PROBLEMA

La quimiogenómica es la disciplina encargada de identificar las interacciones entre moléculas pequeñas y dianas biológicas, un aspecto esencial en el desarrollo de fármacos. La identificación de estas interacciones debe realizarse experimentalmente. Sin embargo, la exploración experimental de todas las posibles interacciones entre compuestos y dianas biológicas es prácticamente imposible de alcanzar por los altos costos y tiempos que implicaría determinar la intersección de los espacios químicos y biológicos relevantes en el diseño de fármacos. Los métodos computacionales han mostrado ser útiles en la predicción de estas interacciones, reduciendo así tanto el tiempo como los recursos invertidos en la realización de experimentos. Estos métodos utilizan la información conocida acerca de las SAR entre compuestos y dianas biológicas para la construcción de modelos que permitan el planteamiento de nuevas hipótesis de interacción. Actualmente, existe una gran variedad de métodos computacionales disponibles para llevar a cabo dichas predicciones, sin embargo, la información acerca de las SAR entre compuestos y dianas biológicas está en constante aumento, por lo que nuevas herramientas y la actualización o refinamiento de las ya existentes que exploten esta creciente disponibilidad de datos deben ser desarrolladas, validadas y documentadas. En beneficio de la “ciencia abierta” (*open science*) y la reproducibilidad de datos, también es conveniente hacer accesibles los recursos computacionales nuevos u optimizados a la comunidad. Esto, no solamente para aplicaciones en la investigación, sino también para que ayude en la difusión, enseñanza y uso adecuado de los métodos computacionales empleados en el descubrimiento y desarrollo de fármacos.

IV. HIPÓTESIS

El análisis computacional de las relaciones estructura-actividad entre compuestos y dianas biológicas disponibles públicamente, permitirá el desarrollo de herramientas computacionales de acceso libre que asistan la búsqueda y selección de nuevos compuestos con potencial actividad sobre dianas biológicas de interés terapéutico.

V. OBJETIVOS

V.1. Objetivo general

Desarrollar un conjunto de herramientas computacionales para asistir la identificación de nuevas interacciones entre moléculas pequeñas y dianas biológicas.

V.2. Objetivos específicos

- Desarrollar nuevas metodologías para el cribado virtual de compuestos sobre diversas dianas biológicas de interés terapéutico.

- Validar el desempeño de las metodologías en bases de datos de compuestos con actividad biológica reportada.

- Implementar las metodologías desarrolladas en un servidor de la Facultad de Química de la UNAM asignado al grupo de investigación de Diseño de Fármacos Asistido por Computadora (DIFACQUIM), como parte de las herramientas libres disponibles en *D-Tools: DIFACQUIM Tools for Cheminformatics* (https://www.difacquim.com/d-tools/).

- Aplicar la nueva plataforma en proyectos de investigación que se tienen como parte de colaboraciones entre el grupo de investigación DIFACQUIM y grupos de investigación experimental, tanto en México como en el extranjero.

VI. RESULTADOS

En esta tesis se presenta el desarrollo, validación e implementación de tres recursos computacionales de acceso libre para la predicción de potenciales interacciones entre moléculas pequeñas y dianas biológicas. Las herramientas nuevas abarcan aspectos asociados a metodologías basadas tanto en la estructura del ligando como en la estructura de la diana, y emplean conceptos de las disciplinas bioinformática y quimioinformática. Los resultados de este trabajo y su discusión están organizados de la siguiente manera:

- El capítulo VII aborda los métodos de representación y procesamiento de estructuras, tanto de los ligandos como de las dianas biológicas empleadas, para su análisis y el desarrollo posterior de las herramientas presentadas.

- En el capítulo VIII, se discute el concepto de *Statistical-Based Database Fingerprint* (SB-DFP) (Sánchez-Cruz and Medina-Franco, 2018), una huella digital para bases de datos de compuestos. En dicho capítulo se detalla la validación de su uso en el cribado virtual basado en el ligando, así como sus aplicaciones en el estudio de relaciones entre dianas biológicas y la representación visual del espacio químico (Chávez-Hernández *et al.*, 2020; Sessions *et al.*, 2020).

- El capítulo IX detalla la validación y comparación de modelos de aprendizaje automático para el cribado virtual inverso enfocado en dianas epigenéticas, así como la implementación de los mejores modelos en una aplicación web denominada *Epigenetic Target Profiler* (ETP) (Sánchez-Cruz and Medina-Franco, 2021a, 2021b).

- En el capítulo X se introduce *Exteneded-Connectivity Interaction Features* (ECIF) (Sánchez-Cruz *et al.*, 2020), una descripción de complejos proteína-ligando y su aplicación en la construcción de una función de puntuación basada en aprendizaje automático para la predicción de su afinidad de unión.

La tesis finaliza con el capítulo XI que aborda las conclusiones generales del proyecto, y el capítulo XII que detalla las perspectivas de este. Los resultados presentados en la tesis están en artículos publicados o en proceso de revisión de pares. Los capítulos siguientes incluyen una breve descripción de los resultados principales de la tesis, mientras que las publicaciones correspondientes están en el ANEXO I. Además, los resultados de este trabajo contribuyeron al desarrollo de otros estudios publicados, cuya primera página se incluye en el ANEXO II.

VII. ESTANDARIZACIÓN DE ESTRUCTURAS

VII.1 Significancia

Existen múltiples formatos de archivos que permiten almacenar la estructura de las moléculas. Las bases de datos quimiogenómicas en el dominio público incluyen moléculas obtenidas a partir de distintos orígenes, lo cual implica el manejo de su estructura mediante distintos tipos de archivos. Esto hace necesario el uso de metodologías que permitan la identificación correcta de moléculas iguales provenientes de distintas fuentes de información. Para abordar este problema se utiliza un proceso denominado estandarización de estructuras, el cual busca generar representaciones de las moléculas que permitan identificarlas sin ambigüedad. Considerando el tamaño actual de las bases de datos quimiogenómicas, la estandarización manual de las estructuras de los ligandos y las dianas biológicas es poco eficiente, por lo que se requiere de algoritmos automáticos para realizar este proceso.

La estandarización de estructuras es un paso fundamental para el análisis de las estructuras de moléculas, particularmente en el desarrollo de modelos predictivos de interacciones diana-ligando. Esto se debe a que los métodos de representación empleados en estos modelos son sensibles a las diferencias en las estructuras de entrada. Por esta razón, aspectos como la tipificación de átomos, la forma tautomérica y el estado de protonación de los grupos funcionales en las moléculas deben ser tratados de tal forma que se garantice que una molécula sea representada siempre de la misma manera. En la sección VII.2 se describen los protocolos de estandarización de estructuras empleados para la construcción de los recursos computacionales presentados en los capítulos VIII, IX y X, mientras que la sección VII.3 aborda aspectos relacionados a su disponibilidad y aplicación en la realización de otros estudios quimioinformáticos.

VII.2 Metodología

Los capítulos VIII y IX describen herramientas basadas en la estructura del ligando, donde solo la estandarización de la estructura de los ligandos es necesaria. Para el desarrollo de estos recursos computacionales se emplearon descripciones 2D, por lo que las moléculas fueron almacenadas y procesadas usando el formato SMILES (*Simplified Molecular-Input Line-Entry System*) (Weininger, 1988), el cual es una notación en forma de cadena de texto para describir la estructura de moléculas pequeñas.

En el capítulo X se describe un recurso basado en la estructura de la diana, por lo que tanto la estructura de las dianas como de los ligandos deben ser estandarizadas. En este caso, para su almacenamiento y procesamiento posterior se usaron los formatos SDF (S*tructure Data File*) y PDB, para ligandos y dianas, respectivamente. A diferencia del formato SMILES, que sólo preserva la conectividad de la molécula, los formatos SDF y PDB permiten almacenar la disposición espacial de todos los átomos que la conforman. Los procesos empleados para la estandarización de estructuras en cada uno de los formatos empleados se describen en las siguientes secciones.

VII.2.1 Estandarización 2D

Para el desarrollo de las herramientas presentadas en los capítulos VIII y IX, cada compuesto almacenado en formato SMILES fue estandarizado empleando un algoritmo propio escrito en Python, usando dos módulos quimioinformáticos de código abierto disponibles para dicho lenguaje de programación: RDKit y MolVS. El proceso empleado fue el siguiente: se verificó el número de componentes en cada molécula, en aquellas con más de un componente (sales), solo el componente más grande fue preservado. Si los compuestos poseían elementos diferentes a H, B, C, N, O, F, Si, P, S, Cl, Se, I y Br, estos eran descartados. En las moléculas con una carga formal diferente de cero, se añadieron o removieron hidrógenos a sus grupos neutralizables hasta alcanzar una carga formal de cero (o lo más cercana posible). Posteriormente, las moléculas se reionizaron, de tal manera que sus grupos ácidos más débiles (de acuerdo con las especificaciones de la FDA) permanecieran no ionizados. Se construyó una forma tautomérica canónica (basada en reglas estructurales predefinidas) para cada compuesto. Finalmente, un SMILES canónico sin información estereoquímica fue generado y utilizado para los procesamientos posteriores.

VII.2.2 Estandarización 3D

Para la generación del recurso computacional presentado en el capítulo X, los compuestos fueron almacenados inicialmente en formato MOL2. Este tipo de formato preserva la información 3D de los compuestos, lo que incluye su estereoquímica y estado de protonación. La estandarización de estas estructuras se realizó utilizando X-Tool (Wang *et al.*, 2002) y Standardizer, JChem 20.11.0, 2020, ChemAxon, dos programas computacionales de acceso libre para uso académico. En este caso, únicamente las moléculas con un solo componente fueron consideradas, excluyendo además a aquellas con átomos diferentes a H, C, N, O, F, P, S, Cl, I y Br. X-Tool fue utilizado para identificar y tipificar a los átomos de acuerdo al campo de fuerza Tripos (Clark *et al.*, 1989). Posteriormente, Standardizer fue empleado para añadir los hidrógenos faltantes (en caso de haberlos), identificar los átomos aromáticos y generar los archivos SDF que serían analizados posteriormente. Por otro lado, las dianas involucradas en el desarrollo de esta herramienta estuvieron limitadas a proteínas, almacenadas en formato PDB, un tipo de archivo ampliamente usado en la bioinformática estructural. Las únicas consideraciones respecto a estas moléculas fueron que sus estructuras contuvieran solamente los 20 aminoácidos naturales y que todos sus átomos pesados tuvieran coordenadas bien definidas, lo cual no requiere de un programa computacional en particular. Este proceso fue verificado mediante código propio escrito en Python.

VII.3 Resultados y discusión

Todas las metodologías empleadas para la estandarización de las moléculas fueron implementadas utilizando tanto programas de cómputo de acceso libre para la academia (X-Tool y Standardizer), como algoritmos escritos en Python, desarrollados como parte de esta tesis. Estos últimos se hicieron de acceso libre y se encuentran disponibles en https://github.com/DIFACQUIM/IFG_General.

El algoritmo de estandarización 2D desarrollado en este trabajo se ha aplicado en diversos estudios adicionales a los presentados en los capítulos VIII y IX. Esto se debe a que es fácil de implementar para usuarios nuevos en la programación en Python y a que con este es posible procesar miles de compuestos en cuestión de minutos usando una computadora de escritorio, facilitando el análisis de quimiotecas de cientos de miles o hasta millones de compuestos. El trabajo donde este algoritmo se usó por primera vez fue el análisis del contenido de grupos funcionales en quimiotecas de productos naturales y compuestos con actividad biológica (Sánchez-Cruz *et al.*, 2019), incluido en el ANEXO 1. Posteriormente, el uso de este algoritmo se extendió al análisis de diversidad estructural (Tran *et al.*, 2020), a la construcción de bibliotecas de fragmentos moleculares a partir de quimiotecas relevantes para el diseño de fármacos (Chávez-Hernández *et al.*, 2020), y al cribado virtual de quimiotecas en la búsqueda de inhibidores de la proteasa principal del SARS-CoV2 (Santibáñez-Morán *et al.*, 2020); trabajos cuya portada se incluye en el ANEXO 2.

En el caso de los procedimientos empleados en la estandarización 3D de las moléculas, su uso depende de la disponibilidad de las licencias académicas de los programas empleados. Por lo que la documentación presentada en este capítulo tiene la intención de proveer al lector (y usuario potencial) con los procedimientos a seguir para el uso adecuado de la herramienta introducida en el capítulo X, lo que a la vez facilitará la reproducibilidad de los datos presentados en dicho capítulo.

VIII. HUELLAS DIGITALES DE QUIMIOTECAS BASADAS EN ESTADÍSTICA

VIII.1 Significancia

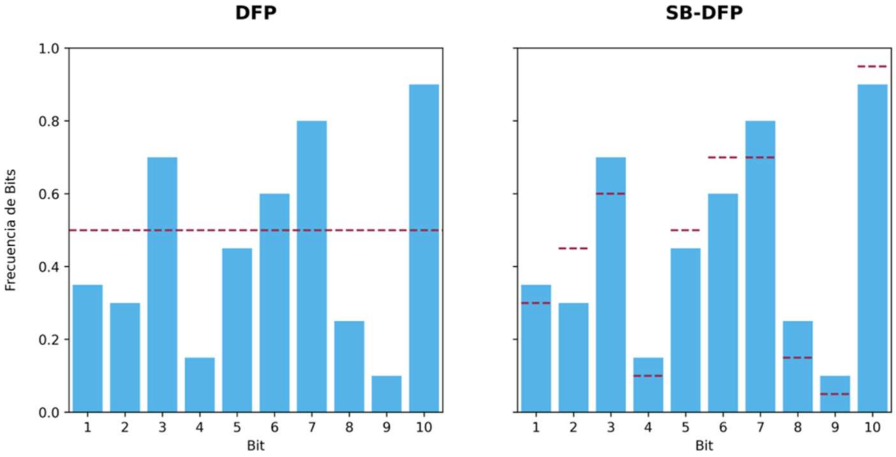
Dentro de los métodos basados en la estructura del ligando para la predicción de interacciones diana-ligando, las búsquedas por similitud son una de las estrategias más usadas. En ellas, se requiere del conocimiento de un ligando de referencia que se usa como plantilla para el cribado virtual de quimiotecas en la búsqueda de compuestos similares a éste, con la hipótesis de que compuestos similares presentaran interacciones similares (con lo misma diana biológica). Con el constante incremento en el tamaño de las bases de datos quimiogenómicas, es común que se conozca más de un ligando para una diana. Cuando esto ocurre, múltiples aproximaciones han sido propuestas para llevar a cabo el cribado virtual, las cuáles asignan un valor de similitud único para cada uno de los compuestos cribados con respecto a la quimioteca de ligandos conocidos. Este valor de similitud puede ser calculado, ya sea como el promedio de la similitud entre el compuesto cribado y los ligandos de referencia, o como el valor máximo de estas comparaciones (vecino más cercano) (Hert *et al.*, 2006; Willett, 2006).

Otra estrategia ha sido la generación de una representación única que capture las características comunes en los ligandos de referencia, para su uso como plantilla en las búsquedas por similitud. Una de las primeras aproximaciones empleadas para tal fin fue la “huella digital modal” (Shemetulskis *et al.*, 1996), una huella digital que captura las características más representativas del conjunto de compuestos de referencia. Esta representación es construida mediante el análisis de los bits presentes en las huellas digitales moleculares para los ligandos conocidos, de tal forma que contenga únicamente los bits encontrados con más frecuencia. El criterio de frecuencia usado para establecer la inclusión o no de un bit dado es variable. La representación con mejores resultados en el cribado virtual incluye solo los bits presentes en más del 50% de los ligandos de referencia (Hert *et al.*, 2004; Duan *et al.*, 2010), y se le ha denominado huella digital de bases de datos (DFP, *database fingerprint*) (Fernández-De Gortari *et al.*, 2017). Esta aproximación involucra la realización de un menor número de comparaciones al llevar a cabo una búsqueda por similitud. Sin embargo, su desempeño para identificar nuevas interacciones diana-ligando es inferior al de las aproximaciones que se basan en las representaciones individuales de los ligandos.

En este capítulo, se introduce el concepto de *Statistical-Based Database Fingerprint* (SB-DFP) (Sánchez-Cruz and Medina-Franco, 2018), una representación alterna a la DFP que considera la comparación estadística entre los bits presentes en una quimioteca de estudio y una de referencia para su construcción. La sección VIII.2 incluye los detalles para su construcción en distintas aplicaciones, mientras que la sección VIII.3 da ejemplos de su aplicación.

VIII.2 Metodología

La construcción de la SB-DFP para una quimioteca de interés requiere del análisis de los bits presentes en las huellas digitales moleculares de los compuestos que la forman. La frecuencia de ocurrencia de cada bit en la quimioteca de interés es comparada con la frecuencia de ocurrencia del mismo bit en una quimioteca de referencia, de tal forma que los bits “1” en la SB-DFP de la quimioteca de estudio serán aquellos cuya frecuencia sea significativamente superior a la observada en la quimioteca de referencia. Una representación esquemática de la SB-DFP y su diferencia con la DFP se muestran en la Figura 5. Se pueden emplear múltiples huellas digitales moleculares y quimiotecas de referencia para la construcción de la SB-DFP, lo cual depende de la aplicación requerida.



[0 0 1 0 0 1 1 0 0 1] [1 0 1 1 0 0 1 1 1 0]

Figura 5. Representaciones condensadas de quimiotecas para una huella digital hipotética de 10 bits. *Debajo de cada gráfica se muestra la huella digital generada bajo las aproximaciones de la DFP y la SB-DFP.*

VIII.2.1 Aplicación en el cribado virtual

La SB-DFP puede ser utilizada como plantilla en búsquedas por similitud para la identificación de compuestos con interacción sobre una diana biológica. Su desempeño en esta tarea fue evaluado sobre quimiotecas asociadas a 28 dianas relacionadas con la regulación epigenética, para las cuales se conocen al menos 50 ligandos. Para cada una de las dianas se seleccionaron al azar diez de sus ligandos conocidos. Con ellos se construyó la SB-DFP y se realizaron búsquedas por similitud con la intención de identificar los ligandos restantes dentro de un conjunto de 1 millón de compuestos usados como señuelos. Este proceso se repitió 100 veces para cada diana y los resultados fueron comparados con los obtenidos mediante búsquedas por similitud empleando la DFP y el método de vecino más cercano (1-NN, *1-Nearest-Neighbour*). Los estudios fueron realizados usando dos huellas digitales moleculares diferentes como representación de los compuestos:

MACCS Keys (166 bits) y ECFP de diámetro 2 (ECFP4 - 2048 bits). En el caso de la SB-DFP, para la comparación de las frecuencias de ocurrencia de los bits se usó como conjunto de referencia una quimioteca con más de 15 millones de compuestos disponibles comercialmente en ZINC (Sterling and Irwin, 2015). Los tres tipos de búsquedas fueron realizados utilizando como métrica de similitud el T_c y sus resultados fueron comparados en términos de la tasa de recuperación (RR, *recovery rate*) de los ligandos conocidos y el área bajo la curva (AUC, *area under the curve*) característica operativa del receptor (ROC, *receiver operating characteristic*).

VIII.2.2 Estudio de relaciones entre dianas

Una SB-DFP puede ser considerada como una representación en forma de huella digital de una diana biológica, siempre que esta sea construida a partir de sus ligandos conocidos. Esta representación implica la posibilidad de comparar distintas dianas biológicas utilizando métricas de similitud empleadas en la comparación de ligandos. Como ejemplo de esta aplicación se seleccionaron 136 dianas epigenéticas para las cuales se conocen al menos 10 ligandos, se construyó una SB-DFP para cada diana utilizando ECFP4 (2048 bits) como representación de los compuestos, y ChEMBL como conjunto de referencia para la comparación de frecuencias de bits. Utilizando la matriz de similitud entre las SB-DFPs, calculada mediante el T_c, se construyó un árbol de mínima cobertura (Probst and Reymond, 2020). Esta representación muestra las SB-DFP de cada diana como nodos conectados por medio de ramas, que representan la relación entre ellas, y en donde las dianas similares se muestran cercanas.

VIII.2.3 Representación del espacio químico

Una tercera aplicación de la SB-DFP es la representación del espacio químico ocupado por un conjunto de compuestos. Para ello, se requiere de la construcción de tantas SB-DFPs como dimensiones se deseen representar, a partir de quimiotecas diferentes. Una primera aproximación de este tipo de representaciones involucró la generación de una SB-DFP para representar una quimioteca con más de 190 mil productos naturales (Sorokina and Steinbeck, 2020; Sorokina *et al.*, 2021), y otra para representar una quimioteca compuesta por más de 15 millones de compuestos sintéticamente accesibles, de tal manera que cualquier compuesto pudiera ser mapeado en dos dimensiones utilizando su similitud, calculada como el T_c, a cada una de estas representaciones. La representación usada para los compuestos fue ECFP4 (1024 bits) y cada una de las quimiotecas fue utilizada como referencia para la construcción de la SB-DFP de la otra. Ambas SB-DFPs fueron generadas usando el 60% de los datos presentes en las quimiotecas, mientras que el 40% restante se usó para generar la representación del espacio químico, al igual que una base de datos de compuestos con actividad biológica reportada y bibliotecas de fragmentos moleculares generados a partir de ellas.

VIII.3 Resultados y discusión

VIII.3.1 Cribado virtual

El código usado para la construcción de las SB-DFPs se encuentra disponible libremente en https://github.com/DIFACQUIM/SB-DFP. Los resultados obtenidos al utilizar la SB-DFP, la DFP y el algoritmo 1-NN en búsquedas por similitud sobre 28 quimiotecas asociadas a dianas epigenéticas fueron comparados. Las búsquedas se realizaron con dos huellas digitales moleculares de distinto diseño, tanto para la construcción de las representaciones condensadas (DFP y SB-DFP) como para las búsquedas por similitud. Se encontró que el método que usa la SB-DFP construida con ECFP presentó el mejor desempeño en estas búsquedas, tanto en términos de RR como de AUC, obteniendo una RR promedio de 50.2%, y una AUC promedio de 0.926 (Figura 6). La publicación con los resultados completos de este estudio está en el ANEXO 1 (Sánchez-Cruz and Medina-Franco, 2018).

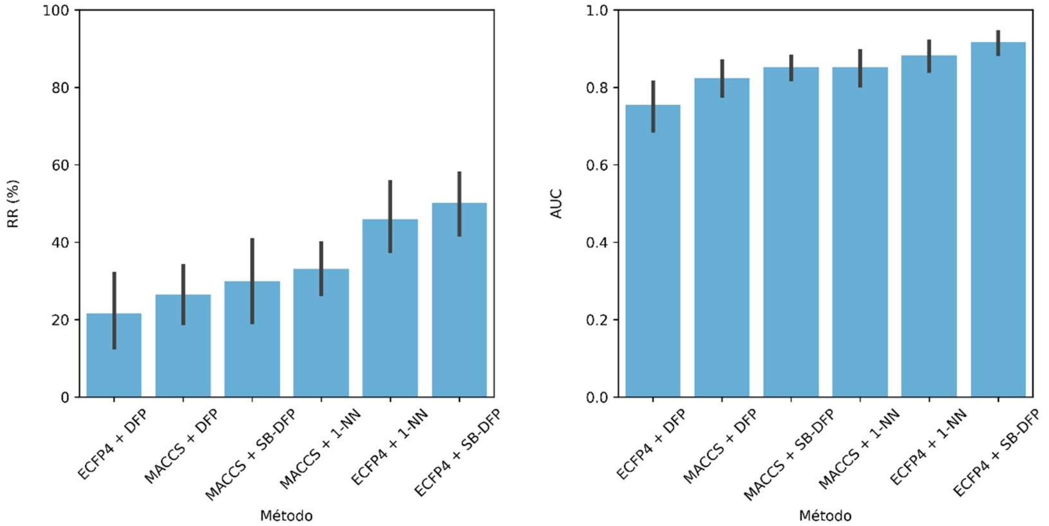


Figura 6. Desempeño de seis estrategias de búsqueda por similitud evaluadas sobre 28 quimiotecas asociadas a dianas epigenéticas. *Cada estrategia se define como un método de búsqueda (1-NN, DFP o SB-DFP) basada en una huella digital molecular. Se muestran los resultados en términos de la tasa de recuperación (RR) y el área bajo la curva ROC (AUC).*

Estos resultados sugieren que el uso de la SB-DFP como plantilla para realizar búsquedas por similitud es una alternativa prometedora para la identificación de nuevas interacciones diana-ligando cuando se conoce más de un ligando de referencia. Sin embargo, se debe considerar que su aplicación debe ser validada con un mayor número de dianas biológicas. Así mismo, el efecto del número de ligandos de referencia usados para su construcción también debe ser evaluado en este tipo de búsquedas. Estos son dos etapas por efectuar como perspectiva del desarrollo y aplicación generalizada de la SB-DFP.

VIII.3.2 Estudio de relaciones entre dianas

Un árbol de mínima cobertura fue construido a partir de la matriz de similitud de 136 dianas epigenéticas, usando una SB-DFP para representar a cada una de ellas (Figura 7). Para facilitar la interpretación de esta representación, las dianas fueron clasificadas manualmente en 14 clases de acuerdo con su función y coloreadas acorde a ello, empleando la siguiente notación: KIN – cinasa, BET – proteína que contiene bromodominio, HDAC – histona deacetilasa, HDM – histona demetilasa, CHR – remodelador de cromatina, HMT – histona metiltransferasa, HAT – histona acetiltransferasa, PRMT- arginina metiltransferasa, PcG – proteína del grupo Polycomb, MRNA – modificador de RNA, DNMT – DNA metiltransferasa, HMR – lector de metilación de histonas, TUD – proteína que contienen el dominio Tudor y Otra – otras proteínas.

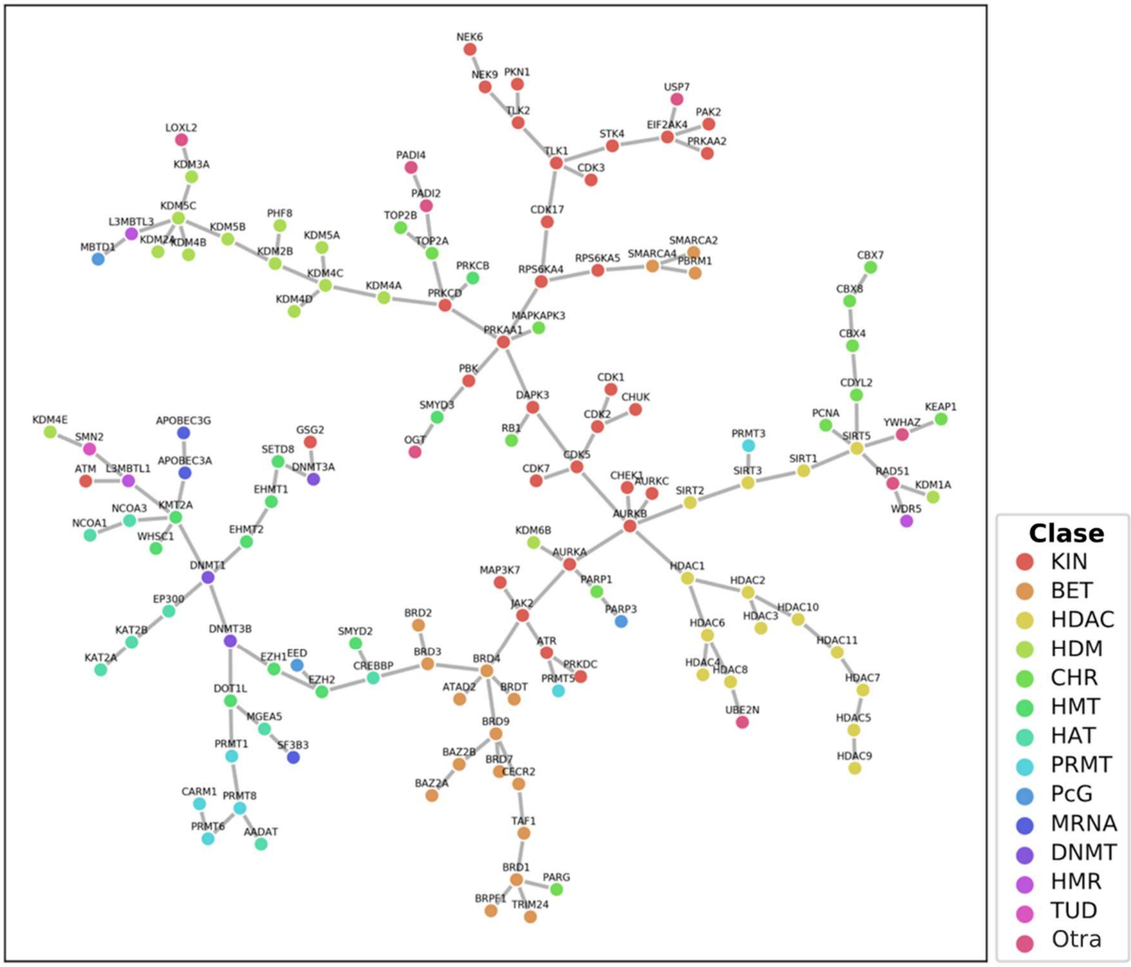


Figura 7. Árbol de mínima cobertura para 136 dianas epigenéticas. *Cada nodo representa la SB-DFP de una diana, generada a partir de sus ligandos conocidos. Adaptada de* (Sessions *et al.*, 2020).

La representación de las dianas epigenéticas en un árbol de mínima cobertura permite concluir que, en general, dianas biológicas con la misma función son agrupadas juntas, ya sea en la misma rama o en ramas cercanas del árbol (Figura 7). Esto es consistente con la dificultad conocida para el diseño de ligandos selectivos dentro de una misma familia de dianas. Por otra parte, en la región más externa del árbol se encuentran dianas relacionadas pobremente con el resto, donde destacan aquellas que no comparten su función con las dianas cercanas, lo que las sugiere como dianas prometedoras para el desarrollo de ligandos selectivos. Estos resultados indican que la representación de las dianas biológicas a través de la SB-DFP de sus ligandos conocidos y el estudio de sus relaciones a través de árboles de mínima cobertura, pueden guiar el diseño tanto de ligandos selectivos, como de ligandos que presenten interacción con múltiples dianas. Estos resultados forman parte de una publicación incluida en el ANEXO 1 (Sessions *et al.*, 2020).

VIII.3.3 Representación del espacio químico

El uso de la similitud de los compuestos a dos SB-DFPs construidas a partir de quimiotecas diferentes, permite la proyección de estos en un plano cartesiano dividido en dos regiones triangulares de igual tamaño, las cuales son pobladas por los compuestos con mayor relación a cada una de las SB-DFPs. En el caso de estudio presentado, estas regiones corresponden a compuestos de origen sintético y productos naturales, respectivamente. La Figura 8 muestra las representaciones generadas para quimiotecas de productos naturales, compuestos sintéticos, compuestos con actividad biológica reportada y bibliotecas de fragmentos generadas a partir de ellas. En dicha figura se observa que la mayor parte de las quimiotecas de productos naturales y compuestos de origen sintético se agrupan en las regiones más cercanas a las SB-DFPs que representan a cada uno de estos grupos. Por su parte, los compuestos con actividad biológica reportada ocupan espacio en ambas secciones, con una tendencia hacia la zona de compuestos sintéticos. Por otra parte, las bibliotecas de fragmentos generadas siguen la misma tendencia que las quimiotecas a partir de las cuáles fueron generadas, sugiriendo la preservación de las características estructurales de los compuestos de los cuales provienen.

Estos resultados indican que la SB-DFP es capaz de capturar las diferencias entre quimiotecas y que estas diferencias pueden ser usadas para generar una representación visual bidimensional del espacio químico ocupado por compuestos. Esta representación tiene la ventaja de ser interpretable fácilmente, permitiendo la clasificación de los compuestos proyectados en ella en grupos asociados a cada uno de sus ejes. Estos resultados forman parte de una publicación incluida en el ANEXO 1 (Chávez-Hernández *et al.*, 2020).

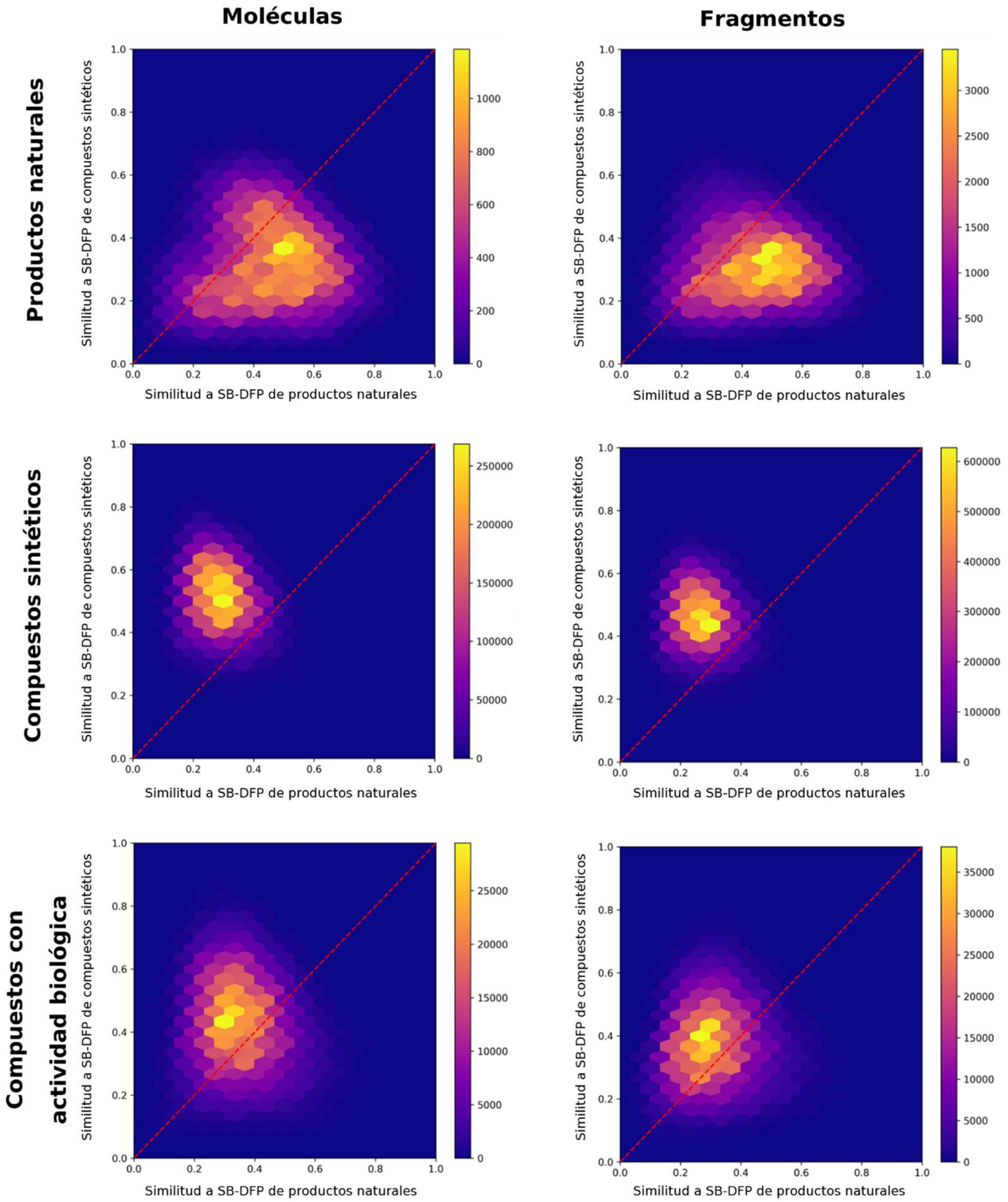


Figura 8. Representación visual del espacio químico basada en la SB-DFP. *Se muestran compuestos en quimiotecas de moléculas completas (izquierda) y los fragmentos que las constituyen (derecha) para productos naturales, compuestos sintéticos y compuestos con actividad biológica. El número de compuestos en cada región de los mapas de calor se indica mediante una escala de color representada en las barras de la derecha.*

IX. NAVEGACIÓN DE ESPACIOS QUIMIOGENÓMICOS EPIGÉNETICOS UTILIZANDO INTELIGENCIA ARTIFICIAL

IX.1 Significancia

La epigenética estudia la adaptación estructural de las regiones cromosómicas para la modificación o mantenimiento de su actividad (Ganesan *et al.*, 2019). Esta adaptación involucra la modificación reversible de los ácidos nucleicos y las histonas, la cual es efectuada por enzimas de múltiples tipos, las dianas epigenéticas. Estas dianas se clasifican en tres grupos principales (Biswas and Rao, 2018): escritoras (*writers*) – aquellas que añaden grupos químicos, como las DNMTs, las HMTs y las HATs, borradoras (*erasers*) – aquellas que remueven los grupos introducidos por las escritoras, como las HDACs y las HDMs, y lectoras (*readers*) – las que contienen dominios especiales capaces de reconocer los cambios hechos por las dianas anteriores, como las BET. Además de estos tres grupos principales, existen otras dianas biológicas que desempeñan un papel importante en la regulación epigenética. Estas dianas incluyen a las chaperonas de histonas y las CHRs, necesarias para el ensamble y reestructuración del nucleosoma (Burgess and Zhang, 2013; Tyagi *et al.*, 2016), e incluso algunas clases de factores de transcripción (Mayran and Drouin, 2018).

El interés en las dianas epigenéticas como blancos terapéuticos ha aumentado con el tiempo, ya que su desregulación está asociada con múltiples enfermedades (Esteller, 2008; Cacabelos and Torrellas, 2014; Küçükali *et al.*, 2015). El ejemplo más claro de esto es la existencia de ocho “epifármacos” (fármacos cuyo efecto terapéutico se debe a la interacción con dianas epigenéticas) aprobados para su uso clínico como tratamiento contra diversos tipos de cáncer, y la evaluación en fases clínicas de más de una decena de potenciales epifármacos. Por estas razones, el número de bases de datos quimiogenómicas públicas relacionadas con este tipo de dianas se ha incrementado en la última década (Sessions *et al.*, 2020). A pesar de ello, la cantidad de información asociada a las dianas epigenéticas resulta mínima en comparación con la información disponible para otras familias de proteínas, como los GPCRs y los canales iónicos (Oprea *et al.*, 2018; Zdrazil *et al.*, 2020). Debido a este desbalance de datos, las dianas epigenéticas están representadas pobremente en las herramientas disponibles actualmente para el cribado virtual inverso, manifestando la necesidad de desarrollar modelos predictivos enfocados en este tipo de dianas.

Los modelos de aprendizaje automático han mostrado su utilidad en múltiples áreas del descubrimiento de fármacos, incluyendo la predicción de interacciones diana-ligando (Lo *et al.*, 2018; Mayr *et al.*, 2018; Vamathevan *et al.*, 2019). Sin embargo, su aplicación en dianas epigenéticas ha sido poco explorada, con la mayoría de las investigaciones enfocadas en familias de proteínas como las HDACs (Norinder *et al.*, 2019) y las BETs (Speck-Planche and Scotti, 2019).

En este capítulo se introduce *Epigenetic Target Profiler* (ETP) (Sánchez-Cruz and Medina-Franco, 2021b), una aplicación web de acceso libre que emplea algoritmos de aprendizaje automático para la predicción del perfil de interacción de moléculas pequeñas sobre 55 dianas epigenéticas. La sección IX.2 describe los conjuntos de datos usados, la construcción de los modelos de aprendizaje automático y su implementación como una aplicación web, mientras que la sección IX.3 muestra el desempeño de los modelos y el uso de la aplicación.

IX.2 Metodología

IX.2.1 Datos quimiogenómicos

Para la construcción y validación de los modelos predictivos se emplearon datos cuantitativos de interacciones diana-ligando asociados a dianas epigenéticas, extraídos de ChEMBL 27 (Mendez *et al.*, 2019) y PubChem (Kim *et al.*, 2019). Se construyeron conjuntos de compuestos asociados a cada diana en donde a cada compuesto se le asignó una clase (activo / inactivo). Los compuestos con una actividad biológica (IC_{50}, EC_{50}, K_i o K_d) menor o igual a 10 µM fueron clasificados como activos y los restantes como inactivos. Únicamente los conjuntos (55 en total) con al menos 30 compuestos activos y 30 compuestos inactivos fueron usados para construir modelos de aprendizaje automático.

IX.2.2 Construcción y validación de los modelos

Los modelos predictivos fueron construidos para la clasificación binaria (activo / inactivo) de los compuestos asociados a cada diana. Estos modelos fueron validados mediante el estudio de su desempeño en dos tareas: la clasificación de los compuestos en el contexto de cada diana (cribado virtual), y la identificación de las dianas asociadas a cada compuesto (cribado virtual inverso).

IX.2.2.1 Cribado virtual

Se construyeron 15 modelos distintos de clasificación binaria para cada uno de los 55 conjuntos de compuestos asociados a dianas epigenéticas. Estos modelos resultaron del entrenamiento de cinco algoritmos de aprendizaje automático: k vecinos más cercanos (k-NN, *k-Nearest Neighbors*) (Altman, 1992), RF (Tin Kam Ho, 1998), árboles de gradiente potenciado (GBT, *Gradient Boosting Trees*) (Friedman, 2001), SVM (Cortes and Vapnik, 1995) y ANN (Hopfield, 1982), entrenados a partir de tres huellas digitales moleculares: ECFP4 (2048 bits), MACCS Keys (166 bits) y RDK (2048 bits). Cada uno de estos modelos individuales es referido como “huella digital::algoritmo”.

Los hiperparámetros de cada modelo fueron optimizados mediante validación cruzada de diez iteraciones, utilizando la exactitud balanceada (BA, *Balanced Accuracy*) como métrica de desempeño. Los 15 modelos optimizados fueron comparados en términos de su BA y las predicciones de los dos mejores fueron combinadas para generar un modelo consenso. Los desempeños de los modelos individuales y el consenso se compararon en términos de su precisión (PPV, *positive predictive value*) y sensibilidad (TPR, *true positive rate*).

Para estimar el dominio de aplicación de los modelos, se calculó la distancia promedio de cada compuesto en el conjunto de prueba a los compuestos en el conjunto de entrenamiento, empleando como métrica la distancia de Jaccard (1- T_c). Con ello, las predicciones fueron clasificadas en cuatro cuartiles de distancia que indican la cercanía de los compuestos predichos al conjunto de compuestos empleados para la generación del modelo (Q1 – Q4), un criterio conocido como distancia al modelo (DM, D*istance-to-Model*) (Tetko *et al.*, 2008; Sushko *et al.*, 2010). Para su comparación, tanto la PPV como la TPR fueron calculados en función de este parámetro de distancia.

IX.2.2.2 Cribado virtual inverso

Como segunda estrategia de validación, se evaluó el desempeño de los dos mejores modelos identificados en la primera estrategia y del modelo consenso para la identificación de las dianas asociadas a 10 muestras de compuestos. Bajo este esquema, un modelo hace referencia a la combinación de 55 clasificadores individuales (uno para cada diana) construidos mediante una combinación de huella digital molecular y un algoritmo de aprendizaje automático. Las 10 muestras de compuestos fueron construidas para contener el mismo número de asociaciones conocidas para cada diana y con ello evitar un sesgo por aquellas con mayor número de asociaciones conocidas. En esta estrategia, solo los dos mejores modelos individuales y el modelo consenso fueron evaluados. Para cada muestra de compuestos, todos los clasificadores binarios fueron reentrenados excluyendo los compuestos de la muestra. El desempeño de estos nuevos modelos fue analizado en términos de su PPV y TPR, calculadas usando el mismo criterio de DM empleado en la estrategia previa.

IX.2.3 Implementación web

El mejor modelo identificado en la segunda estrategia se implementó como una aplicación web de acceso libre para la predicción del perfil de interacción de moléculas pequeñas sobre las 55 dianas epigenéticas. Esta aplicación es soportada por un servidor de la Facultad de Química de la UNAM asignado al grupo DIFACQUIM y fue implementada usando Flask (versión 1.1.1) para Python como entorno web, mientras que la interfaz gráfica de usuario (GUI, *Graphical User Interface*) fue escrita en HTML5, CSS y JavaScript.

IX.3 Resultados y discusión

IX.3.1 Cribado virtual

De los 15 modelos estudiados para la clasificación de los compuestos asociados a 55 dianas epigenéticas, se encontró que los dos modelos con el mejor desempeño en términos de BA (y sin diferencias significativas entre ellos) fueron aquellos basados en SVM y entrenados a partir de las huellas digitales ECFP4 y RDK (ECFP4::SVM y RDK::SVM), con BA medias de 0.830 y 0.827. El modelo consenso construido a partir de estos dos modelos individuales, se generó de tal forma que solo los compuestos predichos como activos por ambos modelos eran considerados activos por este modelo, obteniendo una BA media de 0.835, pero sin diferencias significativas al desempeño de los modelos individuales. Comparando la PPV y la TPR de estos tres modelos en función del parámetro de DM, se encontró que, a costa de una disminución en la TPR, la PPV del modelo consenso fue superior a la de los modelos individuales para todos los cuartiles de DM (Figura 9).

La PPV de los modelos decrece al incrementar la DM, con valores promedio para el modelo consenso que van de 0.923 para Q1 a 0.810 para Q4. Esto sugiere una alta confiabilidad de las predicciones, aun cuando los compuestos evaluados presentan baja similitud con los compuestos del conjunto de entrenamiento.

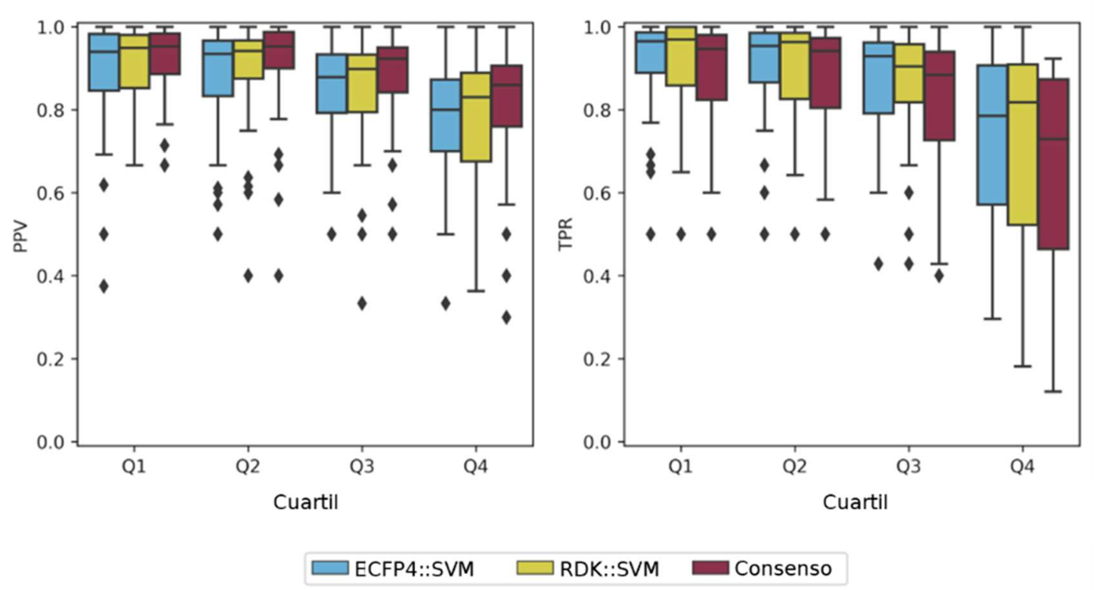


Figura 9. Comparación de 55 clasificadores individuales en función de la distancia al modelo. *Se muestra su desempeño en términos de precisión (PPV) y sensibilidad (TPR).*

IX.3.2 Cribado virtual inverso

Al estudiar el desempeño de la combinación de 55 clasificadores binarios en la identificación de las dianas epigenéticas asociadas a 10 conjuntos de compuestos, se observó la misma tendencia que en la clasificación individual, encontrando que tanto los valores de PPV como de TPR, decrecen a medida que aumenta la DM. Nuevamente, se encontró que el modelo consenso superó a los modelos individuales en términos de PPV para todos los cuartiles de DM, con valores promedio que van de 0.952 para Q1 a 0.773 para Q4, mientras que sus valores de TPR fueron menores que los de los modelos individuales (Figura 10). Si bien, ambas son métricas importantes para el desempeño de un modelo predictivo, en su uso para la búsqueda de nuevas interacciones diana-ligando se prefieren valores altos de PPV, pues esto indica que las predicciones de compuestos activos son más certeras.

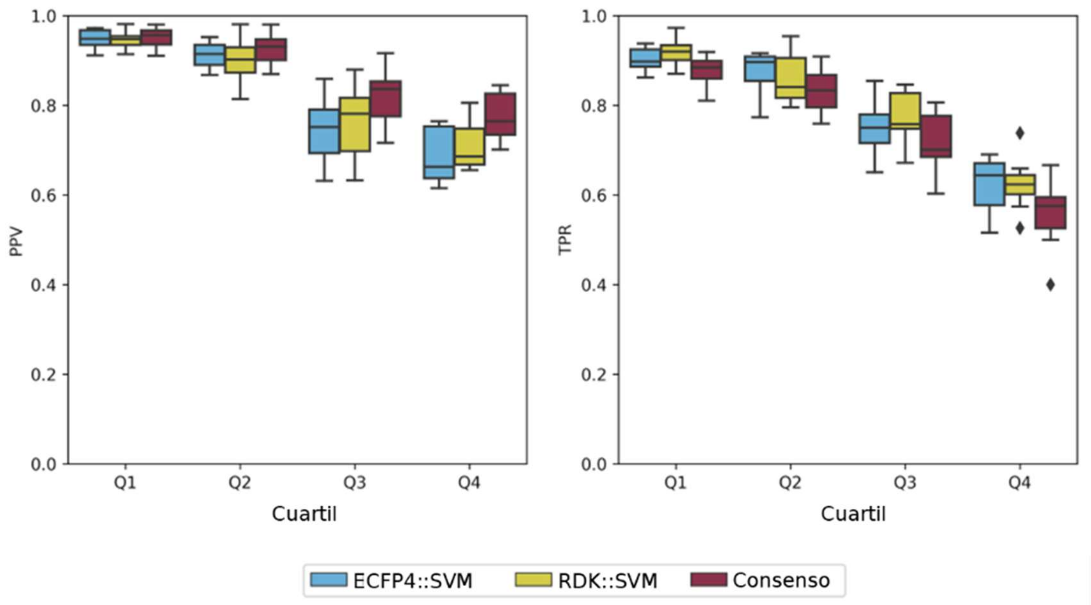


Figura 10. Comparación de la combinación de 55 clasificadores individuales en función de la distancia al modelo. *Se muestra su desempeño en términos de precisión (PPV) y sensibilidad (TPR).*

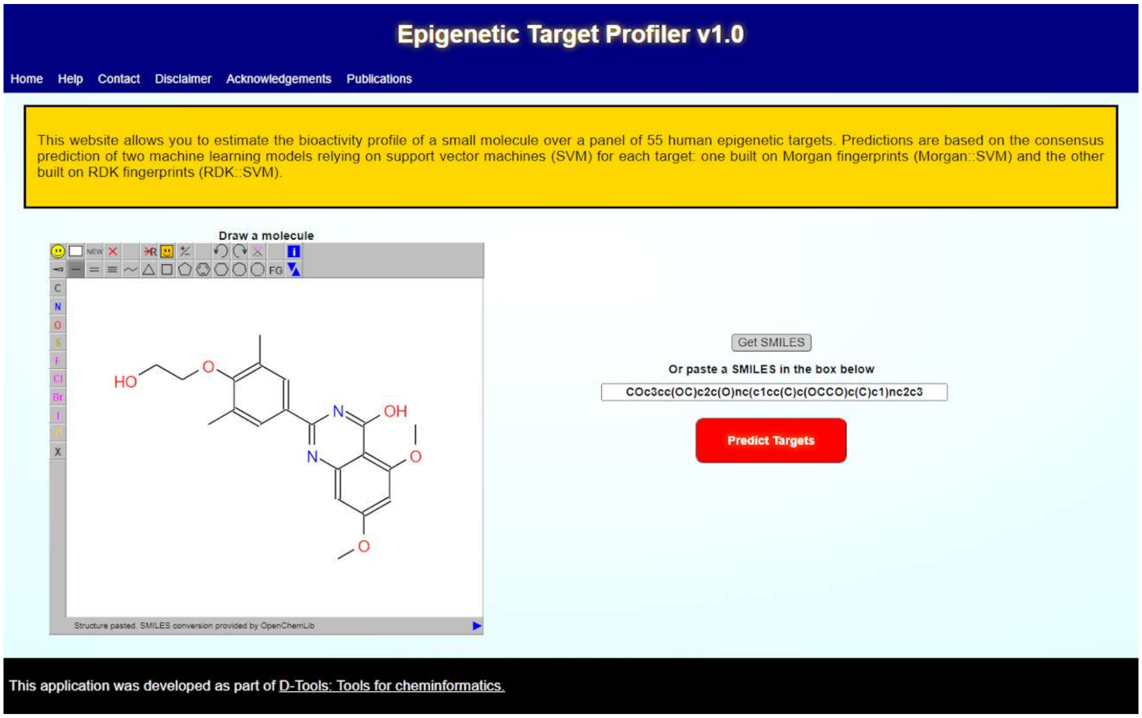
Los hallazgos encontrados mediante ambas estrategias de validación muestran la utilidad potencial de estos modelos para la identificación de nuevas interacciones entre moléculas pequeñas y dianas epigenéticas. Estos resultados forman parte de una publicación en proceso de revisión de pares, incluida en el ANEXO I (Sánchez-Cruz and Medina-Franco, 2021a)

IX.3.3 Aplicación web

Para facilitar el uso del modelo consenso introducido en las secciones anteriores, este fue reentrenado utilizando todos los datos disponibles, e implementado en una aplicación web fácil de usar y de acceso libre, disponible en:

http://www.epigenetictargetprofiler.com/

La GUI de ETP consta de una página de inicio en la que el usuario puede dibujar una molécula de consulta usando el editor molecular JSME (Bienfait and Ertl, 2013) y generar su representación en formato SMILES pulsando el botón “Get SMILES” o introducir directamente el SMILES del compuesto en la celda destinada para ello. Una vez que se tiene el SMILES del compuesto en cuestión, la predicción de sus potenciales dianas epigenéticas puede ser iniciada pulsando el botón “Predict Targets” (Figura 11).



Epigenetic Target Profiler v1.0

This website allows you to estimate the bioactivity profile of a small molecule over a panel of 55 human epigenetic targets. Predictions are based on the consensus prediction of two machine learning models relying on support vector machines (SVM) for each target: one built on Morgan fingerprints (Morgan::SVM) and the other built on RDK fingerprints (RDK::SVM).

Or paste a SMILES in the box below

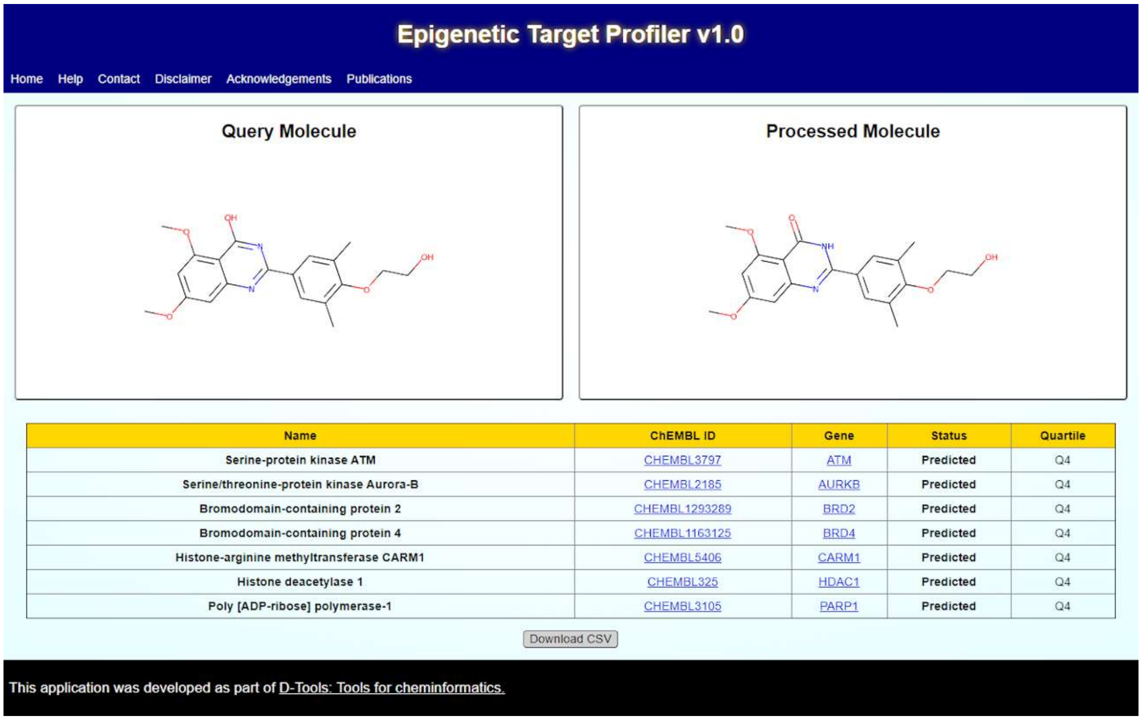
COc3cc(OC)c2c(O)nc(c1cc(C)c(OCCO)c(C)c1)nc2c3

Structure pasted. SMILES conversion provided by OpenChemLib

This application was developed as part of D-Tools: Tools for cheminformatics.

Figura 11. Interfaz gráfica de usuario de Epigenetic Target Profiler. *Se muestra la estructura y SMILES del compuesto RVX-208, un inhibidor de BRD en desarrollo clínico.*

La aplicación estandariza la molécula de entrada de acuerdo con el procedimiento descrito en la sección VII.2.1 y genera las huellas digitales moleculares ECFP4 (2048 bits) y RDK (2048 bits). Para cada una de las dianas epigenéticas, si el compuesto es parte del conjunto de entrenamiento usado para la construcción de los modelos, su asociación conocida es devuelta y no se realiza ningún procedimiento adicional. En caso contrario, se calcula su DM, se clasifica en un cuartil (Q1 – Q4) y se predice por el modelo consenso. Después de este proceso, el usuario es dirigido a la página de resultados.


Epigenetic Target Profiler v1.0
Home Help Contact Disclaimer Acknowledgements Publications
Query Molecule
Processed Molecule
Download CSV

Name	ChEMBL ID	Gene	Status	Quartile
Serine-protein kinase ATM	CHEMBL3797	ATM	Predicted	Q4
Serine/threonine-protein kinase Aurora-B	CHEMBL2185	AURKB	Predicted	Q4
Bromodomain-containing protein 2	CHEMBL1293289	BRD2	Predicted	Q4
Bromodomain-containing protein 4	CHEMBL1163125	BRD4	Predicted	Q4
Histone-arginine methyltransferase CARM1	CHEMBL5406	CARM1	Predicted	Q4
Histone deacetylase 1	CHEMBL325	HDAC1	Predicted	Q4
Poly [ADP-ribose] polymerase-1	CHEMBL3105	PARP1	Predicted	Q4

This application was developed as part of D-Tools: Tools for cheminformatics.

Figura 12. Página de resultados de Epigenetic Target Profiler. *Se muestran las asociaciones conocidas y predichas parta el compuesto RVX-208.*

La página de resultados consta de dos imágenes en la parte superior y una tabla debajo de ellas (Figura 12). La imagen de la izquierda muestra la estructura de la molécula de consulta tal como fue introducida por el usuario, mientras que la imagen de la derecha corresponde a la estructura estandarizada de la misma. La tabla de resultados muestra las dianas epigenéticas asociadas al compuesto de consulta, incluyendo cinco columnas con información adicional. Las primeras tres columnas contienen el nombre de las dianas y los vínculos de acceso a dos bases de datos (ChEMBL y GeneCards) con información sobre la misma (*Name, ChEMBL ID* y *Gene*). Las últimas dos columnas contienen la información acerca de las predicciones (*Status* y *Quartile*), la primera de ellas indica si la interacción es conocida o predicha, mientras que la segunda indica el cuartil de DM al que pertenece la predicción. La lista completa de resultados para las 55 dianas puede descargarse como un archivo delimitado por comas, pulsando el botón “Download CSV” que aparece debajo de la tabla.

Los resultados obtenidos muestran a ETP como una aplicación web de acceso libre y de fácil uso para usuarios no experimentados en el manejo de herramientas computacionales, o para usuarios con mediana o alta experiencia en programación pero que no cuentan con el tiempo suficiente para efectuar los cálculos. Se anticipa la utilidad de ETP en proyectos individuales y multidisciplinarios enfocados en la identificación de nuevas interacciones entre moléculas pequeñas y dianas epigenéticas. Los modelos implementados en esta herramienta fueron generados a partir de los datos quimiogenómicos más actualizados en el dominio público al momento de realizar esta tesis. Sin embargo, estos modelos deberán ser actualizados a medida que nuevos datos quimiogenómicos estén disponibles.

Al estar implementada en un servidor asignado al grupo DIFACQUIM, su presencia en línea está garantizada por cinco años. En el mediano y largo plazo, este servidor puede ser actualizado anualmente, a medida que nueva información quimiogenómica sea publicada en bases de datos como ChEMBL y PubChem. Se postula que el uso del servidor en proyectos de aplicación permitiría la asignación de recursos económicos, ya sea de iniciativa pública o privada, para el mantenimiento y desarrollo de futuras versiones de ETP.

La implementación de este recurso computacional forma parte de una publicación en proceso de revisión de pares, incluido en el ANEXO I (Sánchez-Cruz and Medina-Franco, 2021b).

X. MEJORANDO LAS PREDICCIONES DE AFINIDAD DE UNIÓN PROTEÍNA-LIGANDO MEDIANTE LA DESCRIPCIÓN QUÍMICA DE LOS COMPLEJOS

X.1 Significancia

El acoplamiento molecular es una de las metodologías más usadas en la predicción de interacciones diana-ligando basadas en la información de la estructura 3D de la diana. Estos métodos proponen posibles modos de unión entre moléculas pequeñas y macromoléculas biológicas, por lo que uno de sus componentes principales es la función de puntuación que emplean para la estimación de la afinidad de unión de los complejos propuestos. Las funciones de puntuación clásicas, detalladas en la sección II.4.2, han mostrado ser útiles en la identificación tanto de nuevas interacciones diana-ligando como de sus modos de unión. Sin embargo, sus puntuaciones muestran una baja correlación con las afinidades de unión obtenidas experimentalmente (Su *et al.*, 2019). Por otro lado, las funciones de puntuación basadas en algoritmos de aprendizaje automático han mostrado ser superiores a las funciones clásicas en distintos aspectos, incluyendo la obtención de puntuaciones altamente correlacionadas con los datos experimentales.

La primera función de puntuación basada en un algoritmo de aprendizaje automático que obtuvo un buen desempeño en la predicción de afinidades de unión proteína-ligando fue RF-Score (Ballester and Mitchell, 2010). Esta función emplea un algoritmo de RF entrenado a partir de estructuras de complejos representados como un vector de 36 posiciones, en la que cada posición corresponde al conteo de un par de átomos proteína-ligando, considerando para el conteo a los pares de átomos que se encuentren a una distancia de 12 Å o menos. Desde entonces, múltiples algoritmos de aprendizaje automático y representaciones de los complejos han sido empleados para la construcción de este tipo de funciones de puntuación (Tabla 5). Con el auge del aprendizaje profundo, uno de los algoritmos más explorados son las redes neuronales convolucionales (CNN, *Convolutional Neural Networks*), una clase de ANN ampliamente utilizada en el análisis de imágenes, para lo que se han propuesto distintas formas de representar a los complejos proteína-ligando en forma de imágenes (Jiménez *et al.*, 2018; Hassan-Harrirou *et al.*, 2020). Otras aproximaciones describen a los complejos proteína-ligando mediante descriptores derivados de su análisis mediante topología algebraica (Cang *et al.*, 2018) o geometría diferencial (Nguyen and Wei, 2019b), las cuales han sido usadas para el entrenamiento de algoritmos de GBT. A pesar de la mayor complejidad de estas nuevas funciones de puntuación, su desempeño en la predicción de la afinidad de unión de complejos proteína-ligando no ha mejorado en forma notable.

Tabla 5. Funciones de puntuación basadas en aprendizaje automático.

Función de puntuación	**Algoritmo**	**Descripción de los complejos proteína-ligando**	**Referencia**
RF-Score	RF	Conteos de pares de átomos proteína-ligando	(Ballester and Mitchell, 2010)
NN-Score 2.0	ANN	Términos de interacción de Autodock Vina, conteos de pares de átomos proteína-ligando y términos electrostáticos	(Durrant and McCammon, 2011)
ID-Score	SVM	Nueve categorías de descriptores relacionados a interacciones proteína-ligando	(Li *et al.*, 2013)
SFCscore^{RF}	RF	Términos de interacción de SFCscore	(Zilian and Sotriffer, 2013)
ΔVinaRF$_{20}$	RF	Términos de interacción de Autodock Vina y descriptores moleculares adicionales	(Wang and Zhang, 2017)
RI-Score	RF	Descriptores de índice de rigidez	(Nguyen *et al.*, 2017)
TNet-BP	CNN	Derivados de topología algebraica	(Cang and Wei, 2017)
K$_{DEEP}$	CNN	Descriptores moleculares inscritos en una rejilla tridimensional	(Jiménez *et al.*, 2018)
TopBP-ML	GBT	Derivados de topología algebraica	(Cang *et al.*, 2018)
TopBP-DL	CNN	Derivados de topología algebraica	(Cang *et al.*, 2018)
Pafnucy	CNN	Descriptores moleculares inscritos en una rejilla tridimensional	(Stepniewska-Dziubinska *et al.*, 2018)
PLEC-nn	ANN	Huella digital donde cada posición representa un contacto entre átomos proteína-ligando descritos mediante su ambiente atómico	(Wójcikowski *et al.*, 2019)
EIC-Score	GBT	Derivados de geometría diferencial	(Nguyen and Wei, 2019b)
AGL-Score	GBT	Descriptores estadísticos de las matrices de adyacencia y Laplacianas de subgrafos algebraicos	(Nguyen and Wei, 2019a)
OnionNet	CNN	Contactos entre pares de elementos proteína-ligando libres de rotación y agrupados en distintos intervalos de distancia	(Zheng *et al.*, 2019)
ΔVinaXGB	GBT	Términos de interacción de Autodock Vina y descriptores moleculares adicionales, inclusión de moléculas de agua	(Lu *et al.*, 2019)
NNScore::LD	ANN	Descriptores de NNScore 2.0 y descriptores del ligando	(Boyles *et al.*, 2020)
RosENet	CNN	Términos energéticos del campo de fuerza Rosetta y descriptores moleculares inscritos en una rejilla tridimensional	(Hassan-Harrirou *et al.*, 2020)

En este capítulo se introduce *Extended-Connectivity Interaction Features* (ECIF), un conjunto de descriptores derivados a partir del conteo de pares de átomos proteína-ligando, que toman en cuenta la conectividad química de cada átomo para describirlo y definir en consecuencia los posibles pares. La sección X.2 detalla el algoritmo para su construcción y su uso en la derivación de una nueva función de puntuación basada en aprendizaje automático. La sección X.3 muestra el desempeño de esta función de puntuación en la predicción de la afinidad de unión de complejos proteína-ligando sobre un conjunto de referencia diseñado para tal fin, y su comparación con otras funciones de puntuación disponibles actualmente.

X.2 Metodología

X.2.1 Construcción

En los conteos de pares de átomos proteína-ligando que componen a ECIF, cada uno de los átomos involucrados en estos pares es definido a partir de su conectividad dentro de la molécula a la cual pertenece. Para ello, un átomo es descrito mediante seis características: su símbolo atómico, su valencia explícita, el número de átomos pesados a los que está unido, el número de hidrógenos a los que está unido, su pertenencia a un sistema cíclico y su aromaticidad. Esta definición es similar a la usada en la construcción de las ECFP.

Los tipos de átomo presentes tanto en moléculas pequeñas como en proteínas se obtuvieron del análisis de 9299 estructuras cristalográficas de complejos proteína-ligando con valores de afinidad conocidos (pK_i / pK_d), extraídas de PDBBind. Para los ligandos, se encontraron 70 tipos de átomos diferentes, mientras que, para las proteínas, se definieron 22 posibles tipos de átomos. Bajo estas consideraciones, un complejo proteína-ligando descrito mediante ECIF es representado como un vector de 1540 posiciones. Cada una de estas posiciones representa el conteo de un par de átomos proteína-ligando, considerando para el conteo a los pares de átomos que se encuentren a una distancia predefinida por el usuario (Figura 13).

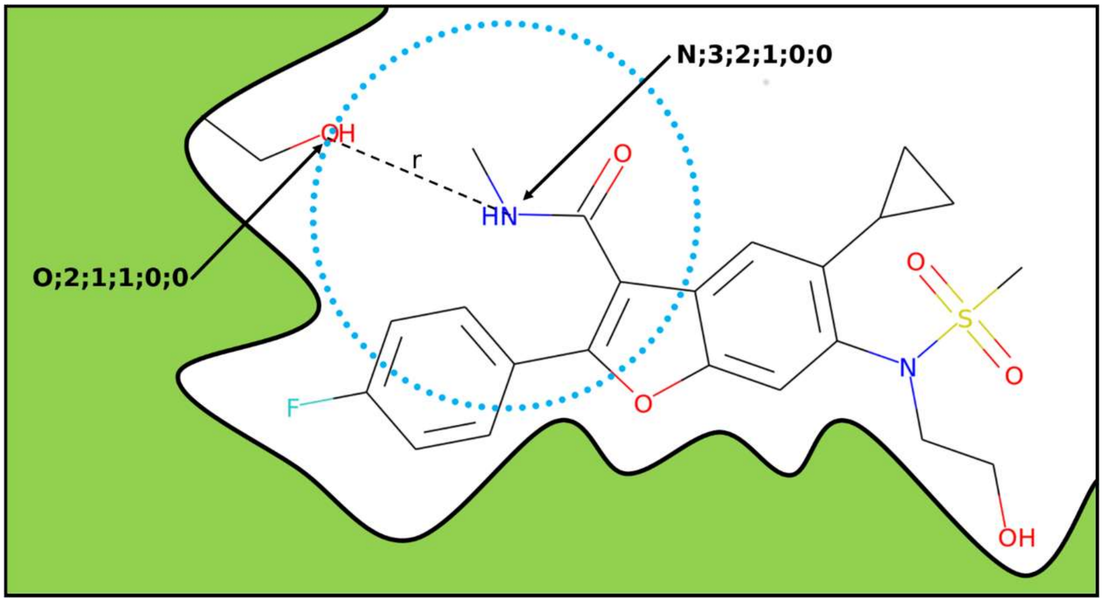


Figura 13. Representación esquemática de un par de tipos de átomo proteína-ligando incluido en ECIF. *Se representa el par “O;2;1;0;0;0” – “N;3;2;1;0;0”. Para cada uno de los átomos en el ligando, todos los átomos de la proteína dentro de una circunferencia delimitada por un radio* ***r*** *son considerados para el conteo de pares. Este radio es el único parámetro ajustable para el cálculo de ECIF.*

X.2.2 Entrenamiento de algoritmos de aprendizaje automático

Para mostrar el poder descriptivo de ECIF, se construyeron funciones de puntuación basadas en aprendizaje automático entrenadas a partir complejos proteína-ligando descritos mediante ECIF. Para ello, se emplearon 9299 complejos descritos en forma de ECIF a distintos valores de distancia, y se entrenaron dos diferentes algoritmos de aprendizaje automático, RF y GBT. Se evaluó el desempeño de estos modelos en la predicción de la afinidad de unión de un conjunto diverso de 285 complejos proteína-ligando que se usa como referencia para la evaluación de este tipo de funciones de puntuación, y cuyas estructuras no forman parte de las 9299 con las que se entrenó al modelo (Su *et al.*, 2019). Los modelos fueron evaluados en términos del coeficiente de correlación de Pearson (R_p) y el error cuadrático medio (RMSE, *Root Mean Square Error*) entre las predicciones de los modelos y los valores experimentales. Puesto que se ha mostrado que la inclusión de descriptores basados puramente en el ligando mejora las predicciones de este tipo de modelos (Boyles *et al.*, 2020), los mismos modelos fueron reentrenados a partir de la combinación de ECIF con 170 descriptores basados en el ligando. Estos nuevos modelos fueron evaluados utilizando los mismos conjuntos de entrenamiento y prueba descritos previamente. Los resultados obtenidos por los modelos basados en ECIF con el mejor desempeño en esta tarea fueron comparados con los de otras funciones de puntuación disponibles actualmente y evaluadas sobre el mismo conjunto de prueba.

X.3 Resultados y discusión

X.2.1 Predicción de afinidades de unión proteína ligando

El código para el cálculo de ECIF a partir de un complejo ligando proteína es de acceso libre y se encuentra disponible en https://github.com/DIFACQUIM/ECIF. Al evaluar las funciones de puntuación basadas en ECIF para la predicción de las afinidades de unión del conjunto de referencia usado, se encontró que las mejores fueron aquellas que usaban GBT como algoritmo de aprendizaje automático, tanto en términos de R_p como de RMSE. La función de puntuación con los mejores resultados en esta tarea fue la construida a partir de ECIF calculadas usando una distancia de corte de 6 Å para el conteo de los pares de átomos (ECIF6-GBT), con un R_p de 0.857 y un RMSE de 1.193. Se demostró también que la inclusión de descriptores basados en el ligando mejora significativamente el desempeño de estas funciones de puntuación, alcanzando un R_p de 0.866 y un RMSE de 1.169 (ECIF6::LD-GBT). La Figura 14 ilustra las predicciones obtenidas por estos modelos.

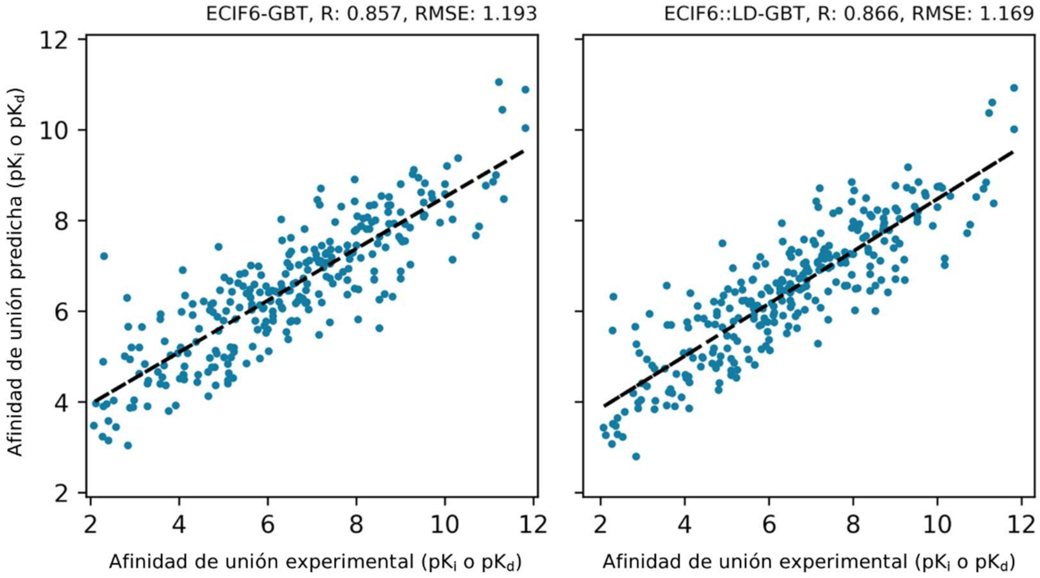


***Figura 14. Predicciones de la afinidad de unión de complejos proteína-ligando.** Se muestran las predicciones obtenidas por los modelos ECIF6-GBT (izquierda) y ECIF6::LD-GBT (derecha). En la parte superior de cada gráfica se indica el coeficiente de correlación de Pearson (R) y el error cuadrático medio (RMSE) entre las predicciones y los valores experimentales.*

X.2.2 Comparación con otras funciones de puntuación

Al comparar los resultados obtenidos por los modelos ECIF6-GBT y ECIF6::LD-GBT con los reportados para otras funciones de puntuación basadas en aprendizaje automático y evaluadas sobre el mismo conjunto de prueba, se encontró que las predicciones obtenidas por las funciones de puntuación basadas en ECIF muestran los mayores valores de R_p y los menores valores de RMSE reportados hasta la fecha para una función de puntuación de este tipo (Figura 15). Estos hallazgos sugieren que una descripción química más completa de los pares de átomos proteína-ligando presentes en la estructura de los complejos, como la incorporada en ECIF, permite la construcción de funciones de puntuación basadas en aprendizaje automático que reproducen mejor los datos de afinidad de unión obtenidos experimentalmente. Esto es un aspecto fundamental en la identificación de nuevas interacciones diana-ligando basados en la estructura de la diana. Los resultados obtenidos forman parte de una publicación incluida en el ANEXO 1 (Sánchez-Cruz *et al.*, 2020). Las funciones de puntuación fueron construidas y evaluadas en complejos proteína-ligando cuya estructura fue determinada experimentalmente. En la identificación de nuevas interacciones esta información no está disponible, por lo que este tipo de funciones de puntuación deben ser evaluadas en modos de unión propuestos por los programas de acoplamiento molecular. La primera página de una publicación que muestra la aplicación de ECIF6::LD-GBT bajo estas condiciones y relacionada con la identificación de nuevos núcleos estructurales como inhibidores de DNMT1, se incluye en el ANEXO 2 (Juárez-Mercado *et al.*, 2020).

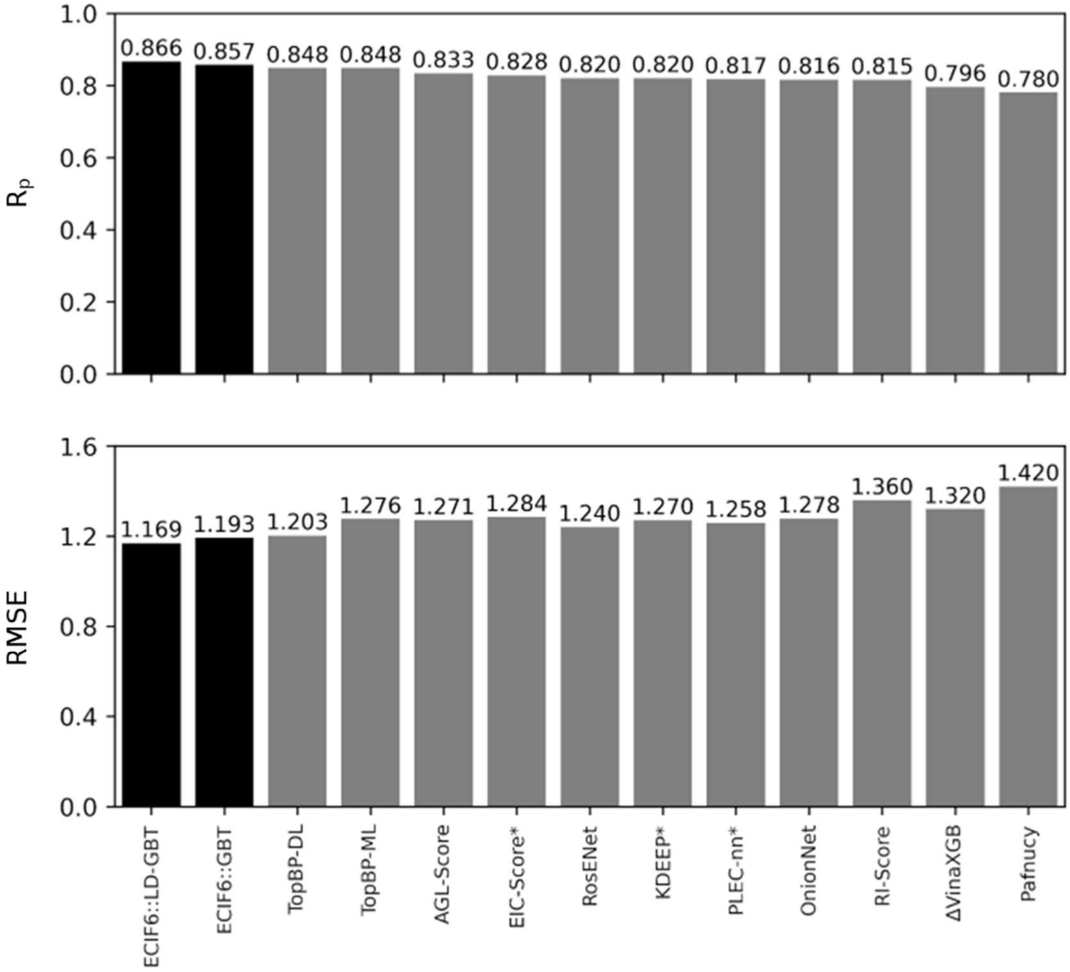


Figura 15. Comparación del desempeño de distintas funciones de puntuación basadas en aprendizaje automático. *Se muestran los resultados en términos del coeficiente de correlación de Pearson (R_p) y el error cuadrático medio (RMSE) entre las predicciones y los valores experimentales de 285 complejos proteína-ligando. Las funciones de puntuación marcadas con * usan una versión alterna del conjunto de prueba que incluye a 290 complejos.*

XI. CONCLUSIONES

Se desarrollaron tres herramientas computacionales para la predicción de nuevas interacciones entre moléculas pequeñas y dianas biológicas. SB-DFP y ETP forman parte de metodologías basadas en la estructura del ligando, mientras que ECIF está asociada con métodos basados en la estructura de la diana. SB-DFP y ECIF tienen aplicaciones en cribado virtual, y ETP se emplea en cribado virtual inverso.

Se propuso la SB-DFP como una nueva representación de quimiotecas a partir de las huellas digitales moleculares de los compuestos que las componen. Se mostró su aplicación en distintas áreas, destacando su uso como plantilla en búsquedas por similitud, cuando se conocen múltiples ligandos para una diana biológica. El desempeño de SB-DFP en la identificación de los ligandos asociados a 28 dianas biológicas de relevancia terapéutica fue superior al obtenido por dos metodologías ampliamente usadas en este tipo de búsquedas, validando así el uso de SB-DFP en el cribado virtual de quimiotecas.

El análisis de la información quimiogenómica pública más actualizada permitió la construcción, la validación y la comparación de modelos de aprendizaje automático para la predicción de las dianas epigenéticas de moléculas pequeñas, un grupo de dianas terapéuticas pobremente representadas en los métodos actuales. La identificación de los modelos con el mejor desempeño en esta tarea condujo al desarrollo e implementación de ETP, una aplicación web para la predicción del perfil de actividad de moléculas pequeñas sobre 55 dianas epigenéticas. ETP forma parte del conjunto de recursos libres D-Tools.

Se propuso ECIF como una nueva representación de complejos proteína-ligando basada en los conteos de pares de átomos presentes en ellas, considerando la conectividad de los átomos para definirlos y determinar así los posibles pares. Se mostró su uso para la construcción de una función de puntuación basada en aprendizaje automático. Esta función de puntuación tuvo la mejor correlación con datos experimentales reportada hasta la fecha para este tipo de metodologías, lo cual fue demostrado al comparar sus desempeños en la predicción de la afinidad de unión de 285 complejos usados como referencia.

Todos los estudios presentados en este trabajo se realizaron a partir de datos disponibles en el dominio público y usando programas computacionales de acceso libre. Por estas razones y en beneficio de la ciencia abierta, el código generado para el desarrollo, uso e implementación de todas las herramientas introducidas, se encuentran disponibles en repositorios públicos y accesibles a través de la web.

XII. PERSPECTIVAS

Las metodologías presentadas en esta tesis se están aplicando en proyectos multidisciplinarios que incluyen la identificación de nuevas interacciones entre moléculas pequeñas y dianas de interés terapéutico. Ejemplos de proyectos son:

- El empleo de la SB-DFP en búsquedas por similitud sobre quimiotecas diversas para la identificación de nuevos inhibidores de DNMT1 y de la proteasa principal del SARS-CoV-2.

- El uso de ETP en la elucidación del potencial mecanismo de acción de agentes desmetilantes del DNA, como parte de una colaboración con la Dra. Laura Álvarez y la Dra. Mayra Antúnez, ambas en la Universidad Autónoma del Estado de Morelos (UAEM).

- La reevaluación de los modelos de acoplamiento de complejos proteína-ligando mediante la función de puntuación basada en ECIF para la identificación de nuevos estabilizadores de tubulina y cristalina gamma. Esto es parte de colaboraciones con el Dr. Carlos M. Cerda García y el Q.C. Edgar López (CINVESTAV), y la Dra. Laura Domínguez (UNAM), respectivamente.

- La combinación de uno de los modelos implementados en ETP y la función de puntuación generada a partir de ECIF para la identificación de inhibidores potentes de DNMT1, lo cual forma parte del proyecto de tesis de dos estudiantes de licenciatura en la Facultad de Química de la UNAM bajo la supervisión del Dr. José Luis Medina: Jocelyn Salazar y Alexis Padilla. ETP y ECIF también serán empleados en el diseño y selección de compuestos sintetizados por el grupo del Dr. Alexandre Gagnon (Universidad de Quebec en Montreal, Canadá) como potenciales inhibidores de DNMTs.

Además, el método de estandarización de estructuras propuesto se está usando como punto de partida para el análisis quimioinformático de quimiotecas de productos naturales y compuestos asociados a factores de transcripción de *Pseudomonas aeruginosa*. Este estudio es realizado en colaboración con el Dr. Fabian López y el Q.F. Felipe Victoria (Universidad Nacional de Colombia).

El uso de la SB-DFP como plantilla en búsquedas por similitud debe ser evaluado en forma sistemática para un mayor número de dianas biológicas y considerando las variables involucradas en su aplicación: el número de compuestos, la huella digital molecular y el conjunto de referencia usados para su construcción, así como la métrica de similitud empleada para realizar las búsquedas.

Tanto los modelos predictivos implementados en ETP y la función de puntuación generada a partir de ECIF fueron construidos usando los datos quimiogenómicos públicos más recientes, por lo que estos modelos deben ser reentrenados y reevaluados a medida que se disponga de nuevos datos. Esto puede hacerse en periodos anuales buscando financiamiento a través de programas de la UNAM, CONACyT u otras entidades públicas o privadas.

XIII. REFERENCIAS

Altman,N.S. (1992) An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression. *Am. Stat.*, **46**, 175–185.

Amaro,R.E. *et al.* (2018) Ensemble Docking in Drug Discovery. *Biophys. J.*, **114**, 2271–2278.

Awale,M. and Reymond,J.-L. (2019) Polypharmacology Browser PPB2: Target Prediction Combining Nearest Neighbors with Machine Learning. *J. Chem. Inf. Model.*, **59**, 10–17.

Backman,T.W.H. *et al.* (2011) ChemMine tools: An Online Service for Analyzing and Clustering Small Molecules. *Nucleic Acids Res.*, **39**, W486–W491.

Bajorath,J. (2013) A perspective on computational chemogenomics. *Mol. Inform.*, **32**, 1025–1028.

Bajusz,D. *et al.* (2015) Why is Tanimoto Index an Appropriate Choice for Fingerprint-Based Similarity Calculations? *J. Cheminform.*, **7**, 20.

Ballester,P.J. (2011) Ultrafast Shape Recognition: Method and Applications. *Future Med. Chem.*, **3**, 65–78.

Ballester,P.J. and Mitchell,J.B.O. (2010) A Machine Learning Approach to Predicting Protein–Ligand Binding Affinity with Applications to Molecular Docking. *Bioinformatics*, **26**, 1169–1175.

Banegas-Luna,A.J. *et al.* (2019) BRUSELAS: HPC Generic and Customizable Software Architecture for 3D Ligand-Based Virtual Screening of Large Molecular Databases. *J. Chem. Inf. Model.*, **59**, 2805–2817.

Bender,A. and Cortes-Ciriano,I. (2021) Artificial Intelligence in Drug Discovery: What is Realistic, What are Illusions? Part 2: A Discussion of Chemical and Biological Data. *Drug Discov. Today*. *En prensa.*

Bento,A.P. *et al.* (2014) The ChEMBL Bioactivity Database: An Update. *Nucleic Acids Res.*, **42**, D1083–D1090.

Bienfait,B. and Ertl,P. (2013) JSME: A Free Molecule Editor in JavaScript. *J. Cheminform.*, **5**, 24.

Biswas,S. and Rao,C.M. (2018) Epigenetic Tools (The Writers, The Readers and The Erasers) and Their Implications in Cancer Therapy. *Eur. J. Pharmacol.*, **837**, 8–24.

Bleicher,K. (2002) Chemogenomics: Bridging a Drug Discovery Gap. *Curr. Med. Chem.*, **9**, 2077–2084.

Booth,B. and Zemmel,R. (2004) Prospects for Productivity. *Nat. Rev. Drug Discov.*, **3**, 451–456.

Boyles,F. *et al.* (2020) Learning From the Ligand: Using Ligand-Based Features to Improve Binding Affinity Prediction. *Bioinformatics*, **36**, 758–764.

Bredel,M. and Jacoby,E. (2004) Chemogenomics: An Emerging Strategy for Rapid Target and Drug Discovery. *Nat. Rev. Genet.*, **5**, 262–275.

Brindisi,M. *et al.* (2020) Old but Gold: Tracking the New Guise of Histone Deacetylase 6 (HDAC6) Enzyme as a Biomarker and Therapeutic Target in Rare Diseases. *J. Med. Chem.*, **63**, 23–39.

Burgess,R.J. and Zhang,Z. (2013) Histone Chaperones in Nucleosome Assembly and Human Disease. *Nat. Struct. Mol. Biol.*, **20**, 14–22.

Cacabelos,R. and Torrellas,C. (2014) Epigenetic Drug Discovery for Alzheimer's Disease. *Expert Opin. Drug Discov.*, **9**, 1059–1086.

Cang,Z. *et al.* (2018) Representability of Algebraic Topology for Biomolecules in Machine Learning Based Scoring and Virtual Screening. *PLOS Comput. Biol.*, **14**, e1005929.

Cang,Z. and Wei,G. (2017) TopologyNet: Topology Based Ddeep Convolutional and Multi-Task Neural Networks for Biomolecular Property Predictions. *PLoS Comput. Biol.*, **13**, 1–27.

Chang,A. *et al.* (2015) BRENDA in 2015: Exciting Developments in its 25th Year of Existence. *Nucleic Acids Res.*, **43**, D439–D446.

Chávez-Hernández,A.L. *et al.* (2020) Fragment Library of Natural Products and Compound Databases for Drug Discovery. *Biomolecules*, **10**, 1518.

Chávez-Hernández,A.L. *et al.* (2020) A Fragment Library of Natural Products and its Comparative Chemoinformatic Characterization. *Mol. Inform.*, **39**, 2000050.

Chemical Computing Group (2018) Molecular Operating Environment.

Chen,H. *et al.* (2018) The Rise of Deep Learning in Drug Discovery. *Drug Discov. Today*, **23**, 1241–1250.

Chong,C.R. and Sullivan,D.J. (2007) New Uses for Old Drugs. *Nature*, **448**, 645–646.

Clark,M. *et al.* (1989) Validation of the General Purpose Tripos 5.2 Force Field. *J. Comput. Chem.*, **10**, 982–1012.

Cortes,C. and Vapnik,V. (1995) Support-Vector Networks. *Mach. Learn.*, **20**, 273–297.

Daina,A. *et al.* (2019) SwissTargetPrediction: Updated Data and New Features for Efficient Prediction of Protein Targets of Small Molecules. *Nucleic Acids Res.*, **47**, W357–W364.

Douguet,D. (2010) e-LEA3D: A computational-Aided Drug Design Web Server. *Nucleic Acids Res.*, **38**, 615–621.

Duan,J. *et al.* (2010) Analysis and Comparison of 2D Fingerprints: Insights into Database Screening Performance Using Eight Fingerprint Methods. *J. Mol. Graph. Model.*, **29**, 157–170.

Durant,J.L. *et al.* (2002) Reoptimization of MDL Keys for Use in Drug Discovery. *J. Chem. Inf. Comput. Sci.*, **42**, 1273–1280.

Durrant,J.D. and McCammon,J.A. (2011) NNScore 2.0: A Neural-Network Receptor-Ligand Scoring Function. *J. Chem. Inf. Model.*, **51**, 2897–2903.

Esteller,M. (2008) Epigenetics in Cancer. *N. Engl. J. Med.*, **358**, 1148–1159.

Fernández-De Gortari,E. *et al.* (2017) Database Fingerprint (DFP): An Approach to Represent Molecular Databases. *J. Cheminform.*, **9**, 1–9.

Finn,P.W. and Morris,G.M. (2013) Shape-Based Similarity Searching in Chemical Databases. *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, **3**, 226–241.

Fleming,N. (2018) How Artificial Intelligence is Changing Drug Discovery. *Nature*, **557**, S55–S57.

Fourches,D. *et al.* (2015) Curation of Chemogenomics Data. *Nat. Chem. Biol.*, **11**, 535–535.

Friedman,J. (2001) Greedy Function Approximation : A Gradient Boosting Machine *Ann. Stat.*, **29**, 1189–1232.

Friesner,R.A. *et al.* (2006) Extra Precision Glide: Docking and Scoring Incorporating a Model of Hydrophobic Enclosure for Protein−Ligand Complexes. *J. Med. Chem.*, **49**, 6177–6196.

Ganesan,A. *et al.* (2019) The Timeline of Epigenetic Drug Discovery: From Reality to Dreams. *Clin. Epigenetics*, **11**, 1–17.

Gasteiger,J. (2020) Chemistry in Times of Artificial Intelligence. *ChemPhysChem*, **21**, 2233–2242.

Gaudet,P. *et al.* (2013) neXtProt: Organizing Protein Knowledge in the Context of Human Proteome Projects. *J. Proteome Res.*, **12**, 293–298.

Gaulton,A. *et al.* (2012) ChEMBL: A Large-Scale Bioactivity Database for Drug Discovery. *Nucleic Acids Res.*, **40**, D1100–D1107.

Gaulton,A. *et al.* (2017) The ChEMBL Database in 2017. *Nucleic Acids Res.*, **45**, D945–D954.

Gilson,M.K. *et al.* (2016) BindingDB in 2015: A Public Database for Medicinal Chemistry, Computational Chemistry and Systems Pharmacology. *Nucleic Acids Res.*, **44**, D1045–D1053.

González-Medina,M. *et al.* (2017) Open Chemoinformatic Resources to Explore the Structure, Properties and Chemical Space of Molecules. *RSC Adv.*, **7**, 54153–54163.

Grant,J.A. and Pickup,B.T. (1995) A Gaussian Description of Molecular Shape. *J. Phys. Chem.*, **99**, 3503–3510.

Guo,L. *et al.* (2014) A Comparison of Various Optimization Algorithms of Protein–Ligand Docking Programs by Fitness Accuracy. *J. Mol. Model.*, **20**, 2251.

Hamad,S. *et al.* (2019) HitPickV2: A Web Server to Predict Targets of Chemical Compounds. *Bioinformatics*, **35**, 1239–1240.

Hassan-Harrirou,H. *et al.* (2020) RosENet: Improving Binding Affinity Prediction by Leveraging Molecular Mechanics Energies with an Ensemble of 3D Convolutional Neural Networks. *J. Chem. Inf. Model.*, **60**, 2791–2802.

Hawkins,P.C.D. *et al.* (2007) Comparison of Shape-Matching and Docking as Virtual Screening Tools. *J. Med. Chem.*, **50**, 74–82.

Hert,J. *et al.* (2004) Comparison of Fingerprint-Based Methods for Virtual Screening Using Multiple Bioactive Reference Structures. *J. Chem. Inf. Comput. Sci.*, **44**, 1177–1185.

Hert,J. *et al.* (2006) New Methods for Ligand-Based Virtual Screening: Use of Data Fusion and Machine Learning to Enhance the Effectiveness of Similarity Searching. *J. Chem. Inf. Model.*, **46**, 462–470.

Hopfield,J.J. (1982) Neural Networks and Physical Systems with Emergent Collective Computational Abilities. *Proc. Natl. Acad. Sci.*, **79**, 2554–2558.

Hu,G. *et al.* (2012) Performance Evaluation of 2D Fingerprint and 3D Shape Similarity Methods in Virtual Screening. *J. Chem. Inf. Model.*, **52**, 1103–1113.

Irwin,J.J. *et al.* (2009) Automated Docking Screens: A Feasibility Study. *J. Med. Chem.*, **52**, 5712–5720.

Irwin,J.J. *et al.* (2020) ZINC20—A Free Ultralarge-Scale Chemical Database for Ligand Discovery. *J. Chem. Inf. Model.*, **60**, 6065–6073.

Jaccard,P. (1901) Étude Comparative de la Distribution Florale Dans une Portion des Alpes et des Jura. *Bull. del la Société Vaudoise des Sci. Nat.*, **37**, 547–579.

Jacob,L. and Vert,J.-P. (2008) Protein-Ligand Interaction Prediction: An Improved Chemogenomics Approach. *Bioinformatics*, **24**, 2149–2156.

Januar,V. *et al.* (2015) Epigenetics and Depressive Disorders: A Review of Current Progress and Future Directions. *Int. J. Epidemiol.*, **44**, 1364–1387.

Jiménez,J. *et al.* (2018) K DEEP : Protein–Ligand Absolute Binding Affinity Prediction via 3D-Convolutional Neural Networks. *J. Chem. Inf. Model.*, **58**, 287–296.

Juárez-Mercado,K.E. *et al.* (2020) Expanding the Structural Diversity of DNA Methyltransferase Inhibitors. *Pharmaceuticals*, **14**, 17-.

Keiser,M.J. *et al.* (2007) Relating Protein Pharmacology by Ligand Chemistry. *Nat. Biotechnol.*, **25**, 197–206.

Kim,S. *et al.* (2019) PubChem 2019 Update: Improved Access to Chemical Data. *Nucleic Acids Res.*, **47**, D1102–D1109.

Klambauer,G. *et al.* (2015) Rchemcpp: A Web Service For Structural Analoging in ChEMBL, Drugbank and the Connectivity Map: Fig. 1. *Bioinformatics*, **31**, 3392–3394.

Klopmand,G. (1992) Concepts and Applications of Molecular Similarity, by Mark A. Johnson and Gerald M. Maggiora, eds., John Wiley &; Sons, New York, 1990, 393 pp. *J. Comput. Chem.*, **13**, 539–540.

Knegtel,R.M.. *et al.* (1997) Molecular Docking to Ensembles of Protein Structures Edited by B. Honig. *J. Mol. Biol.*, **266**, 424–440.

Koes,D.R. and Camacho,C.J. (2012) ZINCPharmer: Pharmacophore Search of the ZINC Database. *Nucleic Acids Res.*, **40**, 409–414.

Kong,R. *et al.* (2020) COVID-19 Docking Server: A Meta Server for Docking Small Molecules, Peptides and Antibodies Against Potential Targets of COVID-19. *Bioinformatics*, **36**, 5109–5111.

Kristensen,T.G. *et al.* (2013) Methods for Similarity-Based Virtual Screening. *Comput. Struct. Biotechnol. J.*, **5**, e201302009.

Küçükali,C.İ. *et al.* (2015) Epigenetics of Multiple Sclerosis: An Updated Review. *NeuroMolecular Med.*, **17**, 83–96.

Labbé,C.M. *et al.* (2015) MTiOpenScreen: A Web Server for Structure-Based Virtual Screening. *Nucleic Acids Res.*, **43**, W448–W454.

Leach,A.R. (1994) Ligand Docking to Proteins with Discrete Side-Chain Flexibility. *J. Mol. Biol.*, **235**, 345–356.

Lee,A. and Kim,D. (2019) CRDS: Consensus Reverse Docking System for Target Fishing. *Bioinformatics*, **36,** 959–960**.**

Lee,K. *et al.* (2017) Utilizing Random Forest QSAR Models with Optimized Parameters for Target Identification and its Application to Target-Fishing Server. *BMC Bioinformatics*, **18**, 567.

de Lera,A.R. and Ganesan,A. (2020) Two-Hit Wonders: The Expanding Universe of Multitargeting Epigenetic Agents. *Curr. Opin. Chem. Biol.*, **57**, 135–154.

Li,G.B. *et al.* (2013) ID-score: A New Empirical Scoring Function Based on a Comprehensive Set of Descriptors Related to Protein-Ligand Interactions. *J. Chem. Inf. Model.*, **53**, 592–600.

Li,H. *et al.* (2020) Machine-Learning Scoring Functions for Structure-Based Drug Lead Optimization. *WIREs Comput. Mol. Sci.*, **10**, 1–20.

Li,Q. (2019) Virtual Screening of Small-Molecule Libraries. In, Trabocchi A. and Lenci E. (eds) *Small Molecule Drug Discovery*., Elsevier Inc., pp. 103–125

Li,Y. *et al.* (2014) Comparative Assessment of Scoring Functions on an Updated Benchmark: 1. Compilation of the Test Set. *J. Chem. Inf. Model.*, **54**, 1700–1716.

Lill,M. (2013) Virtual Screening in Drug Design. In, Kortagere S. (eds) *In Silico Models for Drug Discovery*. Methods in Molecular Biology (Methods and Protocols), vol 993. Humana Press, Totowa, NJ. pp. 1–12.

Lionta,E. *et al.* (2014) Structure-Based Virtual Screening for Drug Discovery: Principles, Applications and Recent Advances. *Curr. Top. Med. Chem.*, **14**, 1923–1938.

Lipinski,C.A. *et al.* (1997) Experimental and Computational Approaches to Estimate Solubility and Permeability in Drug Discovery and Development Settings. *Adv. Drug Deliv. Rev.*, **23**, 3–25.

Liu,J. and Wang,R. (2015) Classification of Current Scoring Functions. *J. Chem. Inf. Model.*, **55**, 475–482.

Liu,T. *et al.* (2007) BindingDB: A Web-Accessible Database of Experimentally Determined Protein-Ligand Binding Affinities. *Nucleic Acids Res.*, **35**, D198–D201.

Liu,Z. *et al.* (2015) PDB-Wide Collection of Binding Data: Current Status of the PDBbind Database. *Bioinformatics*, **31**, 405–412.

Lo,Y.-C. *et al.* (2018) Machine Learning in Chemoinformatics and Drug Discovery. *Drug Discov. Today*, **23**, 1538–1546.

López-López,E. *et al.* (2021) Informatics for Chemistry, Biology, and Biomedical Sciences. *J. Chem. Inf. Model.*, **61**, 26–35.

Lu,J. *et al.* (2019) Incorporating Explicit Water Molecules and Ligand Conformation Stability in Machine-Learning Scoring Functions. *J. Chem. Inf. Model.*, **59**, 4540–4549.

Lyne,P.D. (2002) Structure-Based Virtual Screening: An Overview. *Drug Discov. Today*, **7**, 1047–1055.

Maggiora,G. *et al.* (2020) From Qualitative to Quantitative Analysis of Activity and Property Landscapes. *J. Chem. Inf. Model.*, **60**, 5873–5880.

Al Mahmud,R. *et al.* (2018) A Survey of Web-Based Chemogenomic Data Resources. In: Brown J. (eds) Computational Chemogenomics. Methods in Molecular Biology, vol 1825. Humana Press, New York, NY., pp. 3–62.

Mathias,S.L. *et al.* (2013) The CARLSBAD Database: A Confederated Database of Chemical Bioactivities. *Database*, **2013**, bat044–bat044.

Mayr,A. *et al.* (2018) Large-Scale Comparison of Machine Learning Methods for Drug Target Prediction on ChEMBL. *Chem. Sci.*, **9**, 5441–5451.

Mayran,A. and Drouin,J. (2018) Pioneer Transcription Factors Shape the Epigenetic Landscape. *J. Biol. Chem.*, **293**, 13795–13804.

Medina-Franco,J.L. (2013) Activity Cliffs: Facts or Artifacts? *Chem. Biol. Drug Des.*, **81**, 553–556.

Medina-Franco,J.L. *et al.* (2013) Shifting from the Single to the Multitarget Paradigm in Drug Discovery. *Drug Discov. Today*, **18**, 495–501.

Mendez,D. *et al.* (2019) ChEMBL: Towards Direct Deposition of Bioassay Data. *Nucleic Acids Res.*, **47**, D930–D940.

Meng,X.-Y. *et al.* (2011) Molecular Docking: A Powerful Approach for Structure-Based Drug Discovery. *Curr. Comput. Aided-Drug Des.*, **7**, 146–157.

Montaruli,M. *et al.* (2019) Accelerating Drug Discovery by Early Protein Drug Target Prediction Based on a Multi-Fingerprint Similarity Search †. *Molecules*, **24**, 2233.

Morris,G.M. *et al.* (2009) AutoDock4 and AutoDockTools4: Automated Docking with Selective Receptor Flexibility. *J. Comput. Chem.*, **30**, 2785–2791.

Mullard,A. (2021) 2020 FDA Drug Approvals. *Nat. Rev. Drug Discov.*

Nguyen,D.D. *et al.* (2017) Rigidity Strengthening: A Mechanism for Protein-Ligand Binding. *J. Chem. Inf. Model.*, **57**, 1715–1721.

Nguyen,D.D. and Wei,G.W. (2019a) AGL-Score: Algebraic Graph Learning Score for Protein-Ligand Binding Scoring, Ranking, Docking, and Screening. *J. Chem. Inf. Model.*, **59**, 3291–3304.

Nguyen,D.D. and Wei,G.W. (2019b) DG-GL: Differential Geometry-Based Geometric Learning of Molecular Datasets. *Int. J. Numer. Method. Biomed. Eng.*, **35**, 1–24.

Norinder,U. *et al.* (2019) Conformal Prediction of HDAC Inhibitors. *SAR QSAR Environ. Res.*, **30**, 265–277.

Oprea,T.I. *et al.* (2018) Unexplored Therapeutic Opportunities in the Human Genome. *Nat. Rev. Drug Discov.*, **17**, 317–332.

Pantziarka,P. *et al.* (2018) New Uses for Old Drugs. *BMJ*, **361**, k2701.

Pasznik,P. *et al.* (2019) Potential Off-Target Effects of Beta-Blockers on Gut Hormone Receptors: In Silico Study Including GUT-DOCK—A Web Service for Small-Molecule Docking. *PLoS One*, **14**, e0210705.

Patel,L. *et al.* (2020) Machine Learning Methods in Drug Discovery. *Molecules*, **25**.

Paul,D. *et al.* (2020) Artificial Intelligence in Drug Discovery and Development. *Drug Discov. Today*. *En prensa*.

Paul,S.M. *et al.* (2010) How to improve R&D Productivity: The Pharmaceutical Industry's Grand Challenge. *Nat. Rev. Drug Discov.*, **9**, 203–214.

Pérez-Sánchez,H. *et al.* (2020) DIA-DB: A Database and Web Server for the Prediction of Diabetes Drugs. *J. Chem. Inf. Model.*, **60**, 4124–4130.

Pires,D.E. V *et al.* (2020) EasyVS: A User-Friendly Web-Based Tool for Molecule Library Selection and Structure-Based Virtual Screening. *Bioinformatics*, **36**, 4200–4202.

Prieto-Martínez,F.D. *et al.* (2018) Acoplamiento Molecular: Avances Recientes y Retos. *TIP Rev. Espec. en Ciencias Químico-Biológicas*, **21**, 65-87.

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

Reymond,J.-L. (2015) The Chemical Space Project. *Acc. Chem. Res.*, **48**, 722–730.

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

Rognan,D. (2007) Chemogenomic Approaches to Rational Drug Design. *Br. J. Pharmacol.*, **152**, 38–52.

Ruiz-Carmona,S. *et al.* (2014) rDock: A Fast, Versatile and Open Source Program for Docking Ligands to Proteins and Nucleic Acids. *PLoS Comput. Biol.*, **10**, e1003571.

Saldívar-González,F.I. *et al.* (2017) Getting SMARt in Drug Discovery: Chemoinformatics Approaches for Mining Structure–Multiple Activity Relationships. *RSC Adv.*, **7**, 632–641.

Sánchez-Cruz,N. *et al.* (2020) Extended Connectivity Interaction Features: Improving Binding Affinity Prediction Through Chemical Description. *Bioinformatics*. *En prensa*.

Sánchez-Cruz,N. *et al.* (2019) Functional Group and Diversity Analysis of BIOFACQUIM: A Mexican Natural Product Database. *F1000Research*, **8**, 2071.

Sánchez-Cruz,N. and Medina-Franco,J.L. (2021a) Epigenetic Target Fishing with Accurate Machine Learning Models. *J. Med. Chem.* En prensa.

Sánchez-Cruz,N. and Medina-Franco,J.L. (2021b) Epigenetic Target Profiler: A Web Server to Predict Epigenetic Targets of Small Molecules. *J. Chem. Inf. Model*. En prensa.

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

Sandal,M. *et al.* (2013) GOMoDo: A GPCRs Online Modeling and Docking Webserver. *PLoS One*, **8**, e74092.

Santibáñez-Morán,M.G. *et al.* (2020) Consensus Virtual Screening of Dark Chemical Matter and Food Chemicals Uncover Potential Inhibitors of SARS-CoV-2 Main Protease. *RSC Adv.*, **10**, 25089–25099.

Schneider,M. *et al.* (2020) Towards Accurate High-Throughput Ligand Affinity Prediction by Exploiting Structural Ensembles, Docking Metrics and Ligand Similarity. *Bioinformatics*, **36**, 160–168.

Sessions,Z. *et al.* (2020) Recent Progress on Cheminformatics Approaches to Epigenetic Drug Discovery. *Drug Discov. Today*, **25**, 2268–2276.

Shemetulskis,N.E. *et al.* (1996) Stigmata: An Algorithm to Determine Structural Commonalities in Diverse Datasets. *J. Chem. Inf. Comput. Sci.*, **36**, 862–71.

da Silveira,N.J.F. *et al.* (2019) Web Services for Molecular Docking Simulations., pp. 221–229.

Singh,N. *et al.* (2020) Virtual Screening Web Servers: Designing Chemical Probes and Drug Candidates in the Cyberspace. *Brief. Bioinform. En prensa*.

Skuta,C. *et al.* (2017) Probes & Drugs Portal: An Interactive, Open Data Resource for Chemical Biology. *Nat. Methods*, **14**, 759–760.

Smietana,K. *et al.* (2015) Improving R&D productivity. *Nat. Rev. Drug Discov.*, **14**, 455–456.

Sorokina,M. *et al.* (2021) COCONUT Online: Collection of Open Natural Products Database. *J. Cheminform.*, **13**, 2.

Sorokina,M. and Steinbeck,C. (2020) Review on Natural Products Databases: Where to Find Data in 2020. *J. Cheminform.*, **12**, 20.

Speck-Planche,A. and Scotti,M.T. (2019) BET Bromodomain Inhibitors: Fragment-Based in Silico Design Using Multi-Target QSAR Models. *Mol. Divers.*, **23**, 555–572.

Stepniewska-Dziubinska,M.M. *et al.* (2018) Development and Evaluation of a Deep Learning Model for Protein–Ligand Binding Affinity Prediction. *Bioinformatics*, **34**, 3666–3674.

Sterling,T. and Irwin,J.J. (2015) ZINC 15 – Ligand Discovery for Everyone. *J. Chem. Inf. Model.*, **55**, 2324–2337.

Strausberg,R.L. (2003) From Knowing to Controlling: A Path from Genomics to Drugs Using Small Molecule Probes. *Science*, **300**, 294–295.

Su,M. *et al.* (2019) Comparative Assessment of Scoring Functions: The CASF-2016 Update. *J. Chem. Inf. Model.*, **59**, 895–913.

Sun,J. *et al.* (2017) ExCAPE-DB: An Integrated Large Scale Dataset Facilitating Big Data Analysis in Chemogenomics. *J. Cheminform.*, **9**, 17.

Sunseri,J. and Koes,D.R. (2016) Pharmit: Interactive Exploration of Chemical Space. *Nucleic Acids Res.*, **44**, W442–W448.

Sushko,I. *et al.* (2010) Applicability Domains for Classification Problems: Benchmarking of Distance to Models for Ames Mutagenicity Set. *J. Chem. Inf. Model.*, **50**, 2094–2111.

Sydow,D. *et al.* (2019) Advances and Challenges in Computational Target Prediction. *J. Chem. Inf. Model.*, **59**, 1728–1742.

Syuib,M. *et al.* (2014) Comparison of Similarity Coefficients for Chemical Database Retrieval. *Proc. - 1st Int. Conf. Artif. Intell. Model. Simulation, AIMS 2013*, 129–133.

Tetko,I. V. *et al.* (2008) Critical Assessment of QSAR Models of Environmental Toxicity against Tetrahymena pyriformis: Focusing on Applicability Domain and Overfitting by Variable Selection. *J. Chem. Inf. Model.*, **48**, 1733–1746.

Thafar,M. *et al.* (2019) Comparison Study of Computational Prediction Tools for Drug-Target Binding Affinities. *Front. Chem.*, **7**, 782.

Tin Kam Ho (1998) The Random Subspace Method for Constructing Decision Forests. *IEEE Trans. Pattern Anal. Mach. Intell.*, **20**, 832–844.

Tran,T.D. *et al.* (2020) Lessons from Exploring Chemical Space and Chemical Diversity of Propolis Components. *Int. J. Mol. Sci.*, **21**, 4988.

Trott,O. and Olson,A.J. (2010) AutoDock Vina: improving the Speed and Accuracy of Docking with a New Scoring Function, Efficient Optimization, and Multithreading. *J. Comput. Chem.*, **31**, 455–61.

Tyagi,M. *et al.* (2016) Chromatin Remodelers: We are the Drivers!! *Nucleus*, **7**, 388–404.

Vamathevan,J. *et al.* (2019) Applications of Machine Learning in Drug Discovery and Development. *Nat. Rev. Drug Discov.*, **18**, 463–477.

Wang,C. and Zhang,Y. (2017) Improving Scoring-Docking-Screening Powers of Protein–Ligand Scoring Functions Using Random Forest. *J. Comput. Chem.*, **38**, 169–177.

Wang,F. *et al.* (2019) ACID: A Free Tool for Drug Repurposing Using Consensus Inverse Docking Strategy. *J. Cheminform.*, **11**, 73.

Wang,J.C. *et al.* (2012) idTarget: A Web Server for Identifying Protein Targets of Small Chemical Molecules with Robust Scoring Functions and a Divide-and-Conquer Docking Approach. *Nucleic Acids Res.*, **40**, 393–399.

Wang,R. *et al.* (2002) Further Development and Validation of Empirical Scoring Functions for Structure-Based Binding Affinity Prediction. *J. Comput. Aided. Mol. Des.*, **16**, 11–26.

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

Willett,P. (2006) Similarity-Based Virtual Screening Using 2D Fingerprints. *Drug Discov. Today*, **11**, 1046–1053.

Wishart,D.S. (2006) DrugBank: A Comprehensive Resource for In Silico Drug Discovery and Exploration. *Nucleic Acids Res.*, **34**, D668–D672.

Wishart,D.S. *et al.* (2018) DrugBank 5.0: A Major Update to the DrugBank Database for 2018. *Nucleic Acids Res.*, **46**, D1074–D1082.

Wójcikowski,M. *et al.* (2019) Development of a Protein–Ligand Extended Connectivity (PLEC) Fingerprint and its Application for Binding Affinity Predictions. *Bioinformatics*, **35**, 1334–1341.

Wu,J. *et al.* (2018) WDL-RF: Predicting Bioactivities of Ligand Molecules Acting with G Protein-Coupled Receptors by Combining Weighted Deep Learning and Random Forest. *Bioinformatics*, **34**, 2271–2282.

Yang,S.Q. *et al.* (2020) Current Advances in Ligand-Based Target Prediction. *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, 1–21.

Yao,Z.-J. *et al.* (2016) TargetNet: A Web Service for Predicting Potential Drug–Target Interaction Profiling Via Multi-Target SAR Models. *J. Comput. Aided. Mol. Des.*, **30**, 413–424.

Zdrazil,B. *et al.* (2020) Moving Targets in Drug Discovery. *Sci. Rep.*, **10**, 20213.

Zhang,L. *et al.* (2017) From Machine Learning to Deep Learning: Progress in Machine Intelligence for Rational Drug Discovery. *Drug Discov. Today*, **22**, 1680–1685.

Zheng,L. *et al.* (2019) OnionNet: A Multiple-Layer Intermolecular-Contact-Based Convolutional Neural Network for Protein-Ligand Binding Affinity Prediction. *ACS Omega*, **4**, 15956–15965.

Zilian,D. and Sotriffer,C.A. (2013) SFCscoreRF: A Random Forest-Based Scoring Function for Improved Affinity Prediction of Protein-Ligand Complexes. *J. Chem. Inf. Model.*, **53**, 1923–1933.

Zoete,V. *et al.* (2016) SwissSimilarity: A Web Tool for Low to Ultra High Throughput Ligand-Based Virtual Screening. *J. Chem. Inf. Model.*, **56**, 1399–1404.

ANEXO 1. Artículos que reportan los resultados principales de esta tesis

doi.org/10.26434/chemrxiv.13524575.v1

Epigenetic Target Profiler: A Web Server to Predict Epigenetic Targets of Small Molecules

Norberto Sánchez-Cruz, Jose L. Medina-Franco

Submitted date: 06/01/2021 • Posted date: 07/01/2021

Licence: CC BY 4.0

Citation information: Sánchez-Cruz, Norberto; Medina-Franco, Jose L. (2021): Epigenetic Target Profiler: A Web Server to Predict Epigenetic Targets of Small Molecules. ChemRxiv. Preprint. https://doi.org/10.26434/chemrxiv.13524575.v1

Motivation: The identification of protein targets of small molecules is essential for drug discovery. With the increasing amount of chemogenomic data in the public domain, multiple ligand-based models for target prediction have emerged. However, these models are generally biased by the number of known ligands for different targets, which involves an underrepresentation of epigenetic targets. Epigenetic drug discovery is of increasing importance but there are no open tools for epigenetic target prediction. Results: We introduce Epigenetic Target Profiler (ETP), a freely accessible and easy-to-use web application for the prediction of epigenetic targets of small molecules. For a query compound, ETP predicts its bioactivity profile over a panel of 55 different epigenetic targets. To that aim, ETP uses a consensus model based on two binary classification models for each target, relying on support vector machines and built on molecular fingerprints of different design. A distance-to-model parameter related to the reliability of the predictions is included to facilitate their interpretability and assist the identification of small molecules with potential epigenetic activity.

Epigenetic Target Profiler: a web server to predict epigenetic targets of small molecules

Norberto Sánchez-Cruz[1,*] and José L. Medina-Franco[1,*]

[1]DIFACQUIM research group, Department of Pharmacy, School of Chemistry, Universidad Nacional Autónoma de México, Mexico City 04510, Mexico

*To whom correspondence should be addressed.

Abstract

Motivation: The identification of protein targets of small molecules is essential for drug discovery. With the increasing amount of chemogenomic data in the public domain, multiple ligand-based models for target prediction have emerged. However, these models are generally biased by the number of known ligands for different targets, which involves an underrepresentation of epigenetic targets. Epigenetic drug discovery is of increasing importance but there are no open tools for epigenetic target prediction.

Results: We introduce Epigenetic Target Profiler (ETP), a freely accessible and easy-to-use web application for the prediction of epigenetic targets of small molecules. For a query compound, ETP predicts its bioactivity profile over a panel of 55 different epigenetic targets. To that aim, ETP uses a consensus model based on two binary classification models for each target, relying on support vector machines and built on molecular fingerprints of different design. A distance-to-model parameter related to the reliability of the predictions is included to facilitate their interpretability and assist the identification of small molecules with potential epigenetic activity.

Availability: Epigenetic Target Profiler is freely available at http://www.epigenetictargetprofiler.com/

Contact: norberto.sc90@gmail.com (N.S-C.); medinajl@unam.mx (J.L.M-F.)

1. Introduction

The identification of protein targets for small molecules plays a key role in multiple areas of drug discovery, since it allows the prioritization of compounds for the discovery of novel inhibitors against one or a set of therapeutic targets, as well as the estimation of their off-target effects, which can be useful for the study of the polypharmacology of compounds (Anighoro *et al.*, 2014) and drug repurposing (Oprea *et al.*, 2011). The increase in the publicly available chemogenomic data over the years have led to the construction of multiple ligand-based models to predict the protein targets of small molecules, with some of these models available as web-based tools (Sam and Athri, 2019).

Current ligand-based target prediction methods assign the targets for a given small molecule based on the known targets of the most similar ligands in their training datasets, which represents a target-bias over specific protein families. Considering that chemogenomic data related to epigenetics is minimal in comparison to protein families such as kinases, G protein-coupled receptors, and ion channels (Zdrazil *et al.*, 2020), epigenetic targets are underrepresented in the currently available models. Moreover, despite the increasing relevance of these targets across several different therapeutic areas (Sessions *et al.*, 2020), there are no web-based tools to support epigenetic drug discovery.

Herein, we present Epigenetic Target Profiler (ETP), an easy-to-use and free web application for the prediction of the bioactivity profile of small molecules over a panel of 55 epigenetic targets. ETP implements the best performing model for epigenetic target prediction, as identified from a systematic comparison of machine learning models built on molecular fingerprints (FPs) of different design described in a separate work.

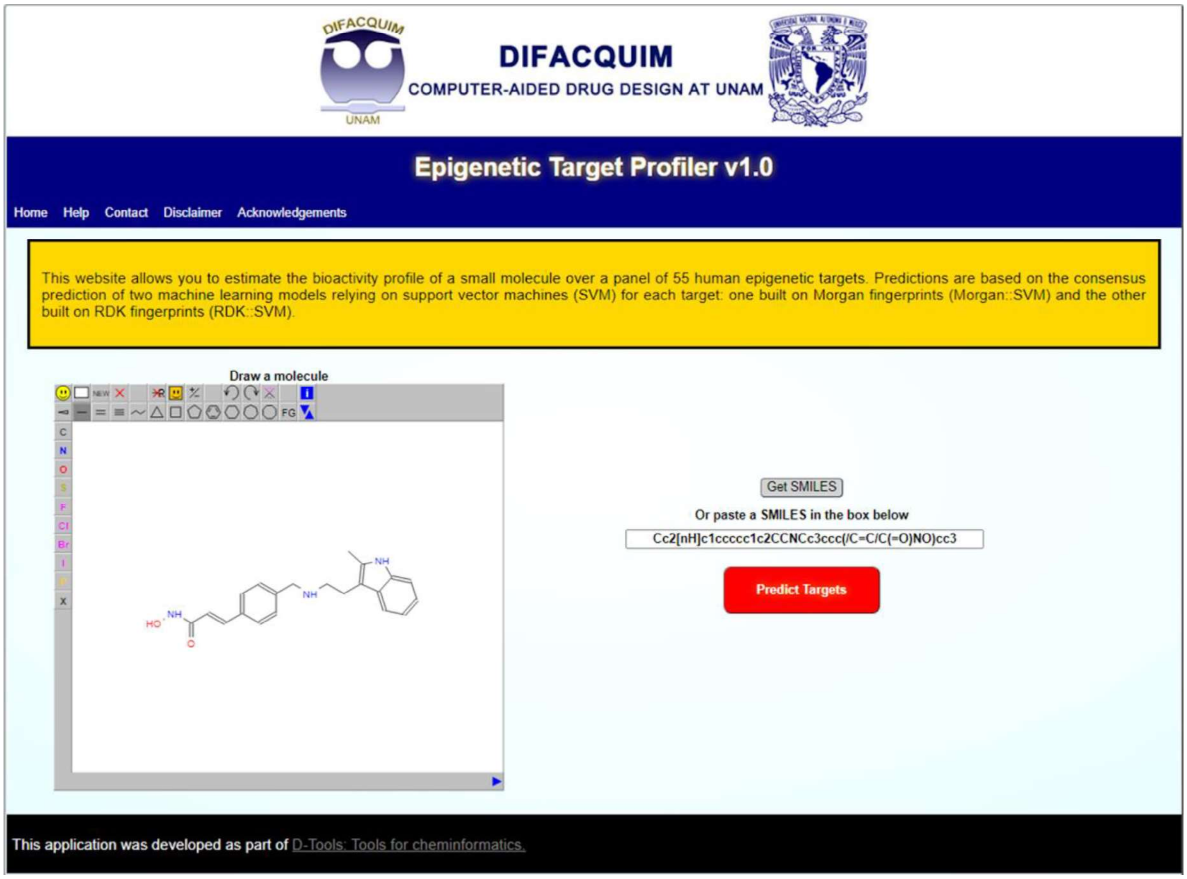
2. Approach

In a separate study, we performed a comprehensive comparison of fifteen machine learning models, derived from the combination of three different molecular FPs as compound representations, and five different machine learning algorithms, for binary classification of compounds over 55 epigenetic targets, using a quantitative measure of biological activity cutoff of 10 μM (IC_{50}, EC_{50}, K_i or K_d). We found support vector machines trained on Morgan FPs of radius 2 (Morgan::SVM) and on RDK FPs (RDK::SVM) as the two best performing models for this task. We built a consensus model by combining the predictions of these two models and examined the performance of the individual models and the derived consensus model on a distance-to-model (DM) basis by classifying each prediction into four quartiles (Q1 - Q4) according to the mean Jaccard's distance of the compound to the corresponding training set. The consensus model showed higher precision than the individual models for the prediction of active compounds for all distance quartiles, with a mean precision of 0.896, ranging from 0.923 for compounds in Q1 to 0.810 for compounds in Q4.

The three models were tested for epigenetic target prediction on ten assembled samples containing the same number of active compounds for each of the epigenetic targets. As a result, the consensus model also showed a superior performance for the correct identification of epigenetic targets, with mean precisions ranging from 0.952 for compounds in Q1 to 0.773 for compounds in Q4. The practical applicability of these model was shown by the retrospective identification of the epigenetic targets of two recently reported epigenetic inhibitors (Wilson *et al.*, 2020; Chen *et al.*, 2020). Supported on these findings, we implemented the consensus model as an easy-to-use web application, described in the following section.

3. The ETP web interface

ETP is freely accessible as a web application at http://www.epigenetictargetprofiler.com/ and all row data to reproduce it is available free of charge at figshare repository (10.6084/m9.figshare.13524368). ETP was implemented using Flask version 1.1.1 for Python as web framework and its graphical user interface (GUI) was written in HTML5, CSS and JavaScript with all major browsers supported. The GUI of ETP starts with an initial page wherein the user can either draw a query molecule using the JavaScript based JSME molecular editor (Bienfait and Ertl, 2013) and generate its corresponding SMILES by clicking on the "Get SMILES" button or directly paste the SMILES of a query compound in the cell provided for that purpose (Figure 1).


DIFACQUIM
COMPUTER-AIDED DRUG DESIGN AT UNAM
Epigenetic Target Profiler v1.0
Home
Help
Contact
Disclaimer
Acknowledgements
Draw a molecule
Get SMILES
Or paste a SMILES in the box below
Predict Targets

This website allows you to estimate the bioactivity profile of a small molecule over a panel of 55 human epigenetic targets. Predictions are based on the consensus prediction of two machine learning models relying on support vector machines (SVM) for each target: one built on Morgan fingerprints (Morgan::SVM) and the other built on RDK fingerprints (RDK::SVM).

Cc2[nH]c1ccccc1c2CCNCc3ccc(/C=C/C(=O)NO)cc3

This application was developed as part of D-Tools: Tools for cheminformatics.

Figure 1. *Graphical User Interface of Epigenetic Target Profiler. Panobinostat is shown in the JSME molecular editor and its SMILES is shown in the cell on the left.*

Following the entry of a query SMILES, the target prediction can be initiated by clicking on the “Predict Targets” button and the user will be directed to the results page. The server standardizes the input compound according to the same process described in Supplementary Section 1 and generates its Morgan and RDK FPs. For each target, if the compound is part of the target-associated compound dataset, no further processing is done and its known association is returned, otherwise the sever computes its mean Jaccard distance to the dataset using Morgan FP, classify it into a quartile accordingly (Supplementary Table S1) and performs the prediction using the Morgan::SVM and RDK::SVM models described in the previous section. The predicted targets for the query compound are those predicted by both models. This process is illustrated in Figure 2. Details on the hyperparameters for each machine learning model and their performances for each target and distance quartile in a 10-fold cross-validation are included in Supplementary Tables S2-S4.

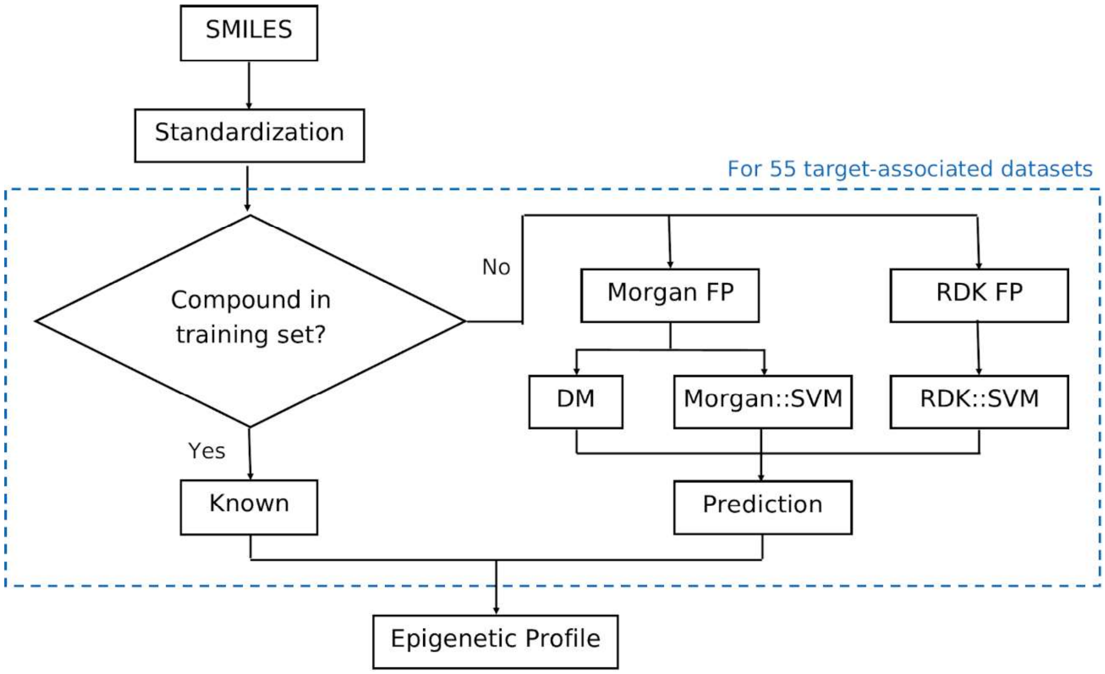
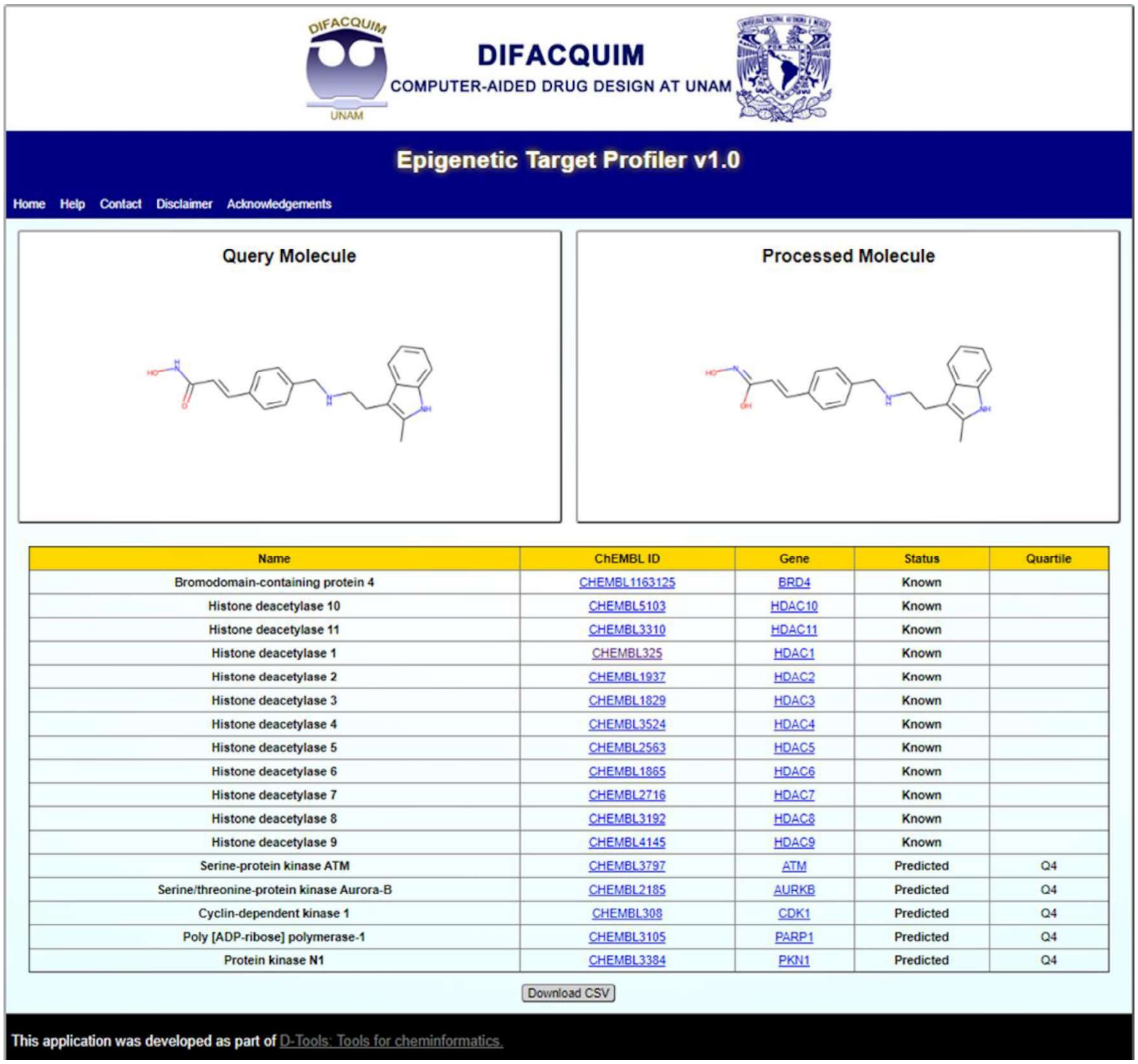


***Figure 2**. Schematic representation of the process performed by ETP.*

Once the predictions have been performed, the user is redirected to the results page, which contains two images at top and a table below them (Figure 3). The image on the left side shows the chemical structure of the query compound as interpreted from the SMILES submitted by the user, while the image on the right side depicts the chemical structure of the standardized compound as processed by the sever. The results table shows the known and predicted targets for the query compound, including five columns with additional information. The first three columns contain the name of the targets and external links to ChEMBL and GeneCards (Name, ChEMBL ID, Gene) and the last two contain information about the predictions (Status and Quartile). The Status column indicates if the association is known or predicted, while the Quartile column indicates the distance quartile (Q1 – Q4) to which the query compound belongs for each of the predictions as an estimation on its reliability. The full list of results including the predictions from the individual models for all 55 targets can be downloaded by clicking on the “Download CSV” button below the table.



DIFACQUIM

COMPUTER-AIDED DRUG DESIGN AT UNAM

Epigenetic Target Profiler v1.0

Query Molecule

Processed Molecule

Name	ChEMBL ID	Gene	Status	Quartile
Bromodomain-containing protein 4	CHEMBL1163125	BRD4	Known	
Histone deacetylase 10	CHEMBL5103	HDAC10	Known	
Histone deacetylase 11	CHEMBL3310	HDAC11	Known	
Histone deacetylase 1	CHEMBL325	HDAC1	Known	
Histone deacetylase 2	CHEMBL1937	HDAC2	Known	
Histone deacetylase 3	CHEMBL1829	HDAC3	Known	
Histone deacetylase 4	CHEMBL3524	HDAC4	Known	
Histone deacetylase 5	CHEMBL2563	HDAC5	Known	
Histone deacetylase 6	CHEMBL1865	HDAC6	Known	
Histone deacetylase 7	CHEMBL2716	HDAC7	Known	
Histone deacetylase 8	CHEMBL3192	HDAC8	Known	
Histone deacetylase 9	CHEMBL4145	HDAC9	Known	
Serine-protein kinase ATM	CHEMBL3797	ATM	Predicted	Q4
Serine/threonine-protein kinase Aurora-B	CHEMBL2185	AURKB	Predicted	Q4
Cyclin-dependent kinase 1	CHEMBL308	CDK1	Predicted	Q4
Poly [ADP-ribose] polymerase-1	CHEMBL3105	PARP1	Predicted	Q4
Protein kinase N1	CHEMBL3384	PKN1	Predicted	Q4

This application was developed as part of D-Tools: Tools for cheminformatics.

Figure S2. *Results page of Epigenetic Target Profiler. Known and predicted associations are shown for Panobinostat.*

4. **Conclusion**

ETP is an easy-to-use and free web-based tool to support epigenetic drug discovery projects.

Acknowledgements

NS-C is thankful to Consejo Nacional de Ciencia y Tecnología (CONACyT), Mexico, for the granted scholarship number 335997.

Funding

This work has been partially supported by the *Dirección General de Cómputo y de Tecnologías de Información y Comunicación* (DGTIC), UNAM that provided computational resources to use Miztli supercomputer with the project LANCAD-UNAM-DGTIC-335. The study was also supported by the *Consejo Nacional de Ciencia y Tecnología* (CONACyT), Mexico, grant 282785.

Conflict of Interest: none declared.

References

Anighoro,A. *et al.* (2014) Polypharmacology: Challenges and Opportunities in Drug Discovery. *J. Med. Chem.*, **57**, 7874–7887.

Bienfait,B. and Ertl,P. (2013) JSME: a free molecule editor in JavaScript. *J. Cheminform.*, **5**, 24.

Chen,J. *et al.* (2020) Discovery of selective HDAC/BRD4 dual inhibitors as epigenetic probes. *Eur. J. Med. Chem.*, **209**, 112868.

Oprea,T.I. *et al.* (2011) Drug repurposing from an academic perspective. *Drug Discov. Today Ther. Strateg.*, **8**, 61–69.

Sam,E. and Athri,P. (2019) Web-based drug repurposing tools: A survey. *Brief. Bioinform.*, **20**, 1–18.

Sessions,Z. *et al.* (2020) Recent progress on cheminformatics approaches to epigenetic drug discovery. *Drug Discov. Today*, **25**, 2268–2276.

Wilson,J.E. *et al.* (2020) Discovery of CPI-1612: A Potent, Selective, and Orally Bioavailable EP300/CBP Histone Acetyltransferase Inhibitor. *ACS Med. Chem. Lett.*, **11**, 1324–1329.

Zdrazil,B. *et al.* (2020) Moving targets in drug discovery. *Sci. Rep.*, **10**, 20213.

doi.org/10.26434/chemrxiv.13522313.v1

Epigenetic Target Prediction with Accurate Machine Learning Models

Norberto Sánchez-Cruz, Jose L. Medina-Franco

Submitted date: 05/01/2021 • Posted date: 06/01/2021

Licence: CC BY 4.0

Citation information: Sánchez-Cruz, Norberto; Medina-Franco, Jose L. (2021): Epigenetic Target Prediction with Accurate Machine Learning Models. ChemRxiv. Preprint. https://doi.org/10.26434/chemrxiv.13522313.v1

Epigenetic targets are a significant focus for drug discovery research, as demonstrated by the eight approved epigenetic drugs for treatment of cancer and the increasing availability of chemogenomic data related to epigenetics. This data represents a large amount of structure-activity relationships that has not been exploited thus far for the development of predictive models to support medicinal chemistry efforts. Herein, we report the first large-scale study of 26318 compounds with a quantitative measure of biological activity for 55 protein targets with epigenetic activity. Through a systematic comparison of machine learning models trained on molecular fingerprints of different design, we built predictive models with high accuracy for the epigenetic target profiling of small molecules. The models were thoroughly validated showing mean precisions up to 0.952 for the epigenetic target prediction task. Our results indicate that the herein reported models have considerable potential to identify small molecules with epigenetic activity. Therefore, our results were implemented as freely accessible and easy-to-use web application.

Epigenetic Target Prediction with Accurate Machine Learning Models

Norberto Sánchez-Cruz* and José L. Medina-Franco*

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

KEYWORDS: Cheminformatics; Chemogenomics; Drug Discovery; Epigenetics; Epi-Informatics; Machine learning; Structure-activity relationships.

ABSTRACT:

Epigenetic targets are a significant focus for drug discovery research, as demonstrated by the eight approved epigenetic drugs for treatment of cancer and the increasing availability of chemogenomic data related to epigenetics. This data represents a large amount of structure-activity relationships that has not been exploited thus far for the development of predictive models to support medicinal chemistry efforts. Herein, we report the first large-scale study of 26318 compounds with a quantitative measure of biological activity for 55 protein targets with epigenetic activity. Through a systematic comparison of machine learning models trained on molecular fingerprints of different design, we built predictive models with high accuracy for the epigenetic target profiling of small molecules. The models were thoroughly validated showing mean precisions up to 0.952 for the epigenetic target prediction task. Our results indicate that the herein reported models have considerable potential to identify small molecules with epigenetic activity. Therefore, our results were implemented as freely accessible and easy-to-use web application.

INTRODUCTION

Since the introduction of the term *epigenetics* by Conrad Waddington in 1942 to denote the mechanisms that relate genotype to phenotype,[1] the term has been used with multiple meanings, going from the classic definition that refers to epigenetics as the study of the alterations in the biological phenotype without underlying changes in the DNA sequence,[2] to one of the most recent and general definitions: "the structural adaptation of chromosomal regions to register, signal, or perpetuate altered activity states."[3] At the molecular level, this adaptation involves the reversible modification of nucleic acids and histones. These modifications are catalyzed by a plethora of proteins, which could be considered as the core epigenetic targets, and that are classified into three main groups: (a) writers - enzymes capable of adding chemical groups to nucleic acids and histones - such as DNA methyltransferases (DNMTs), histone methyltransferases (HMTs) and histone acetyltransferases (HATs), (b) erasers - enzymes capable of removing marks introduced by the writers - such as histone deacetylases (HDACs) and histone demethylases (HDMs), and (c) readers - proteins with specialized domains capable of recognizing these changes - such as the bromodomain and external terminal protein (BET) family.[4] In addition to these core epigenetic targets, a wide range of proteins also play important roles in epigenetic regulation; these proteins include histone chaperones[5] (critical for nucleosome assembly), chromatin remodelers[6,7] (CHRs - responsible for moving, ejecting, and restructuring the nucleosome), and even some classes of transcription factors.[8]

Epigenetics is an essential component in an organism's normal development and responsiveness, so its dysregulation has been associated with altered gene expression patterns related to multiple diseases.[9–12] This makes epigenetic targets a significant focus for drug discovery research. Successful examples can be found in cancer research, with the approval of eight epigenetic drugs (drugs targeting epigenetic proteins) for clinical use: azacytidine and decitabine targeting DNMT1, vorinostat, belinostat, panobinostat, romidepsin and tucsibinostat targeting HDACs, and tazemostat targeting an HMT (EZH2).[3,13] The importance of epigenetics in drug discovery is also illustrated by the increasing availability of chemogenomic databases related to epigenetics over the past decade.[14–18] An example of this is EpiFactors,[16] to the best of our knowledge, the database with the largest number of annotated proteins related to epigenetics reported so far, with a total of 815 different targets. In a recent work,[19] we surveyed the status of the compounds tested against these and other epigenetic targets identified from ChEMBL,[20] Therapeutic Target Database,[21] and scientific literature. We found out that for 136 of these targets, there are more than ten reported inhibitors, which meant a considerable increase in comparison with the 52 targets fulfilling the same criteria in 2017.[18] The rich structure-activity relationships (SAR) contained in these large data sets represents an excellent source of information to develop predictive models that have not been developed thus far on a large-scale basis. In a previous work the authors explored the SAR of epigenetic target data sets using the concept of activity landscape. Although that work was a quantitative study, it was descriptive.[22]

The increase in the publicly available chemogenomic data for all target classes over the years opened up the opportunity for the construction of ligand-based machine learning models to assist target prediction of small molecules. Some of these methods are currently available as easy-to-access web applications, such as Similarity Ensemble Approach[23] (SEA), HitPick,[24,25] Polypharmacology Browser[26,27] (PPB), TargetHunter,[28] and SwissTargetPrediction,[29,30] to name a few examples. These methods usually assign the targets for a given small molecule from the known targets of the most similar ligands in their datasets, employing different descriptions and metrics for the similarity assessment, and often making use of additional statistical models to estimate the significance of the predictions.[23,25,27] Despite of the increasing number of chemogenomic databases related to epigenetics, this data still represents a minimal amount when compared to other protein families such as kinases (KINs), ion channels or G protein-coupled receptors.[31,32] This suggests that epigenetic targets are commonly underrepresented in the current target prediction methods, and that unless the similarity of a known ligand is high enough, they are less likely to be predicted as potential targets of small molecules, which points out the need of developing predictive models focused on epigenetic targets to assist medicinal chemistry efforts in this area.

Machine learning methods have proven to be useful in multiple areas of drug discovery,[33–35] one such being target prediction of small molecules.[25,27,30] For instance, in a retrospective large-scale comparison of machine learning methods for target prediction on ChEMBL (in the context of biochemical assays),[36] deep neural networks were the best performing method for this task when trained on Extended Connectivity Fingerprints[37] (ECFP) of chemical compounds.[38] However, the application of machine learning models for large-scale epigenetic target prediction has been explored on a limited basis, with most works focused on single targets[39,40] or protein families such as HDACs[41] or the BET family.[42]

Herein, we aimed to develop accurate models for epigenetic target prediction based on state-of-the-art machine learning algorithms trained on different fingerprint representation of compounds. We describe the development of predictive models with high precision for 55 epigenetic targets. Derivation of such predictive models is relevant for medicinal chemistry to develop hypothesis for the discovery of new epigenetic probes and drugs. The best models herein generated are implemented in an easy-to-use web application freely available to support medicinal chemistry projects related to epigenetic drug and probe discovery. It is anticipated that this tool will assist epigenetic drug design and development projects in the design and selection of compounds with potential epigenetic activity.

RESULTS

This section is organized into three major parts. First, we described the results of the data sets of epigenetic targets used in this work. The second part, entitled “Epigenetic Target Prediction with Machine Learning,” presents the results of the development of the machine learning models and their validation using two main strategies. The third main section, “Retrospective Identification of Epigenetic Targets,” shows, as a case study, a practical application of the best machine learning model derived in the second part, to identify epigenetic targets for external and recently reported compounds. All the details of the methods used are described in the Experimental Section.

Chemogenomic Data for Epigenetic Targets. Quantitative compound-protein associations were extracted from ChEMBL 27[20] and PubChem[43] to build epigenetic target-associated compound datasets meeting the following criteria: (a) containing at least 30 compounds with a quantitative measure of biological activity (IC_{50}, EC_{50}, K_i or K_d) lower or equal to 10 μM (“active”) and at least 30 compounds with a quantitative measure of biological activity higher than 10 μM (“inactive”), and (b) modelability index (MODI)[44] higher than 0.7 for at least one of the three molecular fingerprints selected as compound representation (see Experimental Section for further details). As illustrated in Figure 1, a total of 55 epigenetic targets were included and distributed as follows: (a) 26 writers, including 16 KINs, six HMTs, three HATs, and DNMT1, (b) 21 erasers, consisting of 12 HDACs, six HDMs, two proteins with dual activity (HDAC/HDM) and one protein related to histone ubiquitination (USP7), (c) four readers, including three bromodomain (BRD) containing proteins and one histone methyl-lysine binding protein (L3MBTL1), and (d) other proteins, consisting of three CHRs and one cofactor involved in DNA demethylation (APEX1). Details on the 55 epigenetic targets and their corresponding target-associated compound datasets are included as Table S1 in the Supporting Information.

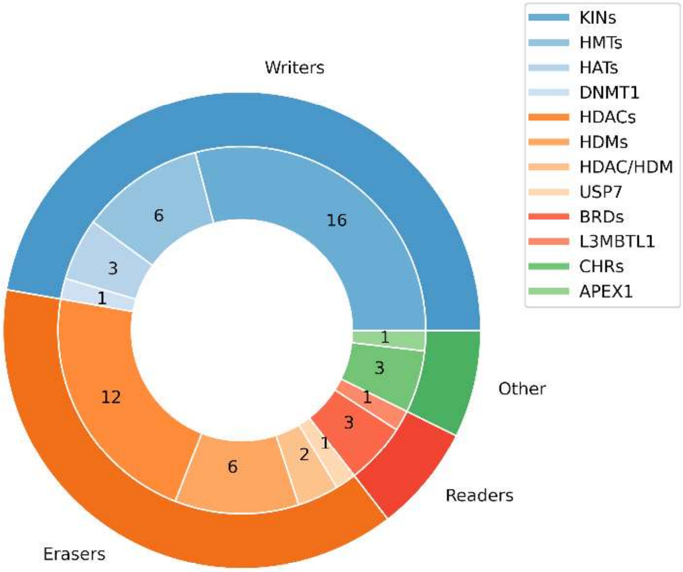


Figure 1. Distribution of Epigenetic Targets included in this work.

The compiled chemogenomic dataset contained 26318 unique compounds and 38129 compound-protein associations, with 28750 of them being labeled as active and 9379 labeled as inactive (due to the natural, although not the best practice of reporting mostly active compounds and not negative -inactive- data in ChEMBL). Consistently with the compound/compound-protein associations ratio, 20318 compounds (77.2%) in the dataset had known associations to a single target, and only 196 compounds (0.7%) had known associations to at least 10 targets, with a maximum of 15 targets for four compounds (Table 1).

Table 1. Distribution of known associations per compound.

Number of known associations	Number of compounds
1	20318
2	3853
3	1004
4	531
5	122
6	127
7	83
8	22
9	62
10	31
11	88
12	15
13	48
14	10
15	4
Total	26318

Epigenetic target-associated compound datasets consisted of 693 compounds on average, with a minimum of 73 for an HDM (KDM4E) and a maximum of 4901 for a KIN (JAK2). In agreement with the class imbalance in the entire dataset, all 55 compound datasets had different class imbalance levels, showing an average proportion of active compounds of 59.3%, with a minimum of 23.2% for a CHR (TOP2A) and a maximum of 92.4% for JAK2 (Figure 2).

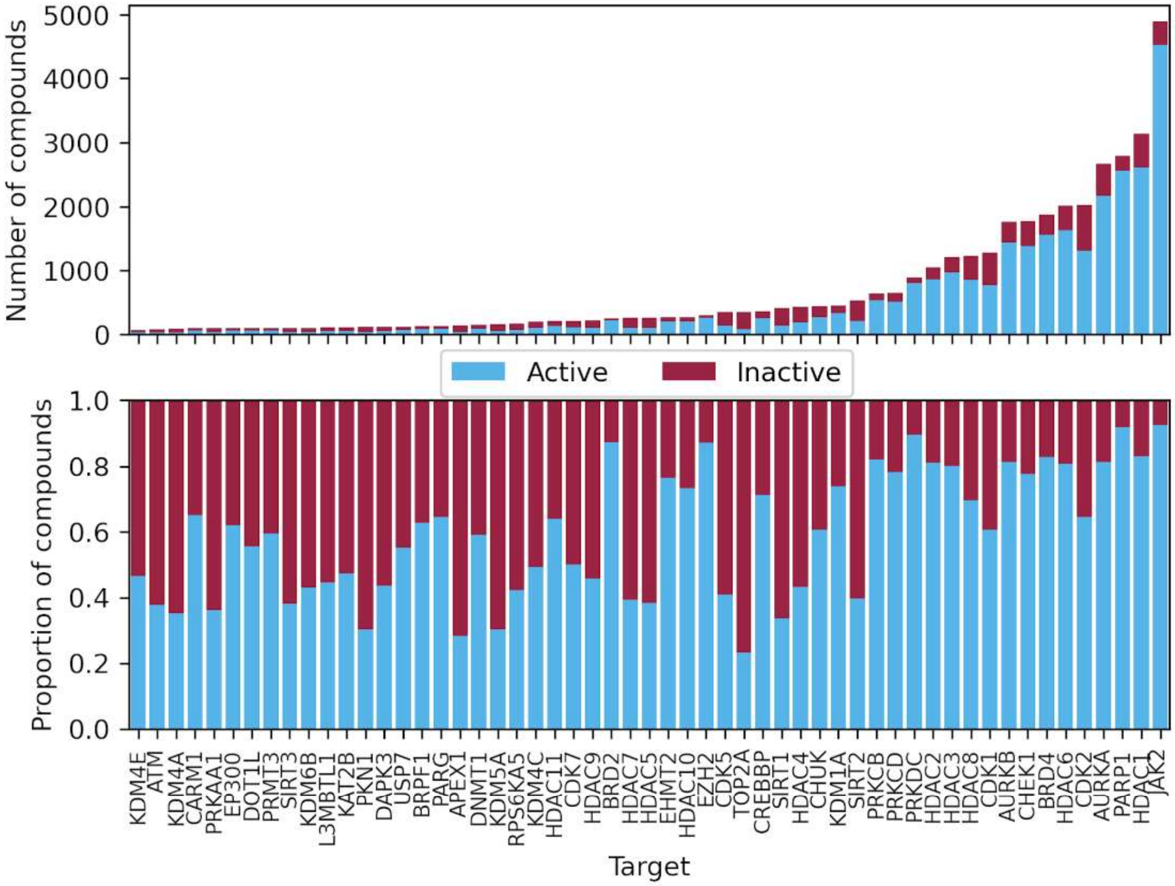


Figure 2. Size and composition of target-associated compound datasets.

Epigenetic Target Prediction with Machine Learning. Predictive models for epigenetic target prediction were built using two validation strategies summarized in Figure 3. The first strategy (Single Target Validation) involved the performance comparison of 15 different models on a stratified 10-fold cross-validation basis in the context of 55 single-target binary classification tasks. The two best performing models were combined to generate a consensus model, and the performance of these three models was assessed on a distance-to-model (DM) basis. The second strategy (Multi-Target Validation) focused on the global performance comparison of the best models identified in the first strategy when evaluated on 10 compound samples with the same number of known active associations for each epigenetic target. The results of each strategy are described in the next two sections.

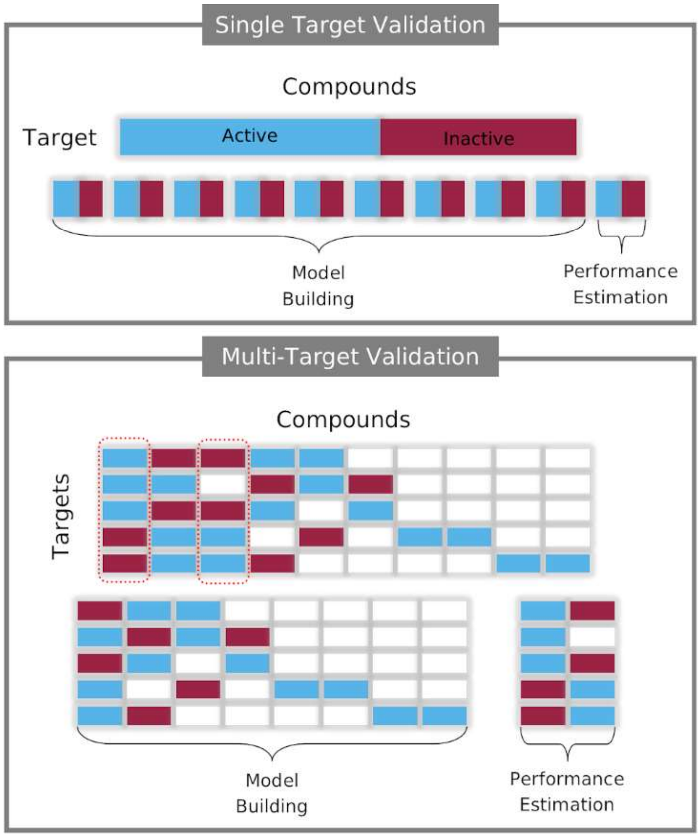


Figure 3. Two validation strategies employed for Epigenetic Target Prediction.

Single Target Validation. Fifteen different binary classification models with optimized hyperparameters were built for each of the 55 target-associated compound datasets. Models were derived from the combination between five state-of-the-art machine learning algorithms: *k*-nearest neighbors (*k*-NN)[45], Random Forest (RF)[46], Gradient Boosting Trees (GBT)[47], Support Vector Machines (SVM)[48], and Feed-Forward Neural Networks (FFNN)[49], and three molecular fingerprints of different design used as compound representations: Molecular ACCess System (MACCS) Keys (166-bit),[50] Morgan fingerprint with radius 2 (2048-bit),[37] and RDK fingerprint (2048-bit). Each model is denoted as a combination of fingerprint and algorithm (fingerprint::algorithm). For each algorithm and target, hyperparameters were optimized from an exhaustive search detailed in the Experimental Section, using the mean balanced accuracy (BA) over a 10-fold cross-validation as the performance metric to select the best set of hyperparameters.

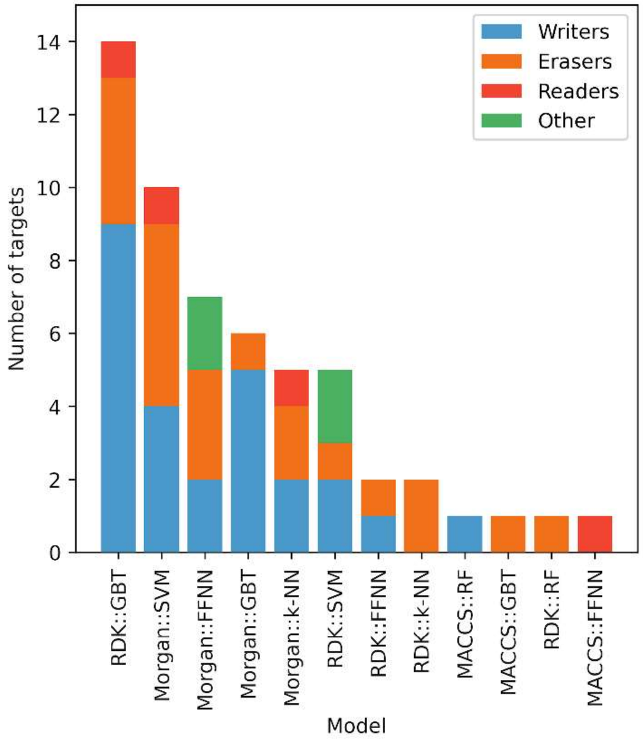


Figure 4. Distribution of best performing model per target class, considering balance accuracy as the evaluation metric.

Figure 4 shows the number of targets for which each model was identified as the best performing, considering the mean BA over the ten folds as a point metric. Under this approach, there is no model, fingerprint, nor machine learning algorithm that could be identified as the best performing for all 55 target datasets considered in this work. Figure 4 shows that RDK::GBT had the highest mean BA for 14 out of the 55 targets, making them the most frequent choice. However, in terms of compound representations only, Morgan fingerprint was the best choice for 28 targets, followed by 24 for RDK fingerprint and three for MACCS. Nevertheless, t-tests comparing the sets of BA scores calculated from the ten validation folds revealed that for all the targets, there is at least another model with no significant difference of performance to the one with the highest mean BA (Table S2 in the Supporting Information). Moreover, the t-test comparison revealed that for 35 out of the 55 targets, there are at least 9 other models with no significant difference of performance to the one with the highest BA (at 95% confidence level), a surprising quantity considering the number of algorithms and compound representations included.

To compare the models herein generated in a more global context, the cross-validated predictions for each optimized model were stored and used to compute single point performance metrics in the context of each target, being BA, F1 score, and Mathews correlation coefficient (MMC). Summary results of the fifteen models' performance are summarized in Table 2, and their distribution across the 55 epigenetic targets is shown in Figure 5.

Table 2. Single Target Validation performance.

Model	**BA**	**F1**	**MCC**
Consensus	0.835 ± 0.067	0.851 ± 0.110	0.676 ± 0.123
Morgan::SVM	0.830 ± 0.065	0.862 ± 0.101	0.680 ± 0.123
RDK::SVM	0.827 ± 0.061	0.862 ± 0.096	0.670 ± 0.116
RDK::GBT	0.824 ± 0.067	0.859 ± 0.107	0.669 ± 0.123
RDK::FFNN	0.822 ± 0.057	0.859 ± 0.092	0.659 ± 0.108
Morgan::FFNN	0.819 ± 0.067	0.856 ± 0.100	0.651 ± 0.132
Morgan::*k*-NN	0.817 ± 0.068	0.859 ± 0.102	0.655 ± 0.134
RDK::RF	0.816 ± 0.067	0.856 ± 0.111	0.666 ± 0.115
Morgan::GBT	0.815 ± 0.073	0.855 ± 0.112	0.659 ± 0.136
RDK::*k*-NN	0.814 ± 0.063	0.855 ± 0.095	0.641 ± 0.124
Morgan::RF	0.811 ± 0.075	0.855 ± 0.118	0.663 ± 0.131
MACCS::SVM	0.807 ± 0.073	0.847 ± 0.118	0.632 ± 0.145
MACCS::GBT	0.806 ± 0.074	0.845 ± 0.117	0.629 ± 0.142
MACCS::RF	0.800 ± 0.072	0.846 ± 0.114	0.626 ± 0.134
MACCS::*k*-NN	0.791 ± 0.066	0.839 ± 0.109	0.600 ± 0.132
MACCS::FFNN	0.785 ± 0.069	0.829 ± 0.115	0.580 ± 0.137

Mean and standard deviation (mean ± SD) of BA, F1 and MCC for 55 single target binary classifiers built on 15 fingerprint::algorithm combinations and a consensus model. Results are sorted by decreasing BA.

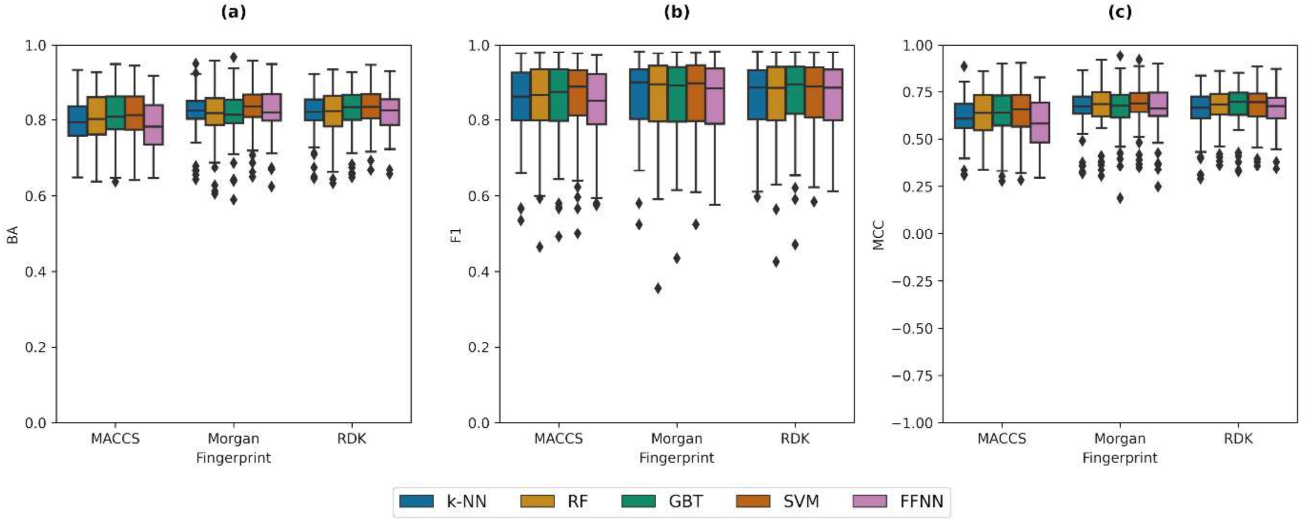


Figure 5. Performance comparison of single target binary classifiers. (a) balanced accuracy (BA), (b) F1 score, (c) Mathews correlation coefficient (MCC). Each boxplot contains the performance metrics for 55 different target-associated compound datasets.

Overall, most of the models performed well in the single-target prediction task, having a mean BA and F1 score higher than 0.5 and mean MCC higher than zero. To identify the global best performing model, we applied Wilcoxon signed-rank tests between all pairs of models for the three metrics of performance. Each test involves a comparison between sets of 55 values. The Morgan::SVM model showed the highest mean values for the three performance metrics and significantly higher values of BA and MCC when compared to all but the RDK::SVM model (at 95% confidence level). F1 score showed the lower differences between models, with the Morgan::SVM having the highest mean value and significantly higher values when compared to all but five models, being RDK::SVM, RDK::GBT, RDK::FNN, RDK::RF and Morgan::*k*-NN (at 95% confidence level). These results suggested Morgan and RDK fingerprints and the SVM algorithm as the best combinations to derive binary classifiers for the current sets of studied epigenetic targets.

Consensus Model. It has been pointed out that the combination of predictive models generally has a higher reliability than the individual models.[51,52] In other to identify the best models combination to construct a consensus model, we performed a hierarchical clustering of the models relying on Morgan and RDK fingerprints by comparing their 38129 cross-validated predictions obtained in the single target validation strategy (vide supra). Jaccard distance was employed as the metric between models and an average linkage was used for the hierarchical clustering calculation as detailed in the Experimental Section. Figure 6 depicts a dendrogram of the hierarchical clustering. Predictions for all models are closely related, with all average distances between groups being lower than 0.1. It should be noted that models relying on the same fingerprint are clustered together before being grouped with models built on a different fingerprint. In the context of each fingerprint, the clustering follows the same order, with models relying on GBT and RF being grouped at first, followed by those built on SVM, FFNN, and *k*-NN.

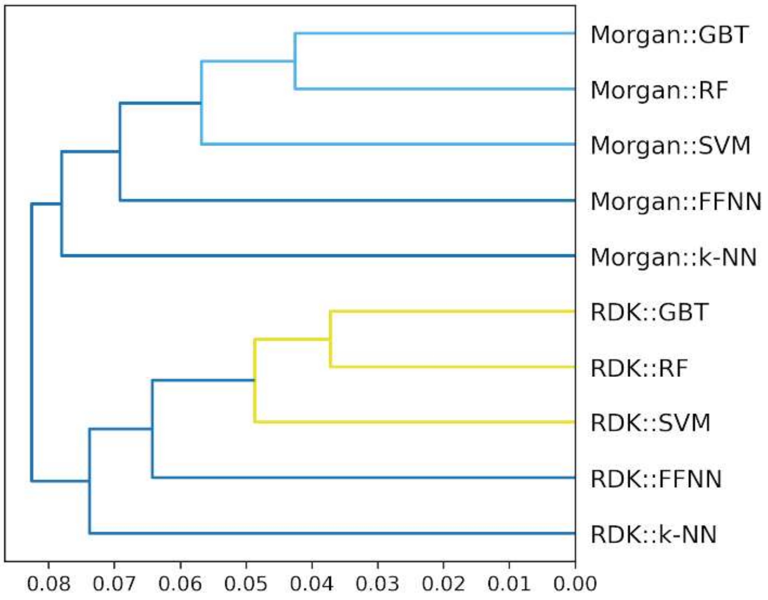


Figure 6. Hierarchical clustering of Morgan and RDK models. Average linkage and Jaccard distance between the models' predictions were used for the calculation.

Based on these findings, the best performing model built on each fingerprint, Morgan::SVM and RDK::SVM, were combined to derive a consensus model. To prioritize the correct identification of active compounds, the consensus model was constructed by combining the predictions of both models so that a compound was predicted as "active" for a given target only if both models agreed in the prediction and “inactive” otherwise. This consensus model showed a mean BA, F1 score, and MCC of 0.835, 0.851, and 0.676, respectively. Wilcoxon signed-rank tests indicated significantly lower values for F1 score than those obtained by the individual models, and no significant difference for BA and MCC values (at 95% confidence level). Since F1 score is defined as the harmonic mean of precision (PPV) and recall (TPR) for the active class, and the consensus model was *a priori* built to have high precision, the significantly lower values obtained for F1 score are explained by a decrease in the TPR of the model (Table S3 and Figure S1 in the Supporting Information), which is related to the decrease in the number of “active” outcomes for the consensus model.

Distance-to-Model. Although BA, F1 score, and MCC are well-suited metrics for model performance estimation on imbalanced datasets, in a practical medicinal chemistry application, the correct identification of active compounds is often more important than the correct identification of inactive ones. To this end, the performance of the individual models and the derived consensus model were studied in terms of PPV, TPR, negative predictive value (NPV), and true negative rate (TNR). To estimate the models’ applicability domain, these metrics were computed on a distance-to-model (DM) basis as detailed in the Experimental Section. All cross-validated predictions were categorized into four quartiles (Q1-Q4) according to their mean Jaccard distances to the training set in the context of each target (Table S4 in the Supporting Information). Summary results of the three models’ performance are presented in Table 3, and their distribution across the 55 epigenetic targets is shown in Figure 7.

Table 3. Single Target Performance (Strategy I) in a distance-to-model basis.

Model	Quartile	PPV	TPR	NPV	TNR
Consensus	Q1	0.923 ± 0.081	0.894 ± 0.121	0.762 ± 0.197	0.777 ± 0.195
	Q2	0.914 ± 0.121	0.872 ± 0.143	0.790 ± 0.184	0.803 ± 0.197
	Q3	0.883 ± 0.114	0.826 ± 0.149	0.805 ± 0.114	0.855 ± 0.134
	Q4	0.810 ± 0.153	0.653 ± 0.242	0.764 ± 0.141	0.869 ± 0.152
Morgan::SVM	Q1	0.912 ± 0.086	0.915 ± 0.113	0.831 ± 0.175	0.741 ± 0.229
	Q2	0.893 ± 0.128	0.897 ± 0.131	0.827 ± 0.161	0.753 ± 0.222
	Q3	0.847 ± 0.141	0.864 ± 0.132	0.834 ± 0.099	0.800 ± 0.149
	Q4	0.781 ± 0.147	0.714 ± 0.226	0.791 ± 0.148	0.820 ± 0.176
RDK::SVM	Q1	0.891 ± 0.131	0.914 ± 0.123	0.792 ± 0.186	0.739 ± 0.203
	Q2	0.878 ± 0.137	0.901 ± 0.127	0.811 ± 0.189	0.721 ± 0.238
	Q3	0.843 ± 0.133	0.866 ± 0.134	0.840 ± 0.111	0.793 ± 0.181
	Q4	0.780 ± 0.137	0.730 ± 0.217	0.825 ± 0.095	0.822 ± 0.146

Mean and standard deviation (mean ± SD) of PPV, TPR, NPV and TNR for 55 single target binary classifiers built on two fingerprint::algorithm combinations and a consensus model.

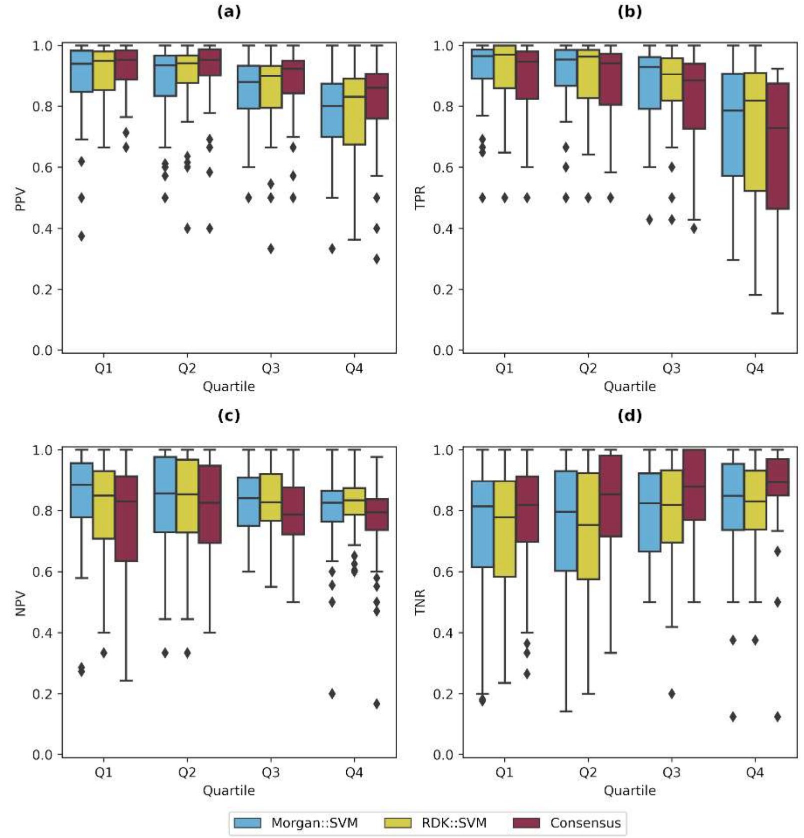


Figure 7. Performance comparison of single target binary classifiers in a distance-to-model basis. (a) positive predictive value (PPV), (b) true positive rate (TPR), (c) negative predictive value (NPV), (d) true negative rate (TNR). Each boxplot contains the performance metrics for up to 55 different target-associated compound datasets.

All performance metrics showed similar trends for the three models. As shown in Figure 7, PPV and TPR decreased as the distance from a compound to the training set increased, while, in the same scenario, NPV and TNR generally decreased. This suggests that predictions, particularly those for active compounds, are more reliable when the predicted compound is closer to the compounds in the training set. Wilcoxon signed-rank tests indicated significantly higher PPV and TNR values, and lower NPV and TPR values for all quartiles when comparing the consensus model to any of the two individual models (at 95% confidence level). These results agree with the lower probability of the consensus model of having an "active" outcome compared to the individual models (since both individual models must agree with the prediction). The lower number of compounds predicted active is associated with the lower recovery of the known active compounds (low TPR) and low precision in predicting inactive compounds (low NPV) compared to the individual models. However, this also implies that the known inactive compounds are well differentiated by the model (high TNR) and the precision in the prediction of active compounds is higher for the consensus model (high PPV), which is desirable in a typical medicinal chemistry project. It should be noted that despite the decrease in the PPV at high DM for the consensus model, the mean values of PPV for all quartiles were higher than 0.8, with a maximum 0.923 at Q1 and a minimum of 0.810 for Q4, suggesting high reliability on the predictions of active compounds, even when the predicted compounds are far from the compounds in the training set (Figure 7). Moreover, regardless of the performance difference in TPR and NPT between the consensus models and the individual models, these performance metrics for the consensus model are still high, showing mean values higher than 0.6 for all quartiles, where the lower mean values were 0.653 and 0.764 for TPR and NPT in Q4, respectively.

Multi-Target Validation. All results in the previous sections were analyzed in the cross-validated predictions of 55 individual binary classifiers. However, given that each classifier was trained and tested on compound datasets of different sizes, assessing the performance of the combination of these 55 predictive models in the epigenetic target prediction task would lead to an incorrect performance estimation, with a bias over the targets with the most populated compound datasets associated. For this reason, 10 compound samples containing exactly six known active compounds for each target were assembled. For each sample, Morgan::SVM, RDK::SVM, and the consensus model were re-trained on the whole compounds datasets, excluding the compounds in the sample and evaluated on the sample initially excluded (as an external set). In this case, only metrics considering the correct identification of active compounds were calculated (PPV and TPR) on a DM basis following the same approach described in the previous section, considering only the predictions with a truly known label. Samples contained between 184 and 229 compounds (210 on average), and no more than 40 repeated compounds among them (Figure S2 in the Supporting Information). Summary results of the three models' performance are presented in Table 4, and their distribution across the 10 samples is shown in Figure 8.

Table 4. Multi-Target Performance (Strategy II) in a distance-to-model basis.

Model	Quartile	PPV	TPR
Consensus	Q1	0.952 ± 0.022	0.879 ± 0.033
	Q2	0.924 ± 0.036	0.833 ± 0.051
	Q3	0.822 ± 0.062	0.719 ± 0.065
	Q4	0.773 ± 0.056	0.558 ± 0.073
Morgan::SVM	Q1	0.948 ± 0.022	0.901 ± 0.027
	Q2	0.912 ± 0.030	0.871 ± 0.054
	Q3	0.744 ± 0.073	0.751 ± 0.058
	Q4	0.688 ± 0.063	0.624 ± 0.060
RDK::SVM	Q1	0.947 ± 0.019	0.918 ± 0.029
	Q2	0.899 ± 0.050	0.862 ± 0.059
	Q3	0.759 ± 0.086	0.772 ± 0.060
	Q4	0.707 ± 0.054	0.624 ± 0.056

Mean and standard deviation (mean ± SD) of PPV and TPR for 10 combinations of 55 single target binary classifiers built on two fingerprint::algorithm combinations and a consensus model.

Under this validation strategy, PPV and TPR showed the same trends as in the single target validation: both decreased as the DM increased for all models. Wilcoxon signed-rank tests indicated significantly lower TPR values and higher PPV values for all quartiles when comparing the consensus model to any of the two individual models (at 95% confidence level) (Table 4 and Figure 8).

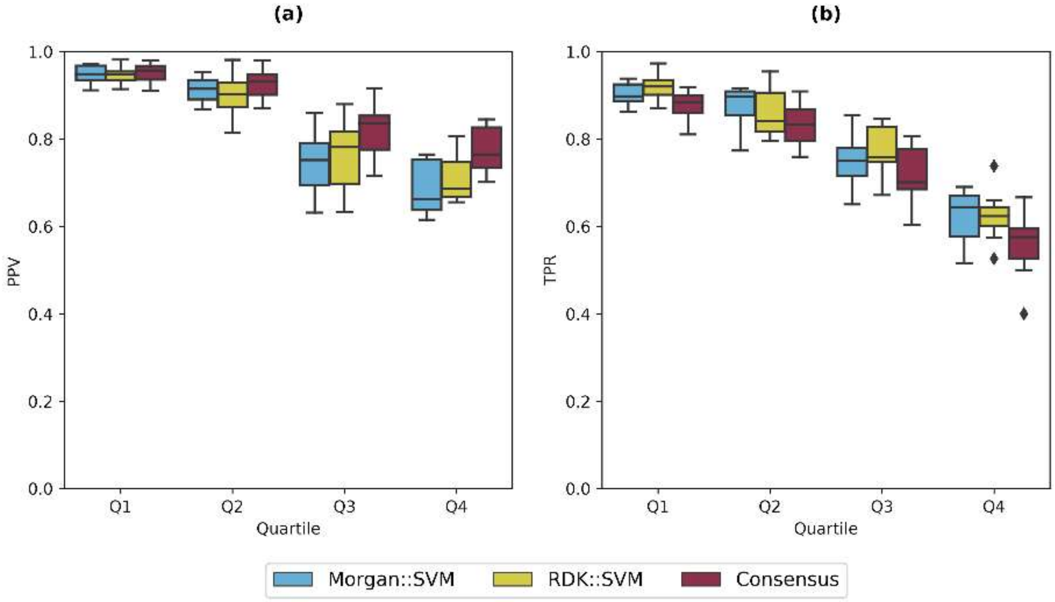


Figure 8. Performance comparison of the combination of 55 single target binary classifiers in a distance-to-model basis. (a) positive predictive value (PPV), (b) true positive rate (TPR). Each boxplot contains the performance metrics for up to 10 different combinations.

Retrospective Identification of Epigenetic Targets. As a proof of concept on the practical applicability of the herein developed consensus model, we employed it in the retrospective identification of the epigenetic targets for two external and recently reported compounds (Figure 9): (1) compound **17**, an inhibitor of EP300 and CREBBP, as targets representative of the less populated compound datasets, and (2) compound **43a**, an inhibitor of HDACs 1, 3, 6, 8 and BRD4, representing targets with the most populated compound datasets. These results were compared to those obtained by four general target prediction tools freely available online when performing this study: HitPickV2, PPB2, SEA, and SwissTargetPrediction. Figure 9 summarizes the results obtained by the consensus model and the four target prediction tools. The full list of predictions is available as Tables S5-S15 in the Supporting Information. The number of targets predicted by each of the web tools is fixed, being 10 for HitPickV2, 20 for PPB2 and SEA, and 100 for SwissTargetPrediction, while the herein reported consensus model was re-fitted using the entire datasets and set up to perform the predictions for the 55 epigenetic targets, with only those involving an “active” outcome considered as the predicted targets.

For compound **17**, our consensus model was the only one able to correctly identify CREBBP and EP300 as its targets (from 12 predicted targets). For compound **43a**, our model (from 18 predicted targets) and SEA identified correctly its five known targets. HitPickV2 and PPB2 predicted correctly the four HDACs but not BRD4, while SwissTargetPrediction predicted correctly only HDACs 1, 6 and 8. Although a more exhaustive external validation is needed, these results suggests that epigenetic targets with large amounts of chemogenomic data associated (such as HDACs) are generally well represented in current target prediction tools, while those with fewer data are not well covered. Moreover, it should be noted that the known epigenetic targets were not always among the top predictions for the available tools. For instance, HDACs 3, 6 and 8 were ranked 12, 13 and 15 by SEA, and HDACs 1, 6 and 8 from SwissTargetPrediction were ranked in positions 42, 43 and 44, so in a practical application, these targets would be hardly prioritized. Although the experimental validation of the predictions from all models would be needed to provide better means of comparison, these findings reinforce the potential usefulness of a tool focused on epigenetic targets for medicinal chemistry applications in drug discovery.

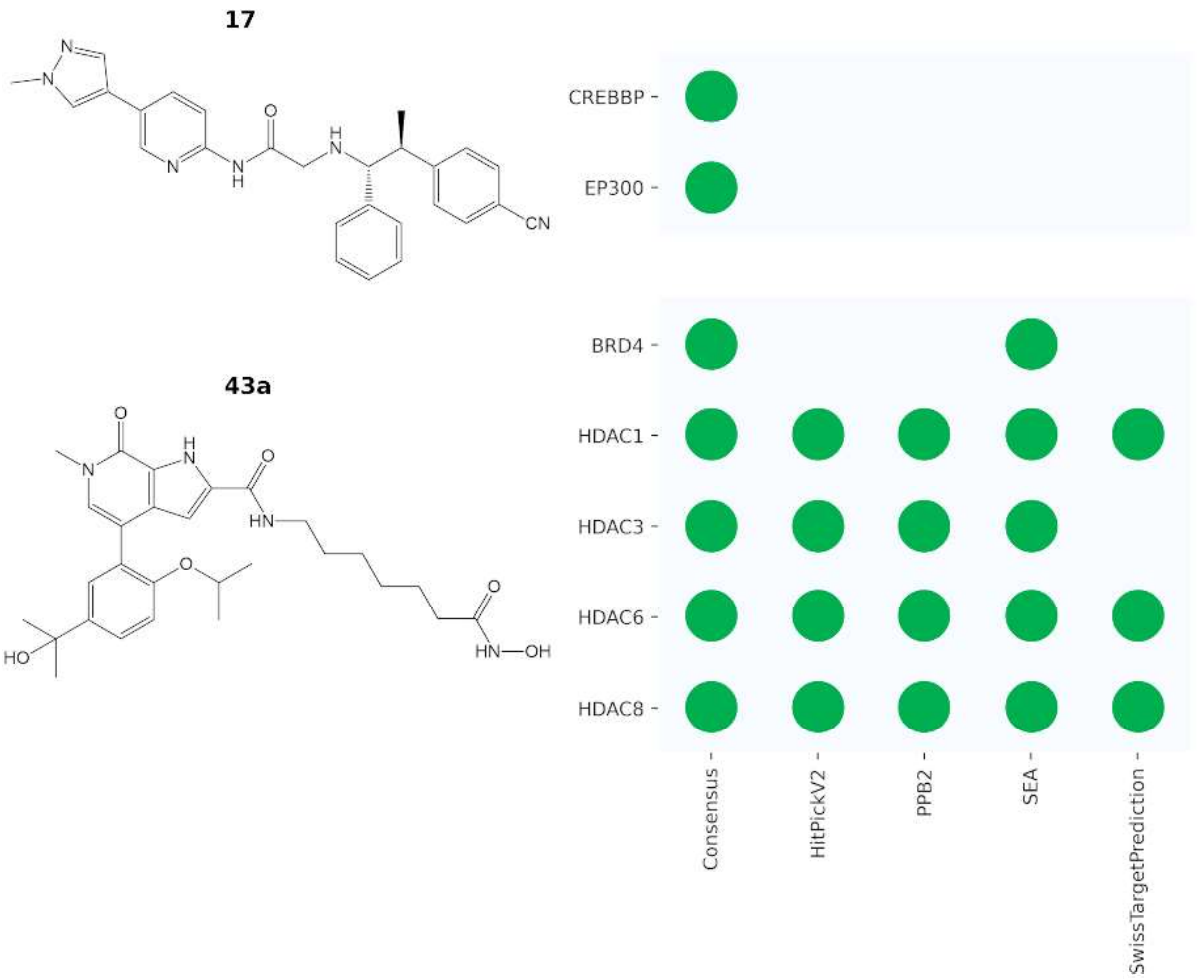


Figure 9. Comparison of target prediction tools for the retrospective identification of epigenetic targets.

Availability and Implementation. All row data to reproduce the results presented in this work is available free of charge at figshare repository (10.6084/m9.figshare.13519580). To encourage the medicinal chemistry community to apply the predictive consensus model developed in this work, the model was re-fitted using the entire datasets and has been implemented as a freely accessible and easy-to-use web application described in a separate work and available at http://www.epigenetictargetprofiler.com/.

DISCUSSION AND CONCLUSIONS

Epigenetic drug discovery is increasingly important across different therapeutic areas. Despite the large amount of SAR data stored in public data sets, that information has not been used on a large scale to develop predictive models that support the medicinal chemistry community's efforts working on these cutting-edge targets. To fill this gap, in this study, we developed and evaluated the performance of five state-of-the-art machine learning algorithms built on three molecular fingerprints of different designs to predict 55 epigenetic targets of small molecules. To the best of our knowledge, this is the first study covering epigenetic targets on a large-scale basis. The performance of the herein reported models was validated using two different approaches, involving their performance estimation for binary classifications in 10-fold cross-validations in the context of each target, as well as the performance of their combination in the epigenetic target prediction task evaluated over 10 balanced samples of compounds containing an equal number of known active compounds for each target.

Although none of the herein reported models was identified as the best performing one for all the 55 targets, our results suggested Morgan and RDK fingerprints as the best representations for the derivation of binary classifiers for the studied targets, particularly when derived using SVM, where no significative difference was found for their performance. This cannot be generalized for other, or even for these targets, since it could be associated with the hyperparameter space employed to optimize the models. Moreover, a model's performance is also dependent on the dataset composition, so the trends herein presented could change as more bioactivity data is published and different sets of hyperparameters are studied.

A consensus model was built by combining the predictions of the best models derived from Morgan and RDK fingerprints (Morgan::SVM and RDK::SVM), also supported on the fact that predictions between models relying on the same fingerprint are more closely related than those relying on different representations as demonstrated by the hierarchical clustering analysis of their cross-validated predictions. The consensus models' performance and the two source models were analyzed on a DM basis, categorizing the predictions according to the Jaccard distance of the compounds in the test set to those in the training set. For the single target binary classification, the consensus model showed a significantly higher precision for identifying active compounds than those obtained by the individual models regardless of the DM. This trend was preserved when the models were evaluated to predict epigenetic targets.

The consensus model showed a mean BA of 0.835 considering the cross-validated predictions of the 55 target-associated binary classifiers, with mean precisions for identifying active compounds ranging from 0.923 for those compounds closer to the training set, to 0.810 for those farther from the training set. For the epigenetic target prediction task, mean precisions ranged from 0.952 to 0.773 under the same scheme.

We showed the consensus model's practical applicability by the retrospective identification of the epigenetic targets of two external and recently reported compounds. These results showed the consensus model as a robust and accurate method for epigenetic target prediction of small molecules, which led us to implement it as an easy-to-use web application available for free. It is hoped that this model will be helpful in practical medicinal chemistry applications for epigenetic drug discovery.

EXPERIMENTAL SECTION

Data Sets. Our primary source of SAR data was ChEMBL 27,[20] we collected all the quantitative compound-protein associations from single protein assays, related to the 136 epigenetic targets identified in our previous work[19] (biological activity reported as IC_{50}, EC_{50}, K_i or K_d). In the context of each target, compounds were labeled as "active" when they had unequivocally assigned activities lower than or equal to 10 µM, and as "inactive" in the opposite case. Compounds whose label could not be unequivocally assigned (e.g., activity <

100 µM or activity > 1 µM) were removed from the data set. The remaining compounds were curated using the open-source cheminformatics toolkit RDKit, version 2020.03.1 and the functions Standardizer, LargestFragmentChoser, Uncharger, Reionizer and TautomerCanonicalizer implemented in the molecule validation and standardization tool MolVS, as described in previous works.[54,55] In short, the Simplified Molecular Input Line Entry System[56] (SMILES) of each compound was standardized, those compounds consisting of multiple components were split and the largest component was retained. Compounds containing 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 without preserved stereochemistry. Once all compounds were standardized, those with molecular weight higher than 800 Da as well as duplicated compounds with contradictory labels were removed. We preserved compound-protein associations only for those targets with at least 30 compounds labeled as “active,” corresponding to 72 different targets. Since chemogenomic data for these epigenetic targets include a higher proportion of associations for “active” compounds (64% on average), we extended our initial data with “inactive” compounds from PubChem.[43] We included only compounds with annotated quantitative data (IC_{50}), all these compounds were curated using the same procedure described above and added only if they were not already included in the datasets. Finally, we kept 58 target-associated datasets containing at least 30 compounds labeled as “inactive.”

Molecular Representations. To develop the machine learning models, we selected three molecular fingerprints of different design: (a) Molecular ACCess System (MACCS) Keys (166-bit)[50] as a dictionary based fingerprint where each position indicates presence or absence of a predefined structure, (b) Morgan fingerprint with radius 2 (2048-bit)[37] as a circular fingerprint where each position represents an atom environment including all atoms connected up to a radius of 2 bonds, and (c) RDK fingerprint (2048-bit) as a topological fingerprint where each position represents a linear substructure including all atoms connected up to a length of 7 bonds. All fingerprints were generated using the open-source cheminformatics toolkit RDKit, version 2020.03.1 for Python.

Data Modelability. To *a priori* estimate the feasibility to obtain predictive binary classification models for each target, we calculated the modelability index (MODI)[44] for each target-associated dataset. MODI is defined as the proportion of compounds in a dataset for which its nearest neighbor belongs to the same class in a given feature space. For its calculation, we selected as compound representation the three different fingerprints described above and as metric to identify the nearest neighbors the Jaccard distance, defined as:

$$J(A,B) = 1 - \frac{c}{a+b-c}$$

where J (A, B) is the Jaccard distance between compounds A and B in a given fingerprint representation, with a and b being the number of “on” bits for compound A and B, respectively, and c being the number of “on” bits for both compounds. Further modeling was performed only for 55 datasets with a MODI higher or equal than 0.7 for at least one molecular representation.

Machine Learning Methods. Binary classification models for each target were generated using five different machine learning algorithms: *k*-nearest neighbors(*k*-NN)[45], Random Forest (RF)[46], Gradient Boosting Trees(GBT)[47], Support Vector Machines(SVM)[48], and Feed-Forward Neural Networks (FFNN)[49]. All machine learning methods were implemented using the Scikit-learn Python library (0.22.1).[57] For model building, training instances were represented by a feature vector (fingerprint) and associated to a class label (“active” / “inactive”). To avoid hyperparameter bias when comparing different models, the hyperparameters for each model were optimized using stratified 10-fold cross-validation in an exhaustive search over a limited hyperparameter space. To keep the search space small, only selected hyperparameters on each algorithm were optimized. Hereunder, we provide brief explanations on each algorithm and the hyperparameters considered for its optimization; all hyperparameters not explicitly indicated in the text were set as default.

In *k*-NN classification, the predicted label of a sample is assigned according to the most common label among its *k* nearest neighbors in the training dataset for a given feature space. For this algorithm, we selected the Jaccard distance as the metric to identify the nearest neighbors using a brute-force search. The optimal number of nearest neighbors was optimized using candidate values of 1, 3, 5, 7, and 9.

RF is one of the so-called ensemble methods relying on decision trees. In RF classification, a fixed number of decision tree classifiers are fitted on various bootstrapped subsamples of the training dataset. For a given sample, each decision tree predicts a label, and the final prediction of the sample is the label predicted by most of the trees. For this algorithm, the number of decision trees was fixed to 1000 and the number of features to consider when searching for the best splits in the individual trees was optimized in a representation-dependent manner using candidate values of 1, 2, 3, 4 and 5 times the square root of the number of features in the fingerprint representation.

GBT is another ensemble method relying on decision trees. In this case, the decision tree classifiers are fitted in stages for the whole training dataset, where each subsequent tree is intended to “correct” the errors made by the previous one in terms of a loss function, usually the deviance of the fitted model with respect to a perfect model. For this algorithm, the number of decision trees was fixed to 1000, the number of features to consider when looking for the best splits in the individual trees was optimized in a representation-dependent manner using candidate values of 1, 2, 3, 4 and 5 times the square root of the number of features in the fingerprint representation, for the maximum depth of the individual trees we used candidate values of 4, 6, 8 and 10, and for the minimum number of samples to split an internal node in the individual trees we used candidate values of 2, 3, 4, and 5.

In SVM classification, the hyper-plane that best separate the two classes in the training dataset is constructed by maximizing the distance between training instances belonging to different classes (margin). As this hyper-plane does not always exist, a limited number of errors is allowed using a “cost” hyperparameter to control the relation between the training errors and the margin size. If linear separation of training classes is not possible in a given feature space, kernel functions are applied to project the data into a higher dimensional space where linear separation is possible. For this algorithm, “cost” was optimized using candidate values of 0.01, 0.1, 1.0, 10.0 and 100.0, and the kernel type to be used was selected from three options being non-kernel (“linear”), radial basis functions (“rbf”), and hyperbolic tangent (“sigmoid”).

A FFNN is composed by different layers of computational neurons: an input layer, one or more hidden layers, and an output layer. Neurons in the input layer are associated to the features describing the data, each neuron in the hidden layer accepts the inputs of all neurons in the input layer and transform them to a weighted sum of the original inputs, then a nonlinear activation function is applied to this weighted sum and the result is passed to the neurons in the output layer, where the prediction is performed. The weights from the network are iteratively adjusted during the training stage on the basis of a cost function to minimize, typically cross entropy. For this algorithm, the solver for weight optimization was set as “lbgfs”, the maximum number of iterations (how many times a training data point is passed to the network) was set to 1000, and the number of hidden layers was fixed to 1. The number of neurons in the hidden layer was optimized in a representation-dependent manner using candidate values of 0.1, 0.2, 0.3, 0.4, 0.5, 0.6 and 0.7 times the number of features in the fingerprint representation, and the activation function was selected from three different options, being logistic sigmoid function (“logistic”), hyperbolic tangent function (“tanh”) and rectified linear unit function (“relu”).

Training and Test Sets. For model building, two different validation strategies were implemented (Figure 3), the first comparing different combinations of fingerprints and machine learning algorithms in single-target binary classification tasks (Single Target Validation), and the second evaluating the best performing models from the first strategy in the epigenetic target prediction task (Multi-Target Validation).

Single Target Validation. Considering that the compound-target bioactivity matrix for the studied targets is sparse, the first strategy involved the construction of target-specific classification models and comparison of their performance across the different combinations of fingerprints and machine learning algorithms. Fifteen different binary classification models were built for each target, resulting from the combinations of the three fingerprints used as molecular representations and the five machine learning algorithms used for model fitting. Hyperparameters for each model were optimized using a stratified 10-fold cross-validation, with balanced accuracy (BA) employed as metric for selection of the best performing set of hyperparameters. The cross-validated predictions of the best model were used for the calculation of different performance metrics and comparison of the models. Each model performance was assessed using three metrics unbiased to the class imbalance in the data, BA, F1 score, and Mathews correlation coefficient (MMC), defined as:

$$BA = \frac{0.5\,TP}{TP + FN} + \frac{0.5\,TN}{TN + FP}$$

$$F1 = 2\,x\,\frac{TP}{2TP + FP + FN}$$

$$MCC = \frac{TP\,x\,TN - FP\,x\,FN}{\sqrt{(TP + FP)\,(TP + FN)(TN + FP)(TN + FN)}}$$

where TP means “true positives”, TN “true negatives”, FP “false positives”, and FN “false negatives”, with “positive” and “negative” refereeing to “active” and “inactive” compound labels, respectively.

We built a consensus model by combining the predictions of the best performing models showing the lower relation among their predictions. For that, we performed a hierarchical clustering with average linkage of the models relying on Morgan and RDK fingerprints (the best performing fingerprints), being described by their cross-validated predictions across all targets. As the distance metric for the construction of the hierarchical clustering, we selected de Jaccard distance defined in the Data Modelability section, where in this case J (A, B) represents the distance between two models, with a and b being the number of "active" predictions for model A and B, respectively, and c being the number of "active" predictions for both models.

We compared the consensus model and the single models of which it is composed using precision (positive predictive value - PPV), sensitivity (true positive rate - TPR), negative predictive value (NPV), and specificity (true negative rate - TNR), defined as:

$$PPV = \frac{TP}{TP + FP}$$

$$TPR = \frac{TP}{TP + FN}$$

$$NPV = \frac{TN}{TN + FN}$$

$$TNR = \frac{TN}{TN + FP}$$

In order to estimate the applicability domain of the models, these metrics were computed on a distance-to-model (DM) basis.[58,59] For that, the mean Jaccard distance from each compound in the test sets to all compounds in the training sets was calculated as the DM metric, using the three different fingerprints employed as molecular representation. These average distances were categorized in four quartiles considering all the cross-validated predictions, and all four metrics were calculated for each target and quartile, when predictions on the corresponding quartile were available.

Multi-Target Validation. The consensus model and the corresponding individual models were compared for the epigenetic target prediction problem. To assess the global performance of the combination of the single-target binary classifiers in the epigenetic target prediction task, ten samples of compounds containing the same number of active compounds for each target were assembled. To reduce the target-bias associated to the different sizes on the target-associated compound datasets, each of the compound samples was constructed by iteratively sampling one compound labeled as "active" from the less populated dataset in the sample (or in alphabetical order according to its gene code when there was more than one less populated sample). This process was performed until the sample contained exactly 6 active compounds (20% of the active compounds for the smaller dataset) for each target. If the addition of a compound yields a target containing more than 6 active compounds, the compound was discarded, and if the equal number of active compounds for each target was not satisfied after 1000 iterations, 10% of the sample was randomly discarded, and the iterative sampling continued. These ten samples were used as validation sets so that compounds in the sample were removed from the original target-associated datasets. The single target binary classifiers were refitted using the hyperparameters selected in the Single Target Validation strategy. The performance for the combination of the single-target binary classifiers was assessed by its capability of identifying the known active compounds among the known compound target associations, using PPV and TPR as metrics in the same DM basis described in the Single Target Validation strategy.

Retrospective Identification of Epigenetic Targets. Two external and recently reported compounds with more than one associated epigenetic target were selected to show the practical applicability of the herein reported models in a retrospective identification of its targets: (1) compound **17**,[60] a dual inhibitor of the HATs CREBBP (IC_{50} = 3.2 nM) and EP300 (IC_{50} = 2.5 nM), and (2) compound **43a**,[61] a *pan*-HDAC/BRD inhibitor with reported activities over BRD4 (IC_{50} = 29.5 nM), HDAC1 (IC_{50} = 19.4 nM), HDAC3 (IC_{50} = 36.4 nM), HDAC6 (IC_{50}

= 5.4 nM) and HDAC8 (IC_{50} = 99.6 nM). The consensus model was re-trained on the whole datasets using the hyperparameters identified in the single target validation strategy and 55 predictions were made for each of the external compounds. The epigenetic targets predicted for a compound were those for which the compound was predicted as "active" for the consensus model. These results were compared to those obtained from four currently freely available ligand-based tools for target prediction, being HitPickV2, SEA, PPB2 and SwissTargetPrediction.

ASSOCIATED CONTENT

Supporting Information

Table S1. Target-associated compound datasets included in this work.

Table S2. Models with no significant difference of performance.

Table S3. Single Target Performance.

Table S4. Distance-to-model quartiles.

Figure S1. Performance comparison of single target binary classifiers.

Figure S2. Compounds overlap between samples employed in the Multi-Target Validation.

Table S5. External compounds employed for retrospective target prediction.

Table S6. Targets predicted by the consensus model for compound 17.

Table S7. Targets predicted by HitPickV2 for compound 17.

Table S8. Targets predicted by PPB for compound 17.

Table S9. Targets predicted by SEA for compound 17.

Table S10. Top 45 targets predicted by SwissTargetPrediction for compound 17.

Table S11. Targets predicted by the consensus model for compound 43a.

Table S12. Targets predicted by HitPickV2 for compound 43a.

Table S13. Targets predicted by PPB for compound 43a.

Table S14. Targets predicted by SEA for compound 43a.

Table S15. Top 45 targets predicted by SwissTargetPrediction for compound 43a.

AUTHOR INFORMATION

Corresponding Authors

José L. Medina-Franco; orcid.org/0000-0003-4940-1107; E-mail: medinajl@unam.mx

Norberto Sánchez-Cruz; orcid.org/0000-0003-2707-3966; E-mail: norberto.sc90@gmail.com

ACKNOWLEDGEMENTS

NS-C is thankful to *Consejo Nacional de Ciencia y Tecnología* (CONACyT), Mexico, for the granted scholarship number 335997. We thank *Dirección General de Cómputo y de Tecnologías de Información y Comunicación* (DGTIC), UNAM that provided computational resources to use Miztli supercomputer with the project LANCAD-UNAM-DGTIC-335. We also thank *Consejo Nacional de Ciencia y Tecnología* (CONACyT), Mexico, grant 282785.

ABBREVIATIONS USED

BA, balanced accuracy; BET, bromodomain and external terminal protein; BRD, bromodomain; CHR, chromatin remodeler; Da, Dalton; Da, Dalton; DM, distance-to-model; DNA, deoxyribonucleic acid; DNMT, DNA methyltransferase; EC50, half maximal effective concentration; ECFP, extended connectivity fingerprints; FFNN, feed-forward neural network; FN, false negatives; FP, false positives; GBT, gradient boosting trees; HAT, histone acetyltransferase; HDAC, histone deacetylase; HDM, histone demethylase; HMT, histone methyltransferase; IC50, half maximal inhibitory concentration; Kd, dissociation constant; Ki, inhibition constant; KIN, kinase; k-NN, k-nearest neighbors; MACCS, molecular access system; MCC, Mathews correlation coefficient; MODI, modelability index; nM, nanomolar; NPV, negative predictive value; PPB, polypharmacology browser; PPV, positive predictive value; RF, random forest; SAR, structure-activity relationships; SD, standard deviation; SEA, similarity ensemble approach; SMILES, simplified molecular input line entry system; SVM, support vector machines; TN, true negatives; TNR, true negative rate; TP, true positives; TPR, true positive rate; μM, micromolar.

REFERENCES

(1) Waddington, C. H. The Epigenotype. *Int. J. Epidemiol.* **2012**, *41* (1), 10–13. https://doi.org/10.1093/ije/dyr184.

(2) Wu, C. -t. Genes, Genetics, and Epigenetics: A Correspondence. *Science (80).* **2001**, *293* (5532), 1103–1105. https://doi.org/10.1126/science.293.5532.1103.

(3) Ganesan, A.; Arimondo, P. B.; Rots, M. G.; Jeronimo, C.; Berdasco, M. The Timeline of Epigenetic Drug Discovery: From Reality to Dreams. *Clin. Epigenetics* **2019**, *11* (1), 1–17. https://doi.org/10.1186/s13148-019-0776-0.

(4) Biswas, S.; Rao, C. M. Epigenetic Tools (The Writers, The Readers and The Erasers) and Their Implications in Cancer Therapy. *Eur. J. Pharmacol.* **2018**, *837* (June), 8–24. https://doi.org/10.1016/j.ejphar.2018.08.021.

(5) Burgess, R. J.; Zhang, Z. Histone Chaperones in Nucleosome Assembly and Human Disease. *Nat. Struct. Mol. Biol.* **2013**, *20* (1), 14–22. https://doi.org/10.1038/nsmb.2461.

(6) Teif, V. B.; Rippe, K. Predicting Nucleosome Positions on the DNA: Combining Intrinsic Sequence Preferences and Remodeler Activities. *Nucleic Acids Res.* **2009**, *37* (17), 5641–5655. https://doi.org/10.1093/nar/gkp610.

(7) Tyagi, M.; Imam, N.; Verma, K.; Patel, A. K. Chromatin Remodelers: We Are the Drivers!! *Nucleus* **2016**, *7* (4), 388–404. https://doi.org/10.1080/19491034.2016.1211217.

(8) Mayran, A.; Drouin, J. Pioneer Transcription Factors Shape the Epigenetic Landscape. *J. Biol. Chem.* **2018**, *293* (36), 13795–13804. https://doi.org/10.1074/jbc.R117.001232.

(9) Esteller, M. Epigenetics in Cancer. *N. Engl. J. Med.* **2008**, *358* (11), 1148–1159. https://doi.org/10.1056/NEJMra072067.

(10) Küçükali, C. İ.; Kürtüncü, M.; Çoban, A.; Çebi, M.; Tüzün, E. Epigenetics of Multiple Sclerosis: An Updated Review. *NeuroMolecular Med.* **2015**, *17* (2), 83–96. https://doi.org/10.1007/s12017-014-8298-6.

(11) Januar, V.; Saffery, R.; Ryan, J. Epigenetics and Depressive Disorders: A Review of Current Progress and Future Directions. *Int. J. Epidemiol.* **2015**, *44* (4), 1364–1387. https://doi.org/10.1093/ije/dyu273.

(12) Brindisi, M.; Saraswati, A. P.; Brogi, S.; Gemma, S.; Butini, S.; Campiani, G. Old but Gold: Tracking the New Guise of Histone Deacetylase 6 (HDAC6) Enzyme as a Biomarker and Therapeutic Target in Rare Diseases. *J. Med. Chem.* **2020**, *63* (1), 23–39. https://doi.org/10.1021/acs.jmedchem.9b00924.

(13) de Lera, A. R.; Ganesan, A. Two-Hit Wonders: The Expanding Universe of Multitargeting Epigenetic Agents. *Curr. Opin. Chem. Biol.* **2020**, *57*, 135–154. https://doi.org/10.1016/j.cbpa.2020.05.009.

(14) Huang, Z.; Jiang, H.; Liu, X.; Chen, Y.; Wong, J.; Wang, Q.; Huang, W.; Shi, T.; Zhang, J. HEMD: An Integrated Tool of Human Epigenetic Enzymes and Chemical Modulators for Therapeutics. *PLoS One* **2012**, *7* (6), e39917. https://doi.org/10.1371/journal.pone.0039917.

(15) Loharch, S.; Bhutani, I.; Jain, K.; Gupta, P.; Sahoo, D. K.; Parkesh, R. EpiDBase: A Manually Curated Database for Small Molecule Modulators of Epigenetic Landscape. *Database* **2015**, *2015*. https://doi.org/10.1093/database/bav013.

(16) Medvedeva, Y. A.; Lennartsson, A.; Ehsani, R.; Kulakovskiy, I. V.; Vorontsov, I. E.; Panahandeh, P.; Khimulya, G.; Kasukawa, T.; Drabløs, F. EpiFactors: A Comprehensive Database of Human Epigenetic Factors and Complexes. *Database* **2015**, *2015*, bav067. https://doi.org/10.1093/database/bav067.

(17) Singh Nanda, J.; Kumar, R.; Raghava, G. P. S. DbEM: A Database of Epigenetic Modifiers Curated from Cancerous and Normal Genomes. *Sci. Rep.* **2016**, *6* (1), 19340. https://doi.org/10.1038/srep19340.

(18) Naveja, J. J.; Medina-Franco, J. L. Insights from Pharmacological Similarity of Epigenetic Targets in Epipolypharmacology. *Drug Discov. Today* **2018**, *23* (1), 141–

150. https://doi.org/10.1016/j.drudis.2017.10.006.

(19) Sessions, Z.; Sánchez-Cruz, N.; Prieto-Martínez, F. D.; Alves, V. M.; Santos, H. P.; Muratov, E.; Tropsha, A.; Medina-Franco, J. L. Recent Progress on Cheminformatics Approaches to Epigenetic Drug Discovery. *Drug Discov. Today* **2020**, *25* (12), 2268–2276. https://doi.org/10.1016/j.drudis.2020.09.021.

(20) Mendez, D.; Gaulton, A.; Bento, A. P.; Chambers, J.; De Veij, M.; Félix, E.; Magariños, M. P.; Mosquera, J. F.; Mutowo, P.; Nowotka, M.; Gordillo-Marañón, M.; Hunter, F.; Junco, L.; Mugumbate, G.; Rodriguez-Lopez, M.; Atkinson, F.; Bosc, N.; Radoux, C. J.; Segura-Cabrera, A.; Hersey, A.; Leach, A. R. ChEMBL: Towards Direct Deposition of Bioassay Data. *Nucleic Acids Res.* **2019**, *47* (D1), D930–D940. https://doi.org/10.1093/nar/gky1075.

(21) Wang, Y.; Zhang, S.; Li, F.; Zhou, Y.; Zhang, Y.; Wang, Z.; Zhang, R.; Zhu, J.; Ren, Y.; Tan, Y.; Qin, C.; Li, Y.; Li, X.; Chen, Y.; Zhu, F. Therapeutic Target Database 2020: Enriched Resource for Facilitating Research and Early Development of Targeted Therapeutics. *Nucleic Acids Res.* **2020**, *48* (D1), D1031–D1041. https://doi.org/10.1093/nar/gkz981.

(22) Naveja, J. J.; Oviedo-Osornio, C. I.; Medina-Franco, J. L. Computational Methods for Epigenetic Drug Discovery: A Focus on Activity Landscape Modeling. In *Advances in Protein Chemistry and Structural Biology*; 2018; Vol. 113, pp 65–83. https://doi.org/10.1016/bs.apcsb.2018.01.001.

(23) Keiser, M. J.; Roth, B. L.; Armbruster, B. N.; Ernsberger, P.; Irwin, J. J.; Shoichet, B. K. Relating Protein Pharmacology by Ligand Chemistry. *Nat. Biotechnol.* **2007**, *25* (2), 197–206. https://doi.org/10.1038/nbt1284.

(24) Liu, X.; Vogt, I.; Haque, T.; Campillos, M. HitPick: A Web Server for Hit Identification and Target Prediction of Chemical Screenings. *Bioinformatics* **2013**, *29* (15), 1910–1912. https://doi.org/10.1093/bioinformatics/btt303.

(25) Hamad, S.; Adornetto, G.; Naveja, J. J.; Chavan Ravindranath, A.; Raffler, J.; Campillos, M. HitPickV2: A Web Server to Predict Targets of Chemical Compounds. *Bioinformatics* **2019**, *35* (7), 1239–1240. https://doi.org/10.1093/bioinformatics/bty759.

(26) Awale, M.; Reymond, J.-L. The Polypharmacology Browser: A Web-Based Multi-Fingerprint Target Prediction Tool Using ChEMBL Bioactivity Data. *J. Cheminform.* **2017**, *9* (1), 11. https://doi.org/10.1186/s13321-017-0199-x.

(27) Awale, M.; Reymond, J.-L. Polypharmacology Browser PPB2: Target Prediction Combining Nearest Neighbors with Machine Learning. *J. Chem. Inf. Model.* **2019**, *59* (1), 10–17. https://doi.org/10.1021/acs.jcim.8b00524.

(28) Wang, L.; Ma, C.; Wipf, P.; Liu, H.; Su, W.; Xie, X.-Q. TargetHunter: An In Silico Target Identification Tool for Predicting Therapeutic Potential of Small Organic Molecules Based on Chemogenomic Database. *AAPS J.* **2013**, *15* (2), 395–406. https://doi.org/10.1208/s12248-012-9449-z.

(29) Gfeller, D.; Grosdidier, A.; Wirth, M.; Daina, A.; Michielin, O.; Zoete, V. SwissTargetPrediction: A Web Server for Target Prediction of Bioactive Small Molecules. *Nucleic Acids Res.* **2014**, *42* (W1), 32–38. https://doi.org/10.1093/nar/gku293.

(30) 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* (W1), W357–W364. https://doi.org/10.1093/nar/gkz382.

(31) Zdrazil, B.; Richter, L.; Brown, N.; Guha, R. Moving Targets in Drug Discovery. *Sci. Rep.* **2020**, *10* (1), 20213. https://doi.org/10.1038/s41598-020-77033-x.

(32) Oprea, T. I.; Bologa, C. G.; Brunak, S.; Campbell, A.; Gan, G. N.; Gaulton, A.; Gomez, S. M.; Guha, R.; Hersey, A.; Holmes, J.; Jadhav, A.; Jensen, L. J.; Johnson, G. L.; Karlson, A.; Leach, A. R.; Ma'ayan, A.; Malovannaya, A.; Mani, S.; Mathias, S. L.; McManus, M. T.; Meehan, T. F.; von Mering, C.; Muthas, D.; Nguyen, D.-T.;

Overington, J. P.; Papadatos, G.; Qin, J.; Reich, C.; Roth, B. L.; Schürer, S. C.; Simeonov, A.; Sklar, L. A.; Southall, N.; Tomita, S.; Tudose, I.; Ursu, O.; Vidović, D.; Waller, A.; Westergaard, D.; Yang, J. J.; Zahoránszky-Köhalmi, G. Unexplored Therapeutic Opportunities in the Human Genome. *Nat. Rev. Drug Discov.* **2018**, *17* (5), 317–332. https://doi.org/10.1038/nrd.2018.14.

(33) Lo, Y.-C.; Rensi, S. E.; Torng, W.; Altman, R. B. Machine Learning in Chemoinformatics and Drug Discovery. *Drug Discov. Today* **2018**, *23* (8), 1538–1546. https://doi.org/10.1016/j.drudis.2018.05.010.

(34) Miljković, F.; Rodríguez-Pérez, R.; Bajorath, J. Machine Learning Models for Accurate Prediction of Kinase Inhibitors with Different Binding Modes. *J. Med. Chem.* **2020**, *63* (16), 8738–8748. https://doi.org/10.1021/acs.jmedchem.9b00867.

(35) Li, H.; Sze, K.; Lu, G.; Ballester, P. J. Machine-learning Scoring Functions for Structure-based Drug Lead Optimization. *WIREs Comput. Mol. Sci.* **2020**, *10* (5), 1–20. https://doi.org/10.1002/wcms.1465.

(36) Gaulton, A.; Hersey, A.; Nowotka, M.; Bento, A. P.; Chambers, J.; Mendez, D.; Mutowo, P.; Atkinson, F.; Bellis, L. J.; Cibrián-Uhalte, E.; Davies, M.; Dedman, N.; Karlsson, A.; Magariños, M. P.; Overington, J. P.; Papadatos, G.; Smit, I.; Leach, A. R. The ChEMBL Database in 2017. *Nucleic Acids Res.* **2017**, *45* (D1), D945–D954. https://doi.org/10.1093/nar/gkw1074.

(37) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50* (5), 742–754. https://doi.org/10.1021/ci100050t.

(38) Mayr, A.; Klambauer, G.; Unterthiner, T.; Steijaert, M.; Wegner, J. K.; Ceulemans, H.; Clevert, D.-A.; Hochreiter, S. Large-Scale Comparison of Machine Learning Methods for Drug Target Prediction on ChEMBL. *Chem. Sci.* **2018**, *9* (24), 5441–5451. https://doi.org/10.1039/C8SC00148K.

(39) Sirous, H.; Campiani, G.; Brogi, S.; Calderone, V.; Chemi, G. Computer-Driven

Development of an in Silico Tool for Finding Selective Histone Deacetylase 1 Inhibitors. *Molecules* **2020**, *25* (8), 1952. https://doi.org/10.3390/molecules25081952.

(40) Li, S.; Ding, Y.; Chen, M.; Chen, Y.; Kirchmair, J.; Zhu, Z.; Wu, S.; Xia, J. HDAC3i-Finder: A Machine Learning-based Computational Tool to Screen for HDAC3 Inhibitors. *Mol. Inform.* **2020**, minf.202000105. https://doi.org/10.1002/minf.202000105.

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

(42) Speck-Planche, A.; Scotti, M. T. BET Bromodomain Inhibitors: Fragment-Based in Silico Design Using Multi-Target QSAR Models. *Mol. Divers.* **2019**, *23* (3), 555–572. https://doi.org/10.1007/s11030-018-9890-8.

(43) Kim, S.; Chen, J.; Cheng, T.; Gindulyte, A.; He, J.; He, S.; Li, Q.; Shoemaker, B. A.; Thiessen, P. A.; Yu, B.; Zaslavsky, L.; Zhang, J.; Bolton, E. E. PubChem 2019 Update: Improved Access to Chemical Data. *Nucleic Acids Res.* **2019**, *47* (D1), D1102–D1109. https://doi.org/10.1093/nar/gky1033.

(44) Golbraikh, A.; Muratov, E.; Fourches, D.; Tropsha, A. Data Set Modelability by QSAR. *J. Chem. Inf. Model.* **2014**, *54* (1), 1–4. https://doi.org/10.1021/ci400572x.

(45) Altman, N. S. An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression. *Am. Stat.* **1992**, *46* (3), 175–185. https://doi.org/10.1080/00031305.1992.10475879.

(46) Tin Kam Ho. The Random Subspace Method for Constructing Decision Forests. *IEEE Trans. Pattern Anal. Mach. Intell.* **1998**, *20* (8), 832–844. https://doi.org/10.1109/34.709601.

(47) Friedman, J. Greedy Function Approximation : A Gradient Boosting Machine Author (s): Jerome H . Friedman Source : The Annals of Statistics , Vol . 29 , No . 5 (Oct ., 2001), Pp . 1189-1232 Published by : Institute of Mathematical Statistics Stable URL :

Http://Www. *Ann. Stat.* **2001**, *29* (5), 1189–1232.

(48) Cortes, C.; Vapnik, V. Support-Vector Networks. *Mach. Learn.* **1995**, *20* (3), 273–297. https://doi.org/10.1007/BF00994018.

(49) Hopfield, J. J. Neural Networks and Physical Systems with Emergent Collective Computational Abilities. *Proc. Natl. Acad. Sci.* **1982**, *79* (8), 2554–2558. https://doi.org/10.1073/pnas.79.8.2554.

(50) 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. https://doi.org/10.1021/ci010132r.

(51) Mansouri, K.; Abdelaziz, A.; Rybacka, A.; Roncaglioni, A.; Tropsha, A.; Varnek, A.; Zakharov, A.; Worth, A.; Richard, A. M.; Grulke, C. M.; Trisciuzzi, D.; Fourches, D.; Horvath, D.; Benfenati, E.; Muratov, E.; Wedebye, E. B.; Grisoni, F.; Mangiatordi, G. F.; Incisivo, G. M.; Hong, H.; Ng, H. W.; Tetko, I. V.; Balabin, I.; Kancherla, J.; Shen, J.; Burton, J.; Nicklaus, M.; Cassotti, M.; Nikolov, N. G.; Nicolotti, O.; Andersson, P. L.; Zang, Q.; Politi, R.; Beger, R. D.; Todeschini, R.; Huang, R.; Farag, S.; Rosenberg, S. A.; Slavov, S.; Hu, X.; Judson, R. S. CERAPP: Collaborative Estrogen Receptor Activity Prediction Project. *Environ. Health Perspect.* **2016**, *124* (7), 1023–1033. https://doi.org/10.1289/ehp.1510267.

(52) Alves, V. M.; Golbraikh, A.; Capuzzi, S. J.; Liu, K.; Lam, W. I.; Korn, D. R.; Pozefsky, D.; Andrade, C. H.; Muratov, E. N.; Tropsha, A. Multi-Descriptor Read Across (MuDRA): A Simple and Transparent Approach for Developing Accurate Quantitative Structure–Activity Relationship Models. *J. Chem. Inf. Model.* **2018**, *58* (6), 1214–1223. https://doi.org/10.1021/acs.jcim.8b00124.

(53) MEDINA-FRANCO, J. L.; Sánchez-Cruz, N. {Supporting Information For. https://doi.org/10.6084/m9.figshare.13519580.

(54) 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. https://doi.org/10.12688/f1000research.21540.2.

(55) 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* (11), minf.202000050. https://doi.org/10.1002/minf.202000050.

(56) Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Model.* **1988**, *28* (1), 31–36. https://doi.org/10.1021/ci00057a005.

(57) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, E. Scikit-Learn: Machine Learning in {P}ython. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.

(58) Tetko, I. V.; Sushko, I.; Pandey, A. K.; Zhu, H.; Tropsha, A.; Papa, E.; Öberg, T.; Todeschini, R.; Fourches, D.; Varnek, A. Critical Assessment of QSAR Models of Environmental Toxicity against Tetrahymena Pyriformis: Focusing on Applicability Domain and Overfitting by Variable Selection. *J. Chem. Inf. Model.* **2008**, *48* (9), 1733–1746. https://doi.org/10.1021/ci800151m.

(59) Sushko, I.; Novotarskyi, S.; Körner, R.; Pandey, A. K.; Cherkasov, A.; Li, J.; Gramatica, P.; Hansen, K.; Schroeter, T.; Müller, K.-R.; Xi, L.; Liu, H.; Yao, X.; Öberg, T.; Hormozdiari, F.; Dao, P.; Sahinalp, C.; Todeschini, R.; Polishchuk, P.; Artemenko, A.; Kuz'min, V.; Martin, T. M.; Young, D. M.; Fourches, D.; Muratov, E.; Tropsha, A.; Baskin, I.; Horvath, D.; Marcou, G.; Muller, C.; Varnek, A.; Prokopenko, V. V.; Tetko, I. V. Applicability Domains for Classification Problems: Benchmarking of Distance to Models for Ames Mutagenicity Set. *J. Chem. Inf. Model.* **2010**, *50* (12), 2094–2111. https://doi.org/10.1021/ci100253r.

(60) Wilson, J. E.; Patel, G.; Patel, C.; Brucelle, F.; Huhn, A.; Gardberg, A. S.; Poy, F.;

Cantone, N.; Bommi-Reddy, A.; Sims, R. J.; Cummings, R. T.; Levell, J. R. Discovery of CPI-1612: A Potent, Selective, and Orally Bioavailable EP300/CBP Histone Acetyltransferase Inhibitor. *ACS Med. Chem. Lett.* **2020**, *11* (6), 1324–1329. https://doi.org/10.1021/acsmedchemlett.0c00155.

(61) Chen, J.; Li, Y.; Zhang, J.; Zhang, M.; Wei, A.; Liu, H.; Xie, Z.; Ren, W.; Duan, W.; Zhang, Z.; Shen, A.; Hu, Y. Discovery of Selective HDAC/BRD4 Dual Inhibitors as Epigenetic Probes. *Eur. J. Med. Chem.* **2021**, *209*, 112868. https://doi.org/10.1016/j.ejmech.2020.112868.

Bioinformatics, 2020, 1–7
doi: 10.1093/bioinformatics/btaa982
Original Paper

Structural Bioinformatics

Extended connectivity interaction features: improving binding affinity prediction through chemical description

Norberto Sánchez-Cruz[1,*], José L. Medina-Franco[1], Jordi Mestres[2,3] and Xavier Barril[4,5]

[1]Department of Pharmacy, School of Chemistry, Universidad Nacional Autónoma de México, Mexico City 04510, Mexico, [2]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, [3]Chemotargets SL, Parc Cientific de Barcelona (PCB), 08028 Barcelona, Catalonia, Spain, [4]Institut de Biomedicina de la Universitat de Barcelona (IBUB) and Facultat de Farmacia, Universitat de Barcelona, 08028 Barcelona, Spain and [5]Catalan Institution for Research and Advanced Studies (ICREA), 08010 Barcelona, Spain

*To whom correspondence should be addressed.

Associate Editor: Arne Elofsson

Received on September 18, 2020; revised on October 27, 2020; editorial decision on November 9, 2020; accepted on November 10, 2020

Abstract

Motivation: Machine-learning scoring functions (SFs) have been found to outperform standard SFs for binding affinity prediction of protein–ligand complexes. A plethora of reports focus on the implementation of increasingly complex algorithms, while the chemical description of the system has not been fully exploited.

Results: Herein, we introduce Extended Connectivity Interaction Features (ECIF) to describe protein–ligand complexes and build machine-learning SFs with improved predictions of binding affinity. ECIF are a set of protein−ligand atom-type pair counts that take into account each atom's connectivity to describe it and thus define the pair types. ECIF were used to build different machine-learning models to predict protein–ligand affinities (pK_d/pK_i). The models were evaluated in terms of 'scoring power' on the Comparative Assessment of Scoring Functions 2016. The best models built on ECIF achieved Pearson correlation coefficients of 0.857 when used on its own, and 0.866 when used in combination with ligand descriptors, demonstrating ECIF descriptive power.

Availability and implementation: Data and code to reproduce all the results are freely available at https://github.com/DIFACQUIM/ECIF.

Contact: norberto.sc90@gmail.com

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Molecular docking plays an important role in structure-based drug design (Lyu *et al.*, 2019). Its main goals are both the prediction of the binding pose and the binding affinity of a small molecule to a macromolecular target, often a protein. A number of works in this field have proven molecular docking to be a valuable tool for the exploration of the chemical space to identify high-affinity compounds or active molecules with novel chemical scaffolds for a particular target (Kuck *et al.*, 2010; Leman *et al.*, 2020). One of the key elements in molecular docking is its scoring function (SF), which is required to estimate the binding affinity of putative protein–ligand complexes. As detailed in the Comparative Assessment of Scoring Functions (CASF) (Cheng *et al.*, 2009; Li *et al.*, 2014a,b, 2018b; Su *et al.*, 2019), a robust SF is one that performs well on multiple tasks, including scoring (obtaining binding scores in a linear correlation with experimental binding data), ranking (correctly rank the known ligands of a given target protein by their binding affinities), docking (identifying the native ligand pose among computer-generated decoys) and screening (identifying the true binders to a given protein among a pool of random molecules). Depending on the goals of the research project in which an SF is used (e.g. lead optimization or virtual screening), one or several of these features are more important.

SFs can be classified into two groups (Ain *et al.*, 2015; Liu and Wang, 2015): classical and machine-learning SFs. Classical SFs are based on a functional form to establish the relationship between the features characterizing a protein–ligand complex and its binding affinity; those SFs include physics-based methods (force fields), linear combinations of empirical terms (empirical SFs) and knowledge-based potentials. Common protein–ligand docking packages such as GLIDE (Friesner *et al.*, 2004; 2006; Halgren *et al.*, 2004), GOLD (Jones *et al.*, 1997), MOE, AutoDock Vina (Trott and Olson, 2009) and rDock (Ruiz-Carmona *et al.*, 2014) make use of classical SFs. It has been extensively demonstrated that this kind of SFs performs

Table 1. Representative machine-learning SFs

Scoring Function	Algorithm	Description of protein–ligand complexes	Reference
RF-Score	RF	Protein–ligand atom-type pair counts	Ballester *et al.* (2010)
NN-Score 2.0	ANN	Autodock Vina interaction terms, protein–ligand atom-type pair counts and electrostatic terms (BINANA)	Durrant and McCammon (2011)
ID-Score	SVM	Nine categories of descriptors related to protein–ligand interactions	Li *et al.* (2013)
SFCscore^{RF}	RF	SFCscore interaction terms	Zilian and Sotriffer (2013)
ΔVinaRF$_{20}$	RF	Autodock Vina interaction terms and additional molecular descriptors	Wang and Zhang (2017)
RI-Score	RF	Rigidity index descriptors	Nguyen *et al.* (2017)
TNet-BP	CNN	Algebraic topology	Cang and Wei (2017)
K$_{DEEP}$	CNN	Molecular descriptors embedded into a 3D grid	Jiménez *et al.* (2018)
TopBP-ML	GBT	Algebraic topology	Cang *et al.* (2018)
TopBP-DL	CNN	Algebraic topology	Cang *et al.* (2018)
Pafnucy	CNN	Molecular descriptors embedded into a 3D grid	Stepniewska-Dziubinska *et al.* (2018)
PLEC-nn	DNN	Hashed fingerprint constructed by pairing ligand and protein atoms according to its environment	Wójcikowski *et al.* (2019)
EIC-Score	GBT	Differential geometry representations	Nguyen and Wei (2019b)
AGL-Score	GBT	Statistical features of the adjacency and Laplacian matrices of multiscale weighted labeled algebraic subgraphs	Nguyen and Wei (2019a)
OnionNet	CNN	Rotation-free element pair-specific contacts between ligands and protein atoms, grouped into different distance ranges	Zheng *et al.* (2019)
ΔVinaXGB	XGBT	Autodock Vina score and molecular descriptors, including water molecules	Lu *et al.* (2019)
NNScore::LD	FFNN	NNScore 2.0 features and RDKit ligand descriptors	Boyles *et al.* (2020)
RosENet	CNN	Molecular mechanics energies from Rosetta force field and molecular descriptors embedded onto a 3D grid	Hassan-Harrirou *et al.* (2020)
ECIF-GBT	GBT	Protein–ligand atom-type pair counts considering each atoms connectivity	This work
ECIF::LD-GBT	GBT	Protein–ligand atom-type pair counts considering each atoms connectivity and RDKit ligand descriptors	This work

well in docking and screening tasks, but scoring and ranking tasks remain to be a challenge for them (Cheng *et al.*, 2009; Li *et al.*, 2014a,b, 2018b; Su *et al.*, 2019). In contrast, machine-learning SFs learn the connection between protein–ligand binding affinity and the features describing the system through a machine-learning algorithm, such as random forest (RF) (Ballester *et al.*, 2010; Wang and Zhang, 2017; Zilian and Sotriffer, 2013), support vector machines (SVM) (Li *et al.*, 2013) or artificial neural networks (ANN) (Durrant and McCammon, 2011). Machine-learning SFs have been found to consistently outperform classical SFs in all four tasks, especially for scoring and ranking and unlike machine-learning SFs, classical SFs are unable to exploit large volumes of structural data, suggesting that this performance gap is expected to increase in the future (Li *et al.*, 2018a, 2019).

The rise of machine-learning SFs since early 2010 is largely due to the increasing availability of structural and binding affinity data of protein–ligand complexes that allow the training of the models (Liu *et al.*, 2015, 2017). RF-Score (Ballester *et al.*, 2010) was arguably the first machine-learning method achieving high performance in the CASF-2007 benchmark in terms of scoring power. This model was derived using a random forest algorithm and employing a combination of 36 protein–ligand atom-type pair counts as features for the description of protein–ligand complexes. Since then, a number of related approaches implementing different machine-learning algorithms and representations have emerged (Table 1). For example, ID-Score (Li *et al.*, 2013) employed a set of 50 descriptors associated to protein–ligand binding and an SVM model. Terms from classical SFs have also been used as descriptors for the development of machine learning SFs, often in combination with additional descriptors. Among these SFs are NNScore 2.0 (Durrant and McCammon, 2011) using ANN, SFCscore^{RF} (Zilian and Sotriffer, 2013) and ΔVinaRF$_{20}$ (Wang and Zhang, 2017) employing RF, and more recently ΔVinaXGB (Lu *et al.*, 2019) derived making use of extreme gradient boosting trees (XGBT).

Recently, a different class of machine learning SFs has been developed. The so-called deep-learning SFs, which differ from standard machine-learning algorithms in the fact that they rely on deep neural networks (DNNs) architectures to derive the models. The key characteristic of DNNs is the use of a large number of hidden layers and neurons for the set-up of the network. For example, PLEC-nn (Wójcikowski *et al.*, 2019) is a model derived from a DNN of three hidden layers with 200 neurons each, that uses a fingerprint-like representation of the protein–ligand complexes. Among the models derived from similar representations, PLEC-nn has shown the best performance in terms of scoring power on the CASF-2016 benchmark. Other algorithms frequently used for the development of deep-learning SFs are convolutional neural networks (CNNs), a class of DNNs commonly applied for image analysis (Krizhevsky, 2017). CNNs have the capability of automatically generating features from the input data, making possible the extraction of features from the crystal structure of protein–ligand complexes. Examples of CNN-based SFs include TNet-BP (Cang and Wei, 2017) and TopBP-DL (Cang *et al.*, 2018) that use algebraic topology descriptors to represent the 3D geometry of the protein–ligand complexes, as well as K$_{DEEP}$ (Jiménez *et al.*, 2018) and Pafnucy (Stepniewska-Dziubinska *et al.*, 2018) that embed molecular descriptors into a 3D image-like representation of the complexes.

Since 2017 there is a clear trend toward the use of deep-learning algorithms for the development of SFs. However, machine-learning algorithms based on decision trees, particularly gradient boosting trees (GBT), continue to be competitive for the task of binding affinity prediction in combination with different types of descriptors. Examples include EIC-Score (Nguyen and Wei, 2019b), where the descriptors are based on differential geometry representations of the protein–ligand complexes, and AGL-Score (Nguyen and Wei, 2019a), where the descriptors are obtained from algebraic graph representations. To the best of our knowledge, the best performing SFs in terms of scoring power are AGL-Score and TopBP-DL, achieving Pearson correlation coefficients of 0.833 and 0.848, respectively, when evaluated on the CASF-2016 benchmark. In-depth reviews of machine learning SFs can be found elsewhere (Ain *et al.*, 2015; Li *et al.*, 2020).

As stated above, multiple descriptions have been implemented for the development of machine-learning SFs, protein−ligand atom-type pair counts being arguably the simplest ones. Although increasingly complex representations have been used, little work has been done on the chemical description of such simple pair counts. It has been shown that using SYBYL atom types and Structural Interaction Fingerprints (SIFts) (Deng *et al.*, 2004) for the construction of the pair counts did not generally improve the predictive power of SFs derived using RF (Ballester *et al.*, 2014). However, a separate study demonstrated that taking into account the connectivity of the atoms could lead to better performances in the same task (Wójcikowski *et al.*, 2019).

Herein, we describe Extended Connectivity Interaction Features (ECIF), a set of protein−ligand atom-type pair counts that take into account each atom's connectivity to describe it and thus define the pair types. To demonstrate the descriptive power of ECIF, we used them to derive machine-learning SFs and compared their scoring power against state-of-the-art machine learning models on the CASF-2016 benchmark.

2 Materials and methods

2.1 Extended connectivity interaction features

ECIF are defined as a set of protein−ligand atom-type pair counts that consider each atom's connectivity to define the atoms involved in the pairs. Atom definitions for ECIF rely on the atom environments concept originally presented in the development of Extended Connectivity Fingerprints (ECFP) (Rogers and Hahn, 2010). As such, we defined an atom type in a molecule considering six atomic features: atom symbol, explicit valence, number of attached heavy atoms, number of attached hydrogens, aromaticity and ring membership. For the ligands, atom types were assigned as interpreted by the Python open source chemoinformatic software *RDKit* (2020.03.1), from the corresponding standard data format (SDF) file. Regarding the proteins, atom types were manually assigned to 22 different atom types in a dictionary-based mapping according to their residue and atom label in the corresponding PDB structural file. The list of mapped PDB-ECIF atom types is included in Supplementary Table S1. It should be noted that atom types were defined purely based on the connectivity of the molecule while the 3D information was not taken into account. These considerations imply that protonation states from the ligand are inferred from the SDF file, while for protein residues they were not considered, since they were assigned in a predefined manner. Once all atom types in both protein and ligand were assigned, the number of each possible pair of protein–ligand atom types was tallied, with the only additional criterion being a predefined distance cutoff. This is schematically represented inFigure 1. Under this scheme, each complex was represented as a vector of 1540 integer-valued features, where each position corresponds to the count of a pair-wise combination of protein–ligand atom types with preserved directionality. For instance, the atom-type pair 'O;2;1;0;0:0'—'N;3;2;1;0;0' from ECIF is different from the pair 'N;3;2;1;0;0'—'O;2;1;0;0;0', since in the first case the oxygen atom refers to a protein atom while the nitrogen atom refers to a ligand atom and vice versa for the second case. Under this description, structural water molecules could be included as part of the protein, but the automatic modeling of these molecules would be prone to error, so although water-bridged interactions play an important role in protein–ligand binding, they were not considered in the present study.

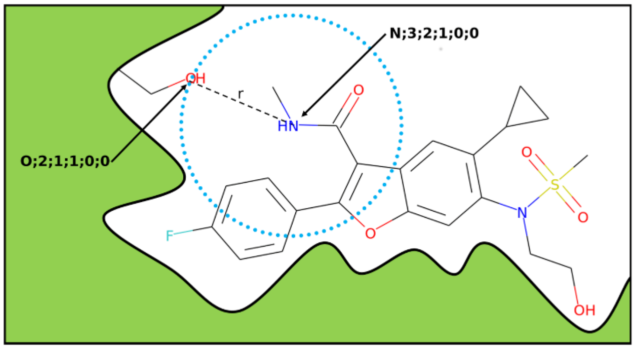


Fig. 1. Schematic 2D representation of a protein−ligand atom-type pair from ECIF. The atom-type pair 'O;2;1;0;0;0' – 'N;3;2;1;0;0' is represented. For each of the ligand atoms, all protein atoms within the circumference delimited by a radius of length **r** are considered for the count of the pairs. This radius is the only adjustable parameter for the calculation of ECIF

As the only parameter to be adjusted for the calculation of ECIF is a distance cutoff, given that for similar approaches different distance criteria have been used (e.g. 12 Å for RF-Score and 4.5 Å for PLEC-nn), we studied the effect of this parameter by computing ECIF using different distance cutoffs ranging from 4.0 to 15.0 Å in increments of 0.5 Å. These sets of descriptors were used to build machine-learning SFs for binding affinity prediction.

2.2 Training and test sets

This work was focused on the task of scoring, the capability of an SF of predicting binding scores in a linear correlation with experimental data. To this end, we employed protein–ligand complexes from the PDBbind database (Liu *et al.*, 2015, 2017), a collection of the experimentally measured binding affinity data for all types of biomolecular complexes deposited in the Protein Data Bank (PDB). Each release of the PDBbind includes a 'general set' containing all protein–ligand structures in the database and a 'refined set' including a subset of structures satisfying strict criteria related mainly to structure quality. Additionally, the 'core set', extracted from the 'refined set', is also included as a representative non-redundant subset and commonly used for benchmark purposes in CASF, being 2016 the latest release. We used the 'refined set' ($n = 4057$) minus 'core set' ($n = 285$) from the PDBbind 2016 as our primary source of training data, and the 'core set' for validation of the models. All protein structures were used without additional treatment, while all ligands were processed using Standardizer, JChem 20.11.0, 2020, *ChemAxon* to add explicit hydrogen atoms (in case they were missing) and to perceive aromaticity in an interpretable way for *RDKit*. Atom types for all ligands in the training set were analyzed and three complexes with ligands containing a single occurrence of atom types among the dataset were discarded (PDB IDs: 2YLC, 3O7U, 3ZNR).

It has been reported that including the lower quality data from the remaining PDBbind 'general set' can improve the performance of machine-learning SFs (Boyles *et al.*, 2020; Li *et al.*, 2015; Lu *et al.*, 2019; Wójcikowski *et al.*, 2019). For that reason, we included data from the 'general set' of the latest release of PDBbind 2019 meeting the following criteria: (i) structures resolved by X-Ray crystallography with a resolution better or equal than 3.0 Å, (ii) binding data reported accurately (not '>', '<' or '~') as inhibition constant (K_i) or dissociation constant (K_d) with a range from 1pM to 10 mM, (iii) atom types of the ligand already included in the 'refined set' and (iv) structures not containing any protein–ligand atom pair at a distance of 2.0 Å or less. The last two criteria were implemented to avoid the inclusion of organometallic and coordination compounds, as well as complexes bound through covalent bonds, and thereby delimit the applicability domain of the herein generated models. In this way, the training set used for this study consisted of 9299 protein–ligand complexes while the validation set consisted of the original 285 structures used in CASF-2016. The full list of complexes contained in the training and validation sets is included in Supplementary Table S2.

2.3 Machine-learning models

To assess the descriptive power of ECIF for binding affinity prediction, two well-established ensemble algorithms for the derivation of machine-learning SFs were implemented in this work: RF and GBT. Since the focus of this study was to show the application of ECIF and not the obtention of an optimal SF, we selected simple and previously described hyperparameters for similar models as opposed to adjusting their hyperparameters. All models herein described were implemented using the Scikit-learn Python library (0.22.1) (Pedregosa *et al.*, 2011). RF models were built using 500 fully grown trees as described for the development of RF-Score (Ballester *et al.*, 2010), but instead of a fixed number for the maximum number of features, we used 0.33 as the fraction of features to look at for the best split given the larger number of descriptors included in ECIF. GBT models were built using the same hyperparameters as those described for the development of TopBP-ML (Cang *et al.*, 2018): 20 000 boosting stages, maximum depth of 8, a learning rate of 0.005, least squares regression as the loss function to be optimized, a fraction of samples to fit the individual learners of 0.7, and 'sqrt' as the fraction of features to look at for the best split. All remaining parameters were set as default. All models were built to predict the binding affinity of the protein–ligand complexes denoted as pK, the negative base-10 logarithm of either K_i or K_d. The performance of the models was assessed in terms of the Pearson correlation coefficient (R_p) and the root mean square error (RMSE) over the 'core set' for CASF-2016.

2.4 Reference features

In order to compare the herein proposed ECIF to related sets of descriptors, we computed the 36 features originally described for the development of RF-Score (herein called ELEMENTS) to build models using the same algorithms described in Section 2.3. Those features were computed using an in-house Python script and considering the same distance cutoff criteria as for ECIF, ranging from 4.0 to 15.0 Å in increments of 0.5 Å. It has been shown that the addition of purely ligand-based descriptors to those obtained from the protein–ligand complexes can improve the performance of machine-learning SFs (Boyles *et al.*, 2020; Cang *et al.*, 2018). To assess if the models herein proposed to follow the same trend, we employed the methodology described by Boyles *et al.* (2020) for the computation of ligand descriptors (LD). In short, 200 molecular descriptors available in the 'Descriptors' module of *RDKit* (2020.03.1) were computed for each ligand; features with zero variance across the dataset features null valued and features with extreme values were excluded. By using this methodology, a total of 170 *RDKit* descriptors were preserved. The full list of descriptors is included in Supplementary Table S3. The set of generated ligand descriptors was added to the previously described sets of features ECIF and ELEMENTS to build new SFs derived from the RF and GBT algorithms described above using the same hyperparameters.

2.5 Feature importance on the best model

The best performing model based on ECIF was selected according to its performance on the 'core set' from CASF-2016. For this model, we tried different combinations of hyperparameters to optimize the RMSE of the predictions. An analysis on the most important features for this optimized model was performed, with a focus on those involving atom-type pair counts from the herein presented ECIF.

2.6 Stability of the results

The 'core set' from CASF-2016 was constructed to be a balanced sample of the protein families and binding affinity values included in the PDBbind database. However, its small size tends to yield overly optimistic results, so additional metrics for performance estimation are needed. To provide more robust metrics for comparison of the herein reported models, we tested all the models built on ELEMENTS and ECIF with 10-fold cross validations over the training set, with each fold consisting of 929 or 930 structures (given that 9299 is not divisible by 10). All comparisons were performed with models trained using the hyperparameters described in Section 2.3.

2.7 Comparison with different scoring functions

The best non-optimized models herein generated for ECIF and its corresponding combination with LD (ECIF::LD) were compared to state-of-the-art SFs evaluated on the same benchmark (CASF-2016), being TopBP-DL, TopBP-ML, AGL-Score, EIC-Score, K_DEEP, RI-Score, RosENet, PLEC-nn, OnionNet, ΔVinaXGB and Pafnucy.

3 Results and discussion

3.1 Extended connectivity interaction features

The first step for the construction of ECIF was the identification of the atom types present in the ligands of the training set to then define the possible protein–ligand atom pairs. According to the six atomic features described in Section 2.1, a set of 70 atom types were identified for the ligands in the training set, while 22 atom types were defined for proteins. Noteworthy, the atom types for proteins are a subset of atom types for ligands and the 70 atom types identified can be considered as a subdivision of the atom types used in ELEMENTS description and they can be mapped in a straightforward manner (Table 2). According to this, ECIF were defined as the set of counts for 1540 (22×70) possible protein–ligand atom pairs; as stated in Section 2.1, the distance cutoff was the only parameter adjusted for the calculation of ECIF.

3.2 Performance of ECIF in binding affinity prediction

To study the effect of the distance cutoff parameter for the calculation of ECIF, we computed multiple sets of ECIF using different distance cutoffs ranging from 4.0 to 15.0 Å in increments of 0.5 Å. These sets of descriptors were used to build machine-learning SFs for binding affinity prediction using two different algorithms, RF and GBT as described in Section 2.3. To compare ECIF to related descriptors, we built the same models using ELEMENTS computed under the same distance cutoff criteria. Figure 2 (left) shows the performance of the models generated using different distance cutoffs over the CASF-2016 benchmark. Performances for each model are shown as R_p between predicted and experimental pK (results for RMSE are included in Supplementary Fig. S1. For all distance cutoff criteria and in the context of each machine-learning algorithm, models built from ECIF outperformed those employing ELEMENTS in terms of both R_p and RMSE. Regarding ECIF, models derived using GBT outperformed consistently those constructed using RF, showing the highest R_p and the lowest RMSE for all distance cutoffs. In the case of ELEMENTS, the best performing algorithm was RF, but none of these models was comparable with the best model based on ECIF. Out of the 23 ECIF-GBT models, the one constructed using a distance cutoff criterion of 6.0 Å for the count of the atom pairs (ECIF6-GBT) had the best performance, with an R_p of 0.857 and an RMSE of 1.193.

3.3 Combination of ECIF with ligand-based features

It has been reported that the combination of features describing protein–ligand complexes with purely ligand-based ones can improve the performance of machine-learning SFs. As detailed in Section 2.4, we computed a set of 170 ligand descriptors and added them to both ECIF and ELEMENTS to train the new models (ECIF::LD and ELEMENTS::LD), using the machine-learning algorithms described in Section 2.3 to derive them. The performance of the models generated using different distance cutoffs over the CASF-2016 benchmark is showed in Figure 2 (right). Performances for each model are shown as R_p between predicted and experimental pK (results for RMSE are included in Supplementary Fig. S1). Regardless of the distance cutoff, the performances in terms of R_p and RMSE improved for all models. Surprisingly, even with the addition of ligand-based descriptors, models built from ELEMENTS do not outperform those built using ECIF alone. Consequently, models combining ECIF with LD were the best-performing ones, particularly when derived from

Table 2. Mapping of ECIF to ELEMENTS atom types for molecules in the refined set of the PDBbind database 2016

ELEMENTS atom type	ECIF atom types
Br	Br;1;1;0;0;0
C	C;3;3;0;1;1—C;4;1;1;0;0—C;4;1;2;0;0—C;4;1;3;0;0 C;4;2;0;0;0—C;4;2;1;0;0—C;4;2;1;0;1—C;4;2;1;1;1 C;4;2;2;0;0—C;4;2;2;0;1—C;4;3;0;0;0—C;4;3;0;0;1 C;4;3;0;1;1—C;4;3;1;0;0—C;4;3;1;0;1—C;4;4;0;0;0 C;4;4;0;0;1—C;5;3;0;0;0—C;5;3;0;1;1—C;6;3;0;0;0
Cl	Cl;1;1;0;0;0
Fl	F;1;1;0;0;0
I	I;1;1;0;0;0
N	N;3;1;0;0;0—N;3;1;1;0;0—N;3;1;2;0;0—N;3;2;0;0;0 N;3;2;0;0;1—N;3;2;0;1;1—N;3;2;1;0;0—N;3;2;1;0;1 N;3;2;1;1;1—N;3;3;0;0;0—N;3;3;0;0;1—N;3;3;0;1;1 N;4;1;2;0;0—N;4;1;3;0;0—N;4;2;1;0;0—N;4;2;2;0;0 N;4;2;2;0;1—N;4;3;0;0;0—N;4;3;0;0;1—N;4;3;1;0;0 N;4;3;1;0;1—N;4;4;0;0;0—N;4;4;0;0;1—N;5;2;0;0;0 N;5;3;0;0;0—N;5;3;0;1;1
O	O;2;1;0;0;0—O;2;1;1;0;0—O;2;2;0;0;0—O;2;2;0;0;1 O;2;2;0;1;1
P	P;5;4;0;0;0—P;6;4;0;0;0—P;6;4;0;0;1—P;7;4;0;0;0
S	S;2;1;0;0;0—S;2;1;1;0;0—S;2;2;0;0;0—S;2;2;0;0;1 S;2;2;0;1;1—S;3;3;0;0;0—S;3;3;0;0;1—S;4;3;0;0;0 S;6;4;0;0;0—S;6;4;0;0;1—S;7;4;0;0;0

GBT. Once again, the best model was obtained using a distance cutoff of 6.0 Å (ECIF6::LD), with an R_p of 0.866 and an RMSE of 1.169. Figure 3 shows the correlation between predicted and experimental pK of ECIF6-GBT and ECIF6::LD-GBT models. According to a Mann–Whitney U-test at 95% confidence over the distribution of 100 bootstrapped R_p and RMSE values, the inclusion of LD led to a marginal but statistically significant improvement (P value $<$ 0.05).

3.4 Feature importance on the best model

With ECIF6::LD-GBT as the best performing model, we tried 1440 different combinations of hyperparameters of the GBT algorithm to optimize the predictions' RMSE. To keep the search space short, we fixed the number of boosting stages at 20 000 and the learning rate at 0.005. The possible choices for the remaining hyperparameters as well as their effects on the performance of the model are included in Supplementary Table S4 and Fig. S2. Hyperparameters yielding to the lower RMSE were a maximum depth of 6, least squares regression as the loss function to be optimized, a fraction of samples to fit the individual learners of 0.8, and 0.25 as the fraction of features to look at for the best split. The RMSE for the ECIF6::LD-GBT model trained using the optimized hyperparameters and evaluated over the CASF-2016 benchmark decreased from 1.169 to 1.154. However, the R_p remained practically with no change, going from 0.866 to 0.867.

An analysis of the feature importance of the optimized ECIF6::LD-GBT model revealed that among the 1710 features, only 373 have an importance higher than a random selection (>1/1710). From them, 276 correspond to pair counts from ECIF and 97 to ligand descriptors. Among the top 15 most important features (Supplementary Table S5), there are three pair counts from ECIF being C;4;1;3;0;0—C;4;3;0;1;1, C;4;3;1;0;0—C;4;3;0;1;1 and C;4;3;1;0;0—C;4;2;1;1;1, which involve aliphatic carbons in proteins (methyl groups and alpha carbons) and aromatic carbons in ligands (with and without substituents). Interestingly, these nonpolar pair counts have a greater importance than those involving polar atoms. This is an example on how machine-learning algorithms can capture patterns in the training data are not commonly explored by classical SFs, but also shows the difficulty on its interpretation,

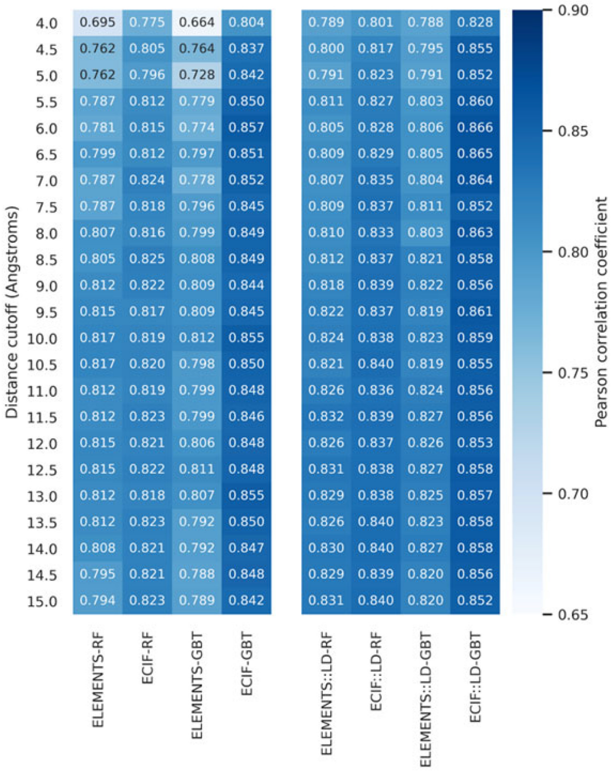


Fig. 2. Heat maps comparing the performance of machine-learning SFs based on ECIF and ELEMENTS. Results are shown in terms of Pearson correlation coefficient between predicted and experimental pK on the 'core set' ($n = 285$) from the CASF-2016. Models were built using multiple distance cutoff criteria and two different algorithms: RF and GBT. Results are shown for the sets of descriptors evaluated on its own (left) and in combination with ligand descriptors (right)

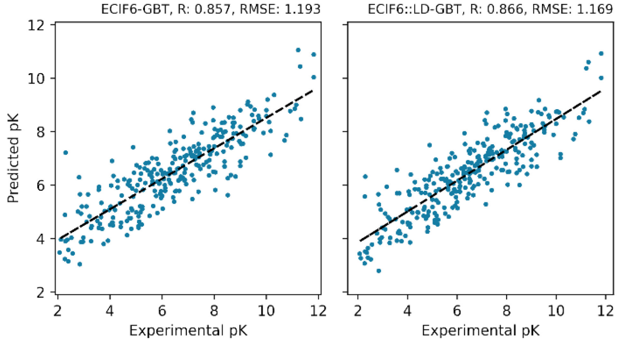


Fig. 3. Binding affinity predictions for ECIF6-GBT (left) and ECIF6::LD-GBT (right) on the 'core set' (n = 285) from the CASF-2016

3.5 Stability of the results

To provide more robust metrics for comparison of the herein reported models, we tested all of them with 10-fold cross validations over the training set. To avoid the hyperparameter selection bias on the best models, all further comparisons involve models trained using the hyperparameters described in Section 2.3. Average R_p and RMSE between predicted and experimental pK for each model are included as Supplementary Figs S3 and S4. Not surprisingly, given the differences in the training and test set sizes, the average R_p decreased for all models and the average RMSE increased. However, results followed the same trends described in Sections 3.2

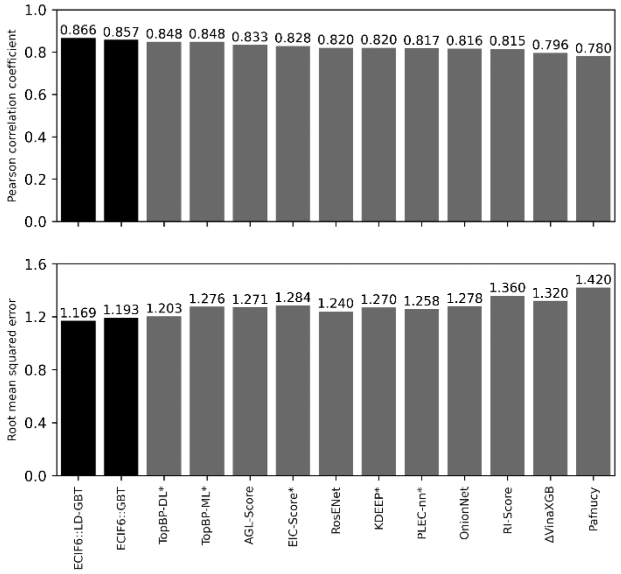


Fig. 4. Performance comparison of different SFs. Results are shown in terms of Pearson correlation coefficient (upper panel) and root mean squared error (lower panel) between predicted and experimental pK on the 'core set' ($n = 285$) from the CASF-2016. The Pearson correlation coefficients of other methods were taken from (Cang *et al.*, 2018; Hassan-Harrirou *et al.*, 2020; Jiménez *et al.*, 2018; Lu et al., 2019; Nguyen and Wei, 2019a,b; Nguyen *et al.*, 2017; Stepniewska-Dziubinska *et al.*, 2018; Wójcikowski *et al.*, 2019; Zheng *et al.*, 2019). SFs marked with * use PDBbind v2016 core set ($n = 290$)

and 3.3, with the models based on ECIF being the best performing ones regardless of the distance cutoff, particularly in combination with ligand descriptors. The only difference was the best distance cutoff, being 8.5 Å for both ECIF and ECIF::LD models, which highlights the importance of tuning this parameter for the construction of models based on ECIF.

3.6 Comparison with different scoring functions

We compared the predictive power of the best SFs derived from ECIF, either alone or in combination with LD, with various state-of-the-art SFs reported in the literature. Specifically, we contrasted the R_p, and RMSE obtained by ECIF6-GBT and ECIF6::LD-GBT to those from recently reported models tested for the scoring power task of the CASF-2016 benchmark. Figure 4 shows the results. Models based on ECIF, either alone or in combination with LD, were the most accurate SFs in this benchmark, achieving the highest R_p and the lowest RMSE of the herein compared models. This comparison highlights the superior descriptive power of ECIF as well as the scoring power of the models derived from them, outperforming models relying on state-of-the-art CNN architectures such as K_{DEEP} (R_p: 0.82, RMSE: 1.27), and TopBP-DL (R_p: 0.848, RMSE: 1.210).

An important aspect to be considered for this comparison is that different machine-learning SFs employed different training set sizes. TopBP-DL, TopBP-ML, AGL-Score, EIC-Score, K_{DEEP} and RI-Score were trained on the 'refined set' minus 'core set' from the PDBbind 2016, RosENet expand this training set with structures from the 'refined set' from the PDBbind 2018, while PLEC-nn, OnionNet, ΔVinaXGB and Pafnucy were trained on larger subsets from the 'general set' from different version of the PDBbind database, including more than 11 000 protein–ligand complexes in these last four cases. In this sense, the effect of the training set size on the performance of all SFs is not straightforward to compare. Although the general trends suggest that increasing the size of the training set has a positive effect on the model performance (Boyles *et al.*, 2020; Lu *et al.*, 2019; Wójcikowski *et al.*, 2019), this cannot be generalized to all models. For instance, authors from K_{DEEP} pointed out that the performance of their CNN model did not improved when using a larger training set (Jiménez *et al.*, 2018). To assess the effect of the training set size on the herein presented models, we trained all the models included in this work on the 3769 protein ligand complexes included in the 'refined set' minus 'core set' from the PDBbind 2016 (2YLC, 3O7U, 3ZNR were excluded as indicated in Section 2.2) and tested them in the 'core set' from CASF-2016. R_p and RMSE between predicted and experimental pK for each model are included in Supplementary Figs S5 and S6. The R_p decreased and the RMSE increased for all models when trained on this smaller set of complexes. However, the trends discussed in Sections 3.2 and 3.3 remained the same, suggesting that a larger training set size has a positive effect for the RF and GBT algorithms regardless of the descriptors employed. According to this, the performance of the models based on ECIF is expected to increase in the future, as more experimental data become available. Regarding our best performing models, R_p went from 0.857 to 0.829 for ECIF6-GBT and from 0.866 to 0.841 for ECIF6::LD-GBT, while RMSE went from 1.193 to 1.278 for ECIF-GBT and from 1.169 to 1.252 for ECIF6::LD-GBT, performances still comparable to those for state-of-the-art SFs.

4 Conclusions

The performance of machine-learning SFs heavily depends on the design and selection of features describing protein–ligand complexes. Herein, we introduced ECIF as a novel description of protein–ligand complexes. Reminiscent of ECFP for small molecules, ECIF are a simple to calculate but detailed set of protein–ligand atom pair counts that consider the connectivity of the involved atoms to define them. In this work, ECIF were defined as a set of counts for 1540 possible protein–ligand atom pairs as delimited by the protein–ligand complexes in the 'refined set' from the PDBBind 2016. However, the underlying idea of ECIF is a general approach that could be easily extended to other types of complexes. The Python code for the computation of ECIF is freely available at https://github.com/DIFACQUIM/ECIF.

We showed the application of ECIF in the construction of machine-learning SFs and evaluated their scoring power performance in the CASF-2016 benchmark with the 'screening power' and 'docking power' tasks yet to be optimized in future research. Using different distance cutoffs for the calculation of the features, we found that models based on ECIF consistently outperformed those derived from ELEMENTS, a related set of descriptors initially proposed for the development of RF-Score. We used RF and GBT as two different machine-learning algorithms to obtain the SFs and demonstrated that GBT performs better than RF for models relying on ECIF. We found 6.0 Å to be the best performing distance cutoff for the construction of SFs based on ECIF, being the only adjustable parameter. Consistently with reports for similar sets of descriptors, we showed that the addition of purely ligand-based descriptors to ECIF for the construction of the SFs significantly improves their performance over the same benchmark. An analysis on the feature importance of an optimized ECIF6::LD-GBT model revealed that non-polar atom pair counts from ECIF are remarkably important for the construction of the models. We found that models built on the combination of ELEMENTS with LD do not outperform those built using ECIF alone. Moreover, we compared the performances of the best performing models based on ECIF, being ECIF6-GBT and ECIF6::LD-GBT, to those from recently reported models evaluated on the same benchmark. To the best of our knowledge, ECIF6-GBT and ECIF6::LD-GBT perform better than any other state-of-the-art machine-learning SF reported to date on the CASF-2016 benchmark, achieving R_p values of 0.857 and 0.866, and RMSE values of 1.193 and 1.169, respectively, highlighting the descriptive power of ECIF. Our results suggest that the use of a more detailed chemical description of atoms for the tallying of protein–ligand atom pairs, such as ECIF, yields to machine-learning SFs with improved predictions of binding affinity, and that considering the increasing availability of experimental data, its performance is expected to increase in the future.

Acknowledgements

N.S.-C. is thankful to *Consejo Nacional de Ciencia y Tecnología* (CONACyT), Mexico, for the granted scholarship number 335997 and to *Becas Iberoamérica Santander de Investigación 2018* for the mobility grant for a research stay in Spain.

Funding

This work was partially supported by the *Dirección General de Cómputo y de Tecnologías de Información y Comunicación* (DGTIC), UNAM that provided computational resources to use Miztli supercomputer with the project LANCAD-UNAM-DGTIC-335.

Conflict of Interest: none declared.

References

Ain,Q.U. *et al.* (2015) Machine-learning scoring functions to improve structure-based binding affinity prediction and virtual screening. *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, **5**, 405–424.

Ballester,P.J. *et al.* (2014) Does a more precise chemical description of protein–ligand complexes lead to more accurate prediction of binding affinity? *J. Chem. Inf. Model.*, **54**, 944–955.

Ballester,P.J. *et al.* (2010) A machine learning approach to predicting protein–-ligand binding affinity with applications to molecular docking. *Bioinformatics*, **26**, 1169–1175.

Boyles,F. *et al.* (2020) Learning from the ligand: using ligand-based features to improve binding affinity prediction. *Bioinformatics*, **36**, 758–764.

Cang,Z. *et al.*. (2018) Representability of algebraic topology for biomolecules in machine learning based scoring and virtual screening. *PLOS Comput. Biol.*, 14, e1005929.

Cang,Z. and Wei,G. (2017) TopologyNet: topology based deep convolutional and multi-task neural networks for biomolecular property predictions. *PLoS Comput. Biol.*, **13**, 1–27.

Cheng,T. *et al.* (2009) Comparative assessment of scoring functions on a diverse test set. *J. Chem. Inf. Model.*, **49**, 1079–1093.

Deng,Z. *et al.* (2004) Structural Interaction Fingerprint (SIFt): a novel method for analyzing three-dimensional protein−ligand binding interactions. *J. Med. Chem.*, **47**, 337–344.

Durrant,J.D. and McCammon,J.A. (2011) NNScore 2.0: a neural-network receptor–ligand scoring function. *J. Chem. Inf. Model.*, **51**, 2897–2903.

Friesner,R.A. *et al.* (2006) Extra precision glide: docking and scoring incorporating a model of hydrophobic enclosure for protein−ligand complexes. *J. Med. Chem.*, **49**, 6177–6196.

Friesner,R.A. *et al.* (2004) Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J. Med. Chem.*, **47**, 1739–1749.

Halgren,T.A. *et al.* (2004) Glide: a new approach for rapid, accurate docking and scoring. 2. Enrichment factors in database screening. *J. Med. Chem.*, **47**, 1750–1759.

Hassan-Harrirou,H. *et al.* (2020) RosENet: improving binding affinity prediction by leveraging molecular mechanics energies with an ensemble of 3D convolutional neural networks. *J. Chem. Inf. Model.*, **60**, 2791–2802.

Jiménez,J. *et al.* (2018) KDEEP: protein–ligand absolute binding affinity prediction via 3D-convolutional neural networks. *J. Chem. Inf. Model.*, **58**, 287–296.

Jones,G. *et al.* (1997) Development and validation of a genetic algorithm for flexible docking 1 1Edited by F. E. Cohen. *J. Mol. Biol.*, **267**, 727–748.

Krizhevsky,A. *et al.*. (2017) ImageNet classification with deep convolutional neural networks. Communications of the ACM, **60**, 84–90.

Kuck,D. *et al.* (2010) Novel and selective DNA methyltransferase inhibitors: docking-based virtual screening and experimental evaluation. *Bioorg. Med. Chem.*, **18**, 822–829.

Leman,J.K. *et al.* (2020) Macromolecular modeling and design in Rosetta: recent methods and frameworks. *Nat. Methods*, **17**, 665–680.

Li,G.B. *et al.* (2013) ID-score: a new empirical scoring function based on a comprehensive set of descriptors related to protein–ligand interactions. *J. Chem. Inf. Model.*, **53**, 592–600.

Li,H. *et al.* (2019) Classical scoring functions for docking are unable to exploit large volumes of structural and interaction data. *Bioinformatics*, **35**, 3989–3995.

Li,H. *et al.* (2015) Low-quality structural and interaction data improves binding affinity prediction via random forest. *Molecules*, **20**, 10947–10962.

Li,H. *et al.* (2020) Machine-learning scoring functions for structure-based drug lead optimization. *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, **10**, 1–20.

Li,H. *et al.* (2018a) The impact of protein structure and sequence similarity on the accuracy of machine-learning scoring functions for binding affinity prediction. *Biomolecules*, **8**, 12.

Li,Y. *et al.* (2018b) Assessing protein–ligand interaction scoring functions with the CASF-2013 benchmark. *Nat. Protoc.*, **13**, 666–680.

Li,Y. *et al.* (2014a) Comparative assessment of scoring functions on an updated benchmark: 1. Compilation of the test set. *J. Chem. Inf. Model.*, **54**, 1700–1716.

Li,Y. *et al.* (2014b) Comparative assessment of scoring functions on an updated benchmark: 2. Evaluation methods and general results. *J. Chem. Inf. Model.*, **54**, 1717–1736.

Liu,J. and Wang,R. (2015) Classification of current scoring functions. *J. Chem. Inf. Model.*, **55**, 475–482.

Liu,Z. *et al.* (2017) Forging the basis for developing protein–ligand interaction scoring functions. *Acc. Chem. Res.*, **50**, 302–309.

Liu,Z. *et al.* (2015) PDB-wide collection of binding data: current status of the PDBbind database. *Bioinformatics*, **31**, 405–412.

Lu,J. *et al.* (2019) Incorporating explicit water molecules and ligand conformation stability in machine-learning scoring functions. *J. Chem. Inf. Model.*, **59**, 4540–4549.

Lyu,J. *et al.* (2019) Ultra-large library docking for discovering new chemotypes. *Nature*, **566**, 224–229.

Nguyen,D.D. *et al.* (2017) Rigidity strengthening: a mechanism for protein–ligand binding. *J. Chem. Inf. Model.*, **57**, 1715–1721.

Nguyen,D.D. and Wei,G.W. (2019a) AGL-score: algebraic graph learning score for protein–ligand binding scoring, ranking, docking, and screening. *J. Chem. Inf. Model.*, **59**, 3291–3304.

Nguyen,D.D. and Wei,G.W. (2019b) DG-GL: differential geometry-based geometric learning of molecular datasets. *Int. J. Numer. Method Biomed. Eng.*, **35**, 1–24.

Pedregosa,F. *et al.* (2011) Scikit-learn: machine learning in {P}ython. *J. Mach. Learn. Res.*, **12**, 2825–2830.

Rogers,D. and Hahn,M. (2010) Extended-connectivity fingerprints. *J. Chem. Inf. Model.*, **50**, 742–754.

Ruiz-Carmona,S. *et al.* (2014) rDock: a fast, versatile and open source program for docking ligands to proteins and nucleic acids. *PLoS Comput. Biol.*, **10**, e1003571.

Stepniewska-Dziubinska,M.M. *et al.* (2018) Development and evaluation of a deep learning model for protein–ligand binding affinity prediction. *Bioinformatics*, **34**, 3666–3674.

Su,M. *et al.* (2019) Comparative assessment of scoring functions: the CASF-2016 update. *J. Chem. Inf. Model.*, **59**, 895–913.

Trott,O. and Olson,A.J. (2009) AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comput. Chem.*, **31**, 455–461.

Wang,C. and Zhang,Y. (2017) Improving scoring-docking-screening powers of protein–ligand scoring functions using random forest. *J. Comput. Chem.*, **38**, 169–177.

Wójcikowski,M. *et al.* (2019) Development of a protein–ligand extended connectivity (PLEC) fingerprint and its application for binding affinity predictions. *Bioinformatics*, **35**, 1334–1341.

Zheng,L. *et al.* (2019) OnionNet: a multiple-layer intermolecular-contact--based convolutional neural network for protein–ligand binding affinity prediction. *ACS Omega*, **4**, 15956–15965.

Zilian,D. and Sotriffer,C.A. (2013) SFCscoreRF: a random forest-based scoring function for improved affinity prediction of protein–ligand complexes. *J. Chem. Inf. Model.*, **53**, 1923–1933.

Recent progress on cheminformatics approaches to epigenetic drug discovery

Zoe Sessions[1,4], Norberto Sánchez-Cruz[2,4], Fernando D. Prieto-Martínez[2], Vinicius M. Alves[1], Hudson P. Santos Jr.[3], Eugene Muratov[1], Alexander Tropsha[1] and José L. Medina-Franco[2]

[1] Laboratory for Molecular Modeling, the UNC Eshelman School of Pharmacy, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA
[2] DIFACQUIM research group, Department of Pharmacy, School of Chemistry, Universidad Nacional Autónoma de México, Avenida Universidad 3000, Mexico City 04510, Mexico
[3] Biobehavioral Laboratory, School of Nursing, the University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA

The ability of epigenetic markers to affect genome function has enabled transformative changes in drug discovery, especially in cancer and other emerging therapeutic areas. Concordant with the introduction of the term 'epi-informatics', the size of the epigenetically relevant chemical space has grown substantially and so did the number of applications of cheminformatic methods to epigenetics. Recent progress in epi-informatics has improved our understanding of the structure–epigenetic activity relationships and boosted the development of models predicting novel epigenetic agents. Herein, we review the advances in computational approaches to drug discovery of small molecules with epigenetic modulation profiles, summarize the current chemogenomics data available for epigenetic targets, and provide a perspective on the greater utility of biomedical knowledge mining as a means to advance the epigenetic drug discovery.

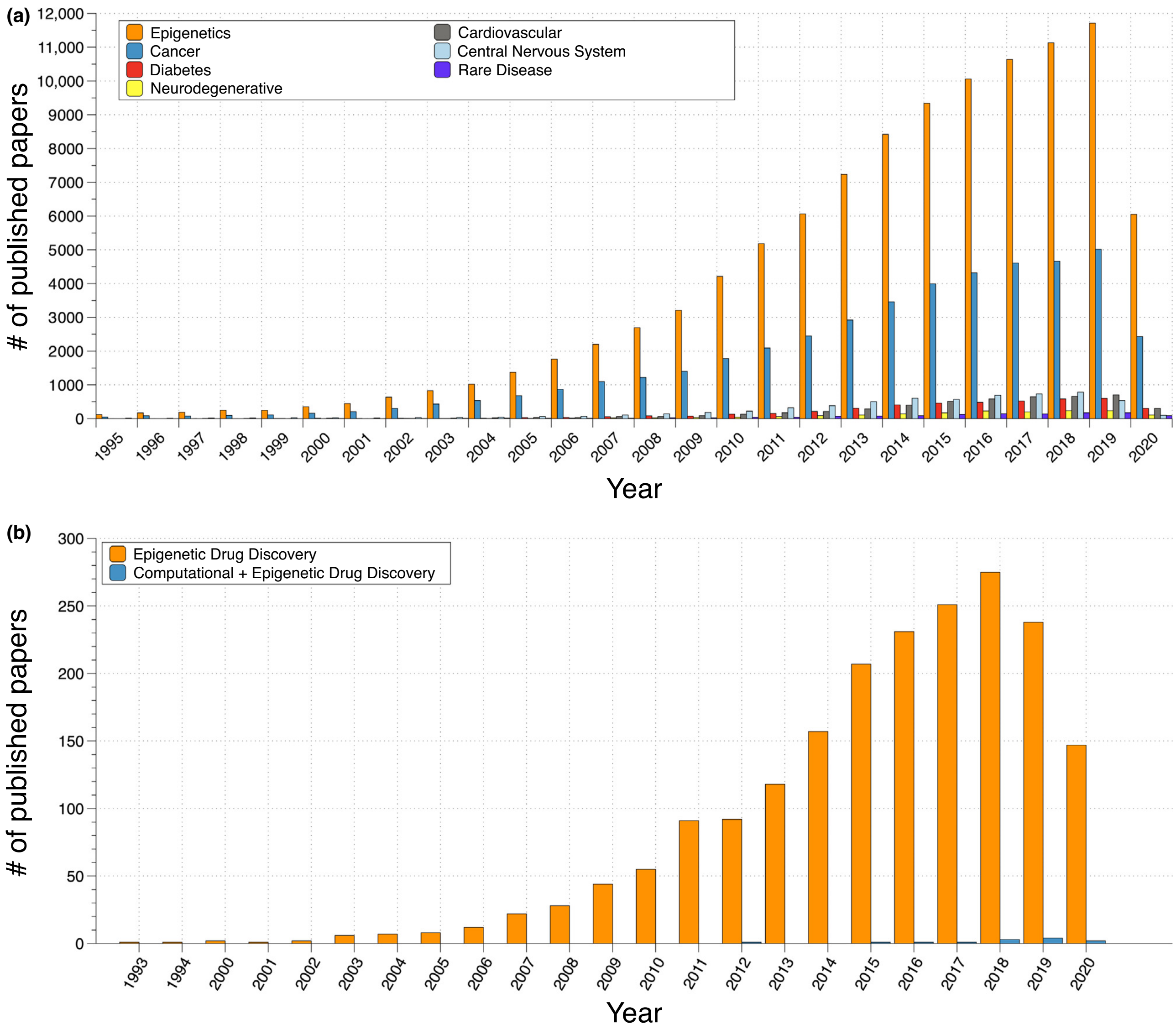
Introduction

The term 'epigenetics' has its historical roots in the work of C. Waddington (1940s) and D.L. Nanney (1950s), 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 [1,2]. However, the term is now used with multiple meanings. For example, epigenetics has been used to describe the (i) heritable phenotype (cellular memory) without modification of DNA sequences [3]; (ii) 'the structural adaptation of chromosomal regions to register, signal, or perpetuate altered activity states' in the genome [4]; and (iii) the mechanism by which the environment conveys its influence to the cell, tissue, or organism [5]. Despite these ambiguous definitions, the growing importance of epigenetics in new therapeutics discovery and understanding disease etiology can be illustrated by the trends in the number of scientific publications on the subject. Fig. 1a shows the number of papers published in the literature with the word 'epigenetics' and six co-occurring diseases per year; now, 25 years after the first six published reports on 'epigenetics' in 1994 [6], there are cumulatively >105 000 relevant publications. Although the most frequent co-occurring term is 'cancer' and most epidrugs are approved for cancer treatment, epigenetic modifications are not limited to malignant conditions. Instead, the data suggest that exceptionally deadly diseases were the first to be studied in the epigenetic field [6].

As one can see from Fig. 1, epigenetics has implications for many different areas of research, as indicated by many extensive reviews on different topics, such as the waves of epigenetic drugs [4], the role of epigenetics in cancer [7], autism diagnosis [8], multiple sclerosis [9], depressive disorders [10], and rare diseases

Corresponding authors: Tropsha, A. (alex_tropsha@unc.edu), Medina-Franco, J.L. (medinajl@unam.mx)
[4] Both authors contributed equally to this study.

1359-6446/© 2020 Elsevier Ltd. All rights reserved.
https://doi.org/10.1016/j.drudis.2020.09.021



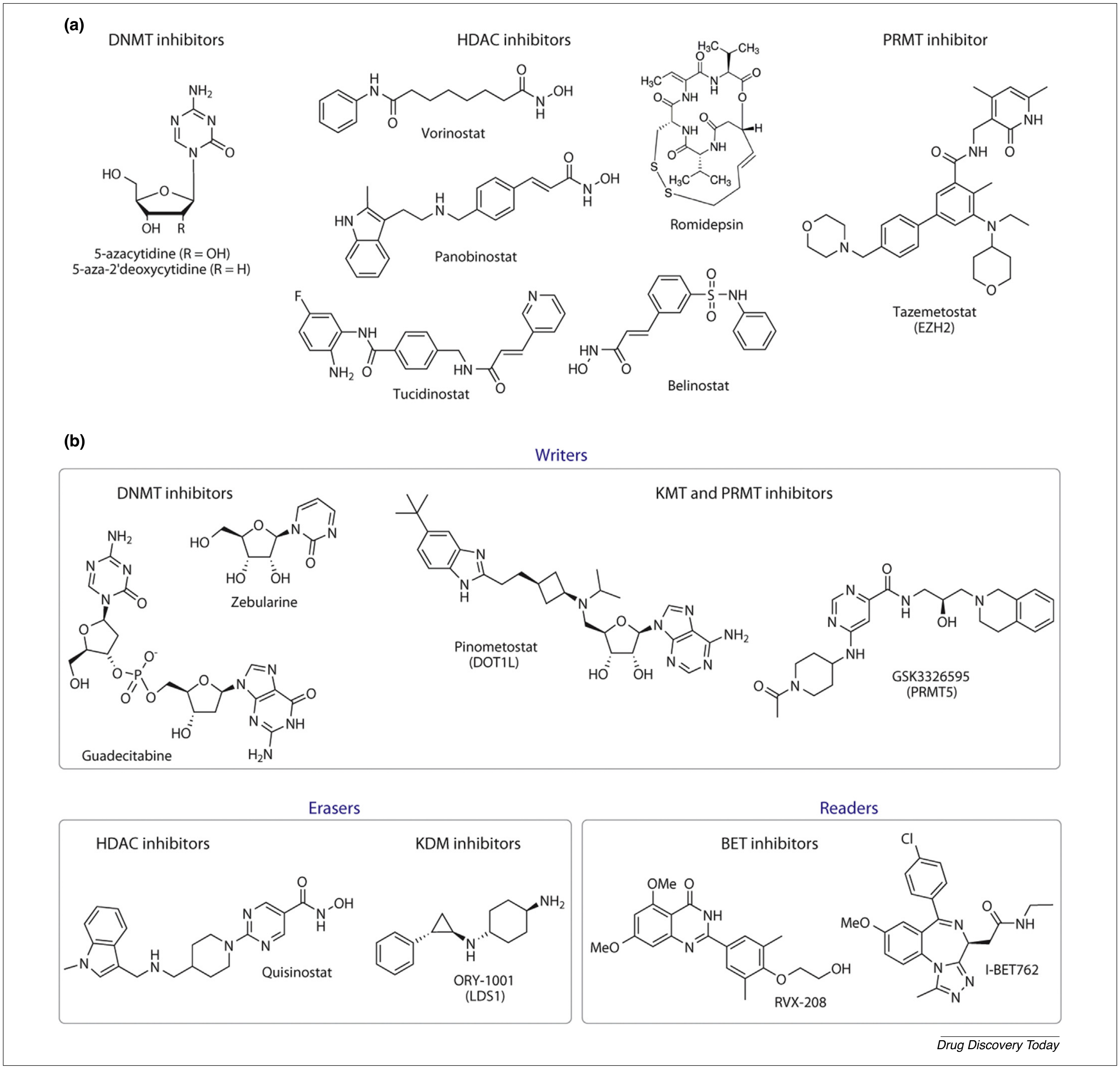
Drug Discovery Today

FIGURE 1

The increasing field of epigenetics. **(a)** The number of papers in PubMed with the keyword 'Epigenetics' alone or co-occurring with multiple diseases (cancer, diabetes, neurodegenerative, cardiovascular, central nervous system, and rare disease) by year. **(b)** The number of papers in PubMed with the keywords 'epigenetics drug discovery' and 'epigenetics + computer-aided drug discovery/computational drug discovery' by year.

[11], to name a few. Although the general trends in drug discovery suggest kinases and G-protein-coupled receptors are the most investigated protein families [12], epigenetic research has predominantly focused on different classes of targets, primarily histone deacetylases (HDACs), histone methyltransferases (HMTs), DNA methyltransferase (DNMT) 1, and other chromodomain and/or tudor domain-containing proteins (TUD). Interest in the bromodomain and external terminal protein (BET) family has shed light on 42 bromodomain (BRD) targets [13], and further investigation of histone acetyltransferases (HATs) led to a distinctive class of lysine acetyltransferases (KATs). Since the report of the first adenine-specific DNMT in *Escherichia coli* [14], the first discovered epigenetic target, 515 human epigenetic targets (136 mappable onto epigenetic phylogenetic groups) have been described, along with respective chemogenomics data (Table S1 in the supplemental information online). Despite the immense growth of the field, several epigenetic proteins associated with important diseases have been largely understudied, such as protein arginine methyltransferases (PRMTs), histone methyl readers (HMRs), and DNMT3; limited funding allocated to the studies of such 'dark' targets could be a contributing factor [15]. Many compounds have been tested in histone demethylase (HDM) assays, and while some of them have reached preclinical and clinical stages, none of them have been approved by the US Food and Drug Administration (FDA) yet. Considering their potential impact on the field, more work needs to be done on dark targets to better understand their mechanisms and expand the number of available targets.

Epigenetics has a major role in the understanding of inheritance, development, and progression of diseases. Consequently, the discovery and development of epidrugs (or epigenetic drugs)



Drug Discovery Today

FIGURE 2

Examples of small molecules relevant to epigenetic drug discovery. **(a)** Chemical structures of epigenetic drugs approved for clinical use. **(b)** Chemical structures of selected compounds in clinical development as epidrugs. Epigenetic targets are grouped as writers, erasers, and readers. A comprehensive review of compounds in clinical development the reader was recently provided by Ganesan *et al*. [4]. For definitions of abbreviations, please see the main text.

have become a major research focus for multiple biotechnology companies [4]. As of August 2020, there were eight epigenetic drugs approved for clinical use [4,16]: two DNMT inhibitors (azacytidine and decitabine), five HDAC inhibitors (vorinostat, belinostat, panobinostat, romidepsin, and tucidinostat), and one HMT inhibitor (tazemetostat). Fig. 2a and b shows the chemical structures of representative FDA-approved drugs and compounds in clinical development as epidrugs, respectively [4]. An ongoing clinical trial is examining the FDA-approved DNMT inhibitor, azacitidine, in combination with a novel BET inhibitor, INCB057643, and a novel lysine dependent demethylase (LSD1) inhibitor, INCB059872, for solid metastatic tumors [17]. An extensive description of ongoing clinical trials related to epigenetic drugs has been compiled recently [4]. Currently, most drugs focus on DNMT1 and HDACs (Fig. 2a). The increased interest in epigenetic applications outside of cancer has not yet translated to a rise in viable drugs or drug candidates, despite the expansive therapeutic benefits of epigenetic modulation.

TABLE 1

Major public databases relevant to epigenetic drug discovery

Database	Last update	Number of targets	Number of compounds	URL[a]	Refs
Human epigenetic enzyme and modulator database (HEMD)	2012	269	4377	http://mdl.shsmu.edu.cn/HEMD/	[22]
EpiDBase	2015	Not available	5784	www.epidbase.org/	[23]
EpiFactors	2015	815	–	http://epifactors.autosome.ru	[24]
Database of Epigenetic Modifiers (dbEM)	2016	167	–	http://crdd.osdd.net/raghava/dbem	[25]
Epigenomics chemical database	2018	54	7820	www.difacquim.com/d-databases/	[18]

[a] Last accessed: 25 August 2020.

In recent years, the substantial growth in epigenetics-related data has prompted the development of cheminformatics methods with application to this field. In 2015, the term 'epi-informatics' was introduced and conceptualized [6] to summarize advances in epigenetic drug and chemical probe discovery driven by computational methods. Since then, molecular modeling and cheminformatics approaches have made substantial contributions to the field (Fig. 1b) [18–20]. Computational methods have helped to explore the mechanism of action of active compounds at the molecular level and guide lead optimization programs. Herein, we review recent advances in computational approaches to epigenetic drug discovery and summarize, to the best of our knowledge, all the publicly available chemogenomics data for epigenetic targets.

Chemogenomics data and databases

Over the past decade, several open-source databases have compiled targets and compounds with epigenetic profiles. However, most do not use the same criteria for target selection/classification or use a systematic drug target ontology [21]. Table 1 summarizes the chemogenomics databases related to epigenetics published thus far, including web links to each resource. Although all of them were accessible as of August 2020, none have been updated since their release. This fact highlights the need for tools that automate the analysis of the constantly increasing body of epigenetic data in the public domain.

One of the first web-accessible databases was the Human Epigenetic Enzyme and Modulator Database (HEMD). This resource, published in 2012, includes 4377 small-molecule modulators and 269 epigenetic targets that are annotated with information on epigenetic mechanisms, catalytic processes, and related diseases [22]. EpiDBase is also a web-accessible database that was released in 2015 [23]. It contains 5784 different ligands annotated with experimental activity (IC_{50}) against writers, erasers, and readers, as well as calculated properties.

EpiFactors [24], also published in 2015, is a manually curated database that provides information on epigenetic regulators, their complexes, targets, and products. The database contains 815 human epigenetic proteins and 69 protein complexes involved in epigenetic regulation. EpiFactors also provides the corresponding genes and their expression levels in 458 human primary cell samples, 255 different cancer cell lines, and 134 human postmortem tissues.

The Database of Epigenetic Modifiers (dbEM) is a web-accessible database released in 2016 [25] that contains the genomic information on 167 epigenetic targets. This resource maintains the information of mutations, copy number variation, and gene expression in tumor samples, cancer cell lines, and healthy samples.

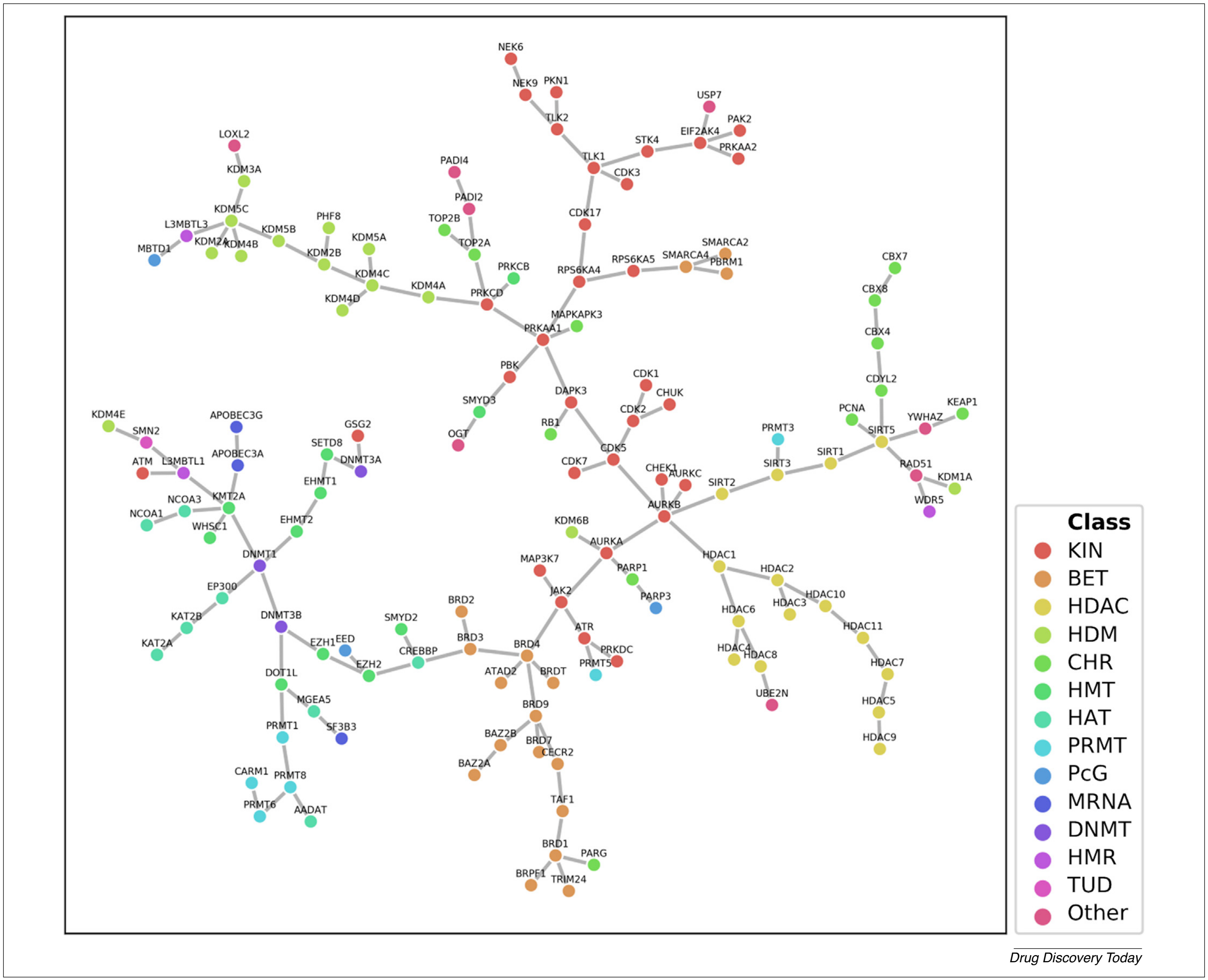
In 2018, a database of epigenetic small molecules inhibitors was released [18] integrated from other public access databases such as ChEMBL (www.ebi.ac.uk/chembl/) and PubChem (https://pubchem.ncbi.nlm.nih.gov/). The epigenomics database includes 7820 unique compounds, of which 3456 have information for more than one epigenetic target. The database has 16,102 compound–target associations, of which 15,887 have quantitative potency data associated with them. The database has associations with 60 epigenetic targets.

Current approaches, models, and best achievements

Epigenetic-relevant chemical space

According to the Chemical Space project, the total number of synthetically feasible organic molecules exceeds 166 billion compounds [26]. Through chemical clustering and visualization, cheminformatics has allowed the navigation of large databases to identify therapeutically relevant chemical spaces. In this context, the 'epigenetic-relevant chemical space' (ERCS) was the first attempt to comprise a list of HDACs, DNMTs, and BET inhibitors (2772 compounds in total) [27]. The design of selective inhibitors remains a challenge in epigenetic drug discovery. For instance, most HDAC inhibitors approved so far are non-isoform selective, even though it has been hypothesized that selective inhibitors might have fewer adverse effects [3,28]. A comprehensive characterization of the ERCS can uncover multitarget relationships as well as guide the design of selective inhibitors, which could significantly improve the effectiveness of drug therapies [29].

We surveyed the status of the compounds tested against the epigenetic targets and currently available in the public domain. Fig. 3 visualizes the current ERCS using a Tree Manifold Approximation and Projection (TMAP) [30], generated with Statistical-Based Database Fingerprints (SB-DFPs) to represent target-associated compound datasets [31]. The SB-DFPs are novel condensed representations of compound data sets that have also been shown to capture these relationships between the compound data sets. This representation is based on statistical comparisons of molecular fingerprint bit proportions among any two data sets. These condensed representations are connected by a minimum spanning tree, as described in detail in the original publication [30]. The TMAP displays the relationships among compounds with biological data for a particular target as nodes that are connected to similar data sets of other targets through branches and subbranches. This map could also help to guide the design of multitarget epigenetic agents. In the map, several targets that are in the same or neighboring branches share the same role in histone modifications, such as kinases (KINs; writers), the BET family (readers), HDACs and HDMs (erasers), as well as chromatin remo-



Drug Discovery Today

FIGURE 3

Visual representation of the current epigenetic-relevant chemical space. The visualization was generated with the recent Tree Manifold Approximation and Projection method and Statistical-Based Database Fingerprints as condensed representations of compounds associated to the epigenetic targets [31]. Individual nodes of the tree are epigenetic targets, colored by their class. The nodes are clustered according to the pairwise similarity of their condensed representations. The branches represent the connection between the chemical space of two different targets. Proximal data sets can contribute to the discovery of compounds acting on multiple targets. For definitions of abbreviations, please see the main text.

delers (CHR). This observation is consistent with the known difficulty of designing a selective inhibitor within each class [4]. The map also shows a close relation between targets related to DNA methylation (i.e., DNMTs) to those involved in RNA modification (mRNAs) as well as different histone writers (HMTs, PRMTs, and HATs). Targets associated with other histone modifications tend to appear in the terminal regions of these branches, illustrating a weak relationship between different target classes. By contrast, Polycomb-group proteins (PcG) and other targets, such as the chromatin remodeler PARG2 and methyltransferases PRMT3 and PRMT5, do not appear to be closely related to targets with the same function, suggesting them as promising targets for the development of selective compounds.

Structure-epigenetic activity relationship analyses

The increasing number of reports on compounds with experimental activity against one or more epigenetic targets has allowed the exploration of the structure–activity landscape for several epigenetic target datasets. Here, we review the respective models developed and published in recent years.

Quantitative structure-activity relationships (QSAR) modeling

QSAR modeling is a major computational approach to drug discovery [32] and, with the growth of chemogenomics datasets for epigenetic targets, its application to find bioactive compounds with epigenetics profile has grown. A recent review discussed 3D-QSAR studies performed with several different sets of HDACs inhibitors [33]. More recently, models were reported for 11 sets of inhibitors of HDACs using conformal prediction [34], showing that these models demonstrated high accuracy for both training and external test sets. The efficiencies for the predictions were >80% for most data sets and >90% for four data sets at different significance levels. Another paper expanded the discussion on HDAC-1 inhibitors, specifically, aminophenyl benzamide derivatives, and was able to produce an externally validated model with a high correlation (R^2) and predictive ($r^2_{ext\_ts}$) values of 0.96 and 0.79, respectively [35]. García-Sánchez *et al.* [36] reviewed more recent examples of QSAR models developed, again, mostly for HDACs and DNMTs. In an interesting effort to obtain novel DNMT1 inhibitors, the authors modified the structures of known inhibitors and used QSAR models to predict the activity of the novel compounds. They also identified electronegativity and the bond information content index as the most influential descriptors [37]. Unfortunately, other epigenetic targets have been minimally explored. BET inhibitors have been studied using QSAR modeling approaches; one study used such models to predict six putative multitarget BET bromodomain inhibitors [38]. Six reversible LSD1 inhibitors were proposed based on 3D-QSAR, molecular docking, and dynamics studies, but these predictions have not been supported by the experimental binding affinity measurements [39]. Similar approaches have been applied to lysine methyltransferase DOTL1 inhibitors, resulting in the computer-assisted design of two compounds that demonstrated inhibition at micromolar levels in confirmatory assays [40].

Activity landscape modeling

Activity landscapes can be defined as representations that compare compound similarity and activity relationships [41]. One of the main goals of this approach is to identify activity cliffs that can help lead optimization efforts [42] as well as data sets with a smooth (continuous) SAR that would be more likely to yield predictive models. Similarly, one can identify epigenetic datasets with a ‘rough’ (i.e., discontinuous) SAR landscape, for which the development of predictive models is expected to be difficult. Based on this concept, the SEAR of various epigenetic target data sets has been analyzed, including DNMTs [43,44], BRDs [36], HDACs [45], and lysine methyltransferases (G9a or EHMT2) [46]. To characterize the epigenetic activity landscape, a SAR analysis of 52 compounds tested against different epigenetic targets was used [47], which showed DNMT3B and DOT1L to have the highest percentage of activity cliffs (i.e., the respective data sets had more discontinuous SAR). The study also revealed that HDACs were the targets with the most continuous SAR, making them most suitable to carry out hit-to-lead optimization. This result is also consistent with the generation of a large number of predictive QSAR models. Finally, this large-scale epigenetic activity landscape study highlighted SMARCA2 and HAT as the epigenetic targets more prone to scaffold hopping (i.e., the search for compounds with different scaffolds but similar activity).

Epigenetic target profiler

Recently, a free online service was developed to predict the potential activity of small molecules against epigenetic targets. Briefly, the Epigenetic Target Profiler is a user-friendly web application that uses binary classification models relying on machine-learning algorithms (support vector machines and artificial neural networks) to predict the most probable epigenetic targets for a small molecule. Classification models were built on the available bioactivity data for 35 epigenetic regulators from ChEMBL26; the development of a new version using the latest release of ChEMBL is in progress. Figure S1 in the supplemental information online shows the graphical user interface of the webserver.

Reshaping the discovery and development of epidrug candidates with current and emerging technologies in cheminformatics

Molecular modeling and cheminformatics have made notable contributions to drug discovery [48]. Several important drugs have been developed with computational methods, including imatinib (a kinase inhibitor used to treat certain types of cancer), dorzolamide (a carboanydrase II inhibitor used to treat high pressure inside the eye, including glaucoma), enfuvirtide (a first-in-class antiretroviral drug used in combination therapy for the treatment of HIV-1), oseltamivir (a neuraminidase inhibitor used to treat flu), and many others [49,50]. The increasing chemogenomic body of data in the field of epigenetics has facilitated the application of several well-established computational methodologies. One example is the development of the Epigenetic Target Profiler discussed in the previous section. However, the application of computational approaches in epigenetic drug discovery is still recent and, so far, no epi-drugs have been discovered with computational strategies. Therefore, the development and application of new epi-informatics methodologies should continue to be an actively growing area of research.

Over the past few years, several groups have used multiple computational approaches to epigenetics research [18–20]. A recent review [51] summarized the advantages and limitations of several structure-based approaches in multitarget drug design, including examples from epigenetics. For example, Kuang *et al.* [52] reported on a comprehensive modeling study of several BRD isoforms with molecular dynamics simulations; the authors exploited the non-negligible ligand-binding kinetics features of these proteins, enhancing the understanding of the binding site of the BRD family. Although these techniques are promising, some limitations associated with molecular simulations [53] need to be overcome to enable their broader application to epigenetic research. For instance, an improvement in parametrization protocols for both the simulation of drug-like molecules [54] and protein–protein interfaces [55] is necessary.

Computational approaches can be used to find hidden allosteric binding sites [56] and protein–protein interaction hotspots [57] for epigenetic targets. Computational approaches can also help improve both the pharmacodynamics and pharmacokinetics of compounds in the hit-to-lead stage. Recently, Letfus *et al.* [58] used quantum/molecular mechanics (QM/MM) approaches to improve toxoflavin-based inhibitors of the human histone lysine demethylase enzymes of the subfamily KDM4C (a molecular target related to prostate and breast cancer). These studies led to compounds with activity in the low nM range in biochemical assays and mM activity in cell-based assays that also had enhanced pharmacokinetics properties in vitro. QM/MM methods were used because conventional force fields used in docking programs cannot model most metalloenzymes and metal–ligand interactions properly.

Epigenetic drug discovery can mostly benefit from cheminformatics to identify and prioritize hits for experimental validation. Recently, Tao *et al*. [59] reported the identification of a novel small molecule with in vitro enzymatic inhibition at a submicromolar level for PRMT5, an anticancer therapeutic target, using molecular docking and pharmacophore-based virtual screening. In another study, Song *et al*. [60] used a combination of pharmacophore searches, molecular docking, and molecular dynamic simulations and found compounds with selective activity against the proliferation of cancer cells.

We expect that QSAR modeling will continue to be an important methodology to accelerate the discovery of epidrugs. This approach has been substantially enriched over the past few years with modern machine learning and artificial intelligence algorithms [32]. Although some studies have reported the development of QSAR models and the identification of virtual hits for epigenetic targets [37,38], these hits were not experimentally tested. However, compounds have been identified computationally that were later validated experimentally [61–66], confirming how this approach could help with identifying promising bioactive compounds for epigenetic targets. We have summarized available libraries that have been designed for epigenetics targeted using computational approaches, and a comprehensive list of these libraries is available in Table S2 in the supplemental information online. Meanwhile, structural diversity and scaffold content of such libraries, as well as their in silico pharmacokinetic and toxicological profiles, still need to be estimated [67].

Molecular docking and molecular dynamics simulations as well as artificial intelligence methods have gained attention because of increased computational capabilities and efficient scaling with the advent of graphics processing units. For instance, Lyu *et al.* [68] recently ran molecular docking on a 170-million compound library leading to the experimental discovery of a new scaffold of phenolate inhibitors of AmpC. After optimization, one of the compounds showed inhibitory activity of 77 nM. The authors also identified 30 compounds with submicromolar activity for the D_4 dopamine receptor. Gentile *et al.* [69] developed Deep Docking, a hybrid platform built on deep learning QSAR models trained on docking scores, which afforded rapid accurate prediction of docking scores for billions of molecular structures. Several recent applications of deep learning and generative neural network models have been described in the literature. For instance, a deep reinforcement learning algorithm, termed ReLEASE (Reinforcement Learning for Structural Evolution) was developed that can help to design chemical libraries with a bias toward desired inhibitory activity [70]. By using a similar approach, Zhavoronkov *et al.* [71] reported the discovery of a potent candidate for DDR1, a kinase target implicated in fibrosis and other diseases, noting that the entire project was accomplished in 21 days. Although this work received criticism because the new molecule was similar to an approved drug [72], the new compound was *de facto* not available in any open database. The application of these advanced and accelerated computational approaches to epigenetic targets is pending.

The potential of polypharmacological profiles allowed by epigenetic targets is an important consideration of direct relevance to epigenetic drug discovery. Very often, drugs hitting undesired targets are the primary source of toxicity, also known as 'off-target' effects [73]. Indeed, toxicity and lack of efficacy are the major causes of drug attrition [74]. However, recent studies highlighted the advantages of multitarget design (MTD), taking advantage of the 'selective synergism' [75]. Structure-based and machine-learning methods have been successfully applied in MTDs [76]. Epigenetic modulation is suitable for MTD because several targets are involved in the same pathway, which could lead to synergic results [18,77]. For the purposes of epigenetic drug discovery, promising MTD strategies should focus on designing compounds that inhibit different aspects (reading, writing, and erasing) of the same gene modification, such as (i) reader/writer (e.g., BET/HDAC inhibition) [78]; (ii) writer/writer of distinct changes (e.g., HDAC/DNMT) [79]; and (iii) writer/writer of the same modification (e.g., G9a/DNMT) [46,80]. Therefore, it is evident that epigenetic polypharmacology, or even the combination therapy of different epidrugs, offer a promising avenue for future epigenetic drug discovery [16,18].

Concluding remarks

Over the past 25 years, the number of scientific publications related to epigenetics has increased impressively from six papers in 1994 to >100,000 overall. Although cancer remains the primary therapeutic area associated with epigenetic drug discovery, other conditions, such as cardiovascular, neurodegenerative, central nervous system-related diseases, diabetes, and rare diseases, have been rapidly gaining interest in relation to epigenetics. In addition, we have seen a slight uptick in computational approaches used in this field. So far, epi-informatics has allowed the creation and maintenance of target-compound databases, exploring the increasing ERCS and SEARs, which eventually led to the development of the Epigenetic Target Profiler, a webserver to generate a predicted profile of potential inhibition of small molecules across a panel of epigenetic targets of pharmaceutical relevance.

Herein, we have highlighted recent advances in the use of cheminformatics in epigenetic drug discovery as well as discussed recent cutting-edge computer-aided drug design technologies that could be used to enable breakthrough discoveries in the field. Although no epigenetic drug has been discovered by computational approaches yet, the diversity of therapeutic areas associated with epigenetic targets, the growth of epigenetic chemical space, and the proliferation of robust computational approaches to drug discovery promise a substantial increase in the number of publications and the emergence of novel epigenetic drug candidates discovered by cheminformatics methods.

Declaration of Competing Interest

The authors report no declarations of interest.

Acknowledgments

We thank Consejo Nacional de Ciencia y Tecnología (CONACyT), Mexico (grant 282785), and the National Institutes of Health, USA (grant R01GM114015). F.D.P-M. and N.S-C. are thankful to CONACyT for the PhD scholarships number, 660465/576637 and 335997, respectively.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at https://doi.org/10.1016/j.drudis.2020.09.021.

References

1 Waddington, C.H. (2012) The epigenotype, endeavor, 1942, vol. 1 (pg. 18-20) reprinted. *Int. J. Epidemiol.* 41, 10–13

2 Greally, J.M. (2018) A user's guide to the ambiguous word 'epigenetics'. *Nat. Rev. Mol. Cell Biol.* 19, 207–208

3 Wu, C. and Morris, J.R. (2001) Genes, genetics, and epigenetics: a correspondence. *Science* 293, 1103–1105

4 Ganesan, A. *et al.* (2019) The timeline of epigenetic drug discovery: from reality to dreams. *Clin. Epigenet.* 11, 174

5 Cavalli, G. and Heard, E. (2019) Advances in epigenetics link genetics to the environment and disease. *Nature* 571, 489–499

6 Dueñas-González, A. *et al.* (2016) Introduction of epigenetic targets in drug discovery and current status of epi-drugs and epi-probes. In *Epi-informatics* (Medina-Franco, J.L., ed.), pp. 1–20, Academic Press

7 Esteller, M. (2008) Molecular origins of cancer: epigenetics in cancer. *N. Engl. J. Med.* 358, 1148–1159

8 Waye, M.M.Y. and Cheng, H.Y. (2018) Genetics and epigenetics of autism: a review. *Psychiatry Clin. Neurosci.* 72, 228–244

9 Küçükali, C.İ. *et al.* (2015) Epigenetics of multiple sclerosis: an updated review. *Neuro. Mol. Med.* 17, 83–96

10 Januar, V. *et al.* (2015) Epigenetics and depressive disorders: a review of current progress and future directions. *Int. J. Epidemiol.* 44, 1364–1387

11 Brindisi, M. *et al.* (2020) Old but gold: tracking the new guise of histone deacetylase 6 (HDAC6) enzyme as a biomarker and therapeutic target in rare diseases. *J. Med. Chem.* 63, 23–39

12 Zdrazil, B. *et al.* (2019) Moving targets: monitoring target trends in drug discovery by mapping targets, go terms, and diseases. *bioRxiv* 691550

13 Fujisawa, T. and Filippakopoulos, P. (2017) Functions of bromodomain-containing proteins and their roles in homeostasis and cancer. *Nat. Rev. Mol. Cell Biol.* 18, 246–262

14 Kühnlein, U. *et al.* (1969) Host specificity of DNA produced by *Escherichia coli*. XI. In vitro modification of phage fd replicative form. *Proc. Natl. Acad. Sci. U. S. A.* 63, 556–562

15 Oprea, T.I. *et al.* (2018) Unexplored therapeutic opportunities in the human genome. *Nat. Rev. Drug Discov.* 17, 317–332

16 de Lera, A.R. and Ganesan, A. (2020) Two-hit wonders: the expanding universe of multitargeting epigenetic agents. *Curr. Opin. Chem. Biol.* 57, 135–154

17 Incyte, Cor. (2016) *Azacitidine Combined With Pembrolizumab and Epacadostat in Subjects With Advanced Solid Tumors (ECHO-206).* US National Library of Medicine

18 Naveja, J.J. and Medina-Franco, J.L. (2018) Insights from pharmacological similarity of epigenetic targets in epipolypharmacology. *Drug Discov. Today* 23, 141–150

19 Lim, S.J. *et al.* (2010) Computational epigenetics: the new scientific paradigm. *Bioinformation* 4, 331–337

20 Lu, W. *et al.* (2018) Computer-aided drug design in epigenetics. *Front. Chem.* 6, 57

21 Lin, Y. *et al.* (2017) Drug target ontology to classify and integrate drug discovery data. *J. Biomed. Sem.* 8, 50

22 Huang, Z. *et al.* (2012) HEMD: an integrated tool of human epigenetic enzymes and chemical modulators for therapeutics. *PLoS ONE* 7, e39917

23 Loharch, S. *et al.* (2015) Epidbase: a manually curated database for small molecule modulators of epigenetic landscape. *Database* 2015, bav013

24 Medvedeva, Y.A. *et al.* (2015) Epifactors: a comprehensive database of human epigenetic factors and complexes. *Database 2015* bav067

25 Singh Nanda, J. *et al.* (2016) dbEM: a database of epigenetic modifiers curated from cancerous and normal genomes. *Sci. Rep.* 6, 19340

26 Reymond, J.-L. (2015) The chemical space project. *Acc. Chem. Res.* 48, 722–730

27 Prieto-Martinez, F.D. *et al.* (2016) A chemical space odyssey of inhibitors of histone deacetylases and bromodomains. *RSC Adv.* 6, 56225–56239

28 Chang, P. *et al.* (2018) Histone deacetylase inhibitors: Isoform selectivity improves survival in a hemorrhagic shock model. *J. Trauma Acute Care Surg.* 84, 795–801

29 Loharch, S. and Parkesh, R. (2019) Epigenetic drug discovery: systematic assessment of chemical space. *Fut. Med. Chem.* 11, 2803–2819

30 Probst, D. and Reymond, J.-L. (2020) Visualization of very large high-dimensional data sets as minimum spanning trees. *J. Cheminf.* 12, 12

31 Sánchez-Cruz, N. and Medina-Franco, J.L. (2018) Statistical-based database fingerprint: chemical space dependent representation of compound databases. *J. Cheminf.* 10, 55

32 Muratov, E.N. *et al.* (2020) QSAR without borders. *Chem. Soc. Rev.* 49, 3525–3564

33 Aguayo-Ortiz, R. and Fernández-de Gortari, E. (2016) Overview of computer-aided drug design for epigenetic targets. In *Epi-informatics* (Medina-Franco, J.L., ed.), pp. 21–52, Academic Press

34 Norinder, U. *et al.* (2019) Conformal prediction of HDAC inhibitors. *SAR QSAR Environ. Res.* 30, 265–277

35 Sirous, H. *et al.* (2020) Computer-driven development of an in silico tool for finding selective histone deacetylase 1 inhibitors. *Molecules* 25, 1952

36 García-Sánchez, M.O. *et al.* (2017) Quantitative structure–epigenetic activity relationships. *Adv. QSAR Model.* 24, 303–338

37 Phanus-Umporn, C. *et al.* (2020) QSAR-driven rational design of novel DNA methyltransferase 1 inhibitors. *EXCLI J.* 19, 458–475

38 Speck-Planche, A. and Scotti, M.T. (2019) BET bromodomain inhibitors: fragment-based in silico design using multi-target QSAR models. *Mol. Diver.* 23, 555–572

39 Zhang, X. *et al.* (2020) Molecular docking, 3D-QSAR, and molecular dynamics simulations of thieno[3,2-b]pyrrole derivatives against anticancer targets of KDM1A/LSD1. *J. Biomol. Struct. Dyn.* . http://dx.doi.org/10.1080/07391102.2020.1726819

40 Sabatino, M. *et al.* (2018) Disruptor of telomeric silencing 1-like (DOT1L): disclosing a new class of non-nucleoside inhibitors by means of ligand-based and structure-based approaches. *J. Comp. Aided Mol. Des.* 32, 435–458

41 Iqbal, J. *et al.* (2020) Activity landscape image analysis using convolutional neural networks. *J. Cheminf.* 12, 34

42 Hu, H. *et al.* (2019) Systematic identification of target set-dependent activity cliffs. *Future Sci. OA* 5, FSO363

43 Naveja, J.J. and Medina-Franco, J.L. (2015) Activity landscape of DNA methyltransferase inhibitors bridges chemoinformatics with epigenetic drug discovery. *Expert Opin. Drug Discov.* 10, 1059–1070

44 Naveja, J.J. and Medina-Franco, J.L. (2015) Activity landscape sweeping: into the mechanism of inhibition and optimization of dnmt1 inhibitors. *RSC Adv.* 5, 63882–63895

45 Saldívar-González, F.I. *et al.* (2017) Getting SMARt in drug discovery: chemoinformatics approaches for mining structure–multiple activity relationships. *RSC Adv.* 7, 632–641

46 López-López, E. *et al.* (2020) Towards the understanding of the activity of G9a inhibitors: an activity landscape and molecular modeling approach. *J. Comp. Aided Mol. Des.* 34, 659–669

47 Naveja, J.J. *et al.* (2018) Computational methods for epigenetic drug discovery: a focus on activity landscape modeling. *Adv. Protein Chem. Struct. Biol.* 113, 65–83

48 Martinez-Mayorga, K. *et al.* (2020) The impact of chemoinformatics on drug discovery in the pharmaceutical industry. *Expert Opin. Drug Discov.* 15, 293–306

49 Talele, T.T. *et al.* (2010) Successful applications of computer aided drug discovery: moving drugs from concept to the clinic. *Curr. Top. Med. Chem.* 10, 127–141

50 Leelananda, S.P. and Lindert, S. (2016) Computational methods in drug discovery. *Beilstein J. Org. Chem.* 12, 2694–2718

51 Sivakumar, K.C. *et al.* (2020) Prospects of multitarget drug designing strategies by linking molecular docking and molecular dynamics to explore the protein–ligand recognition process. *Drug Dev. Res.* 81, 685–699

52 Kuang, M. *et al.* (2015) Binding kinetics versus affinities in BRD4 inhibition. *J. Chem. Inf. Model.* 55, 1926–1935

53 Chodera, J.D. *et al.* (2011) Alchemical free energy methods for drug discovery: progress and challenges. *Curr. Opin. Struct. Biol.* 21, 150–160

54 Zanette, C. *et al.* (2019) Toward learned chemical perception of force field typing rules. *J. Chem. Theory. Comput.* 15, 402–423

55 Best, R.B. (2019) Atomistic force fields for proteins. *Methods Mol. Biol.* 2022, 3–19

56 Lu, S. *et al.* (2018) Discovery of hidden allosteric sites as novel targets for allosteric drug design. *Drug Discov. Today* 23, 359–365

57 Yang, C.Y. and Wang, S. (2010) Computational analysis of protein hotspots. *ACS Med. Chem. Lett.* 1, 125–129

58 Letfus, V. *et al.* (2020) Rational design, synthesis and biological profiling of new KDM4C inhibitors. *Bioorg. Med. Chem.* 28, 115128

59 Tao, H. *et al.* (2019) Discovery of novel PRMT5 inhibitors by virtual screening and biological evaluations. *Chem. Pharm. Bull.* 67, 382–388

60 Song, Q. *et al.* (2019) An improved protocol for the virtual screening discovery of novel histone deacetylase inhibitors. *Chem. Pharm. Bull.* 67, 1076–1081

61 Alves, V. *et al.* (2020) QSAR modeling of SARS-CoV Mpro inhibitors identifies sufugolix, cenicriviroc, proglumetacin and other drugs as candidates for repurposing against SARS-CoV–2. *Mol. Inf.* . http://dx.doi.org/10.1002/minf.202000113 Published online July 28, 2020

62 Capuzzi, S.J. *et al.* (2018) Computer-aided discovery and characterization of novel ebola virus inhibitors. *J. Med. Chem.* 61, 3582–3594

63 Hajjo, R. *et al.* (2010) Development, validation, and use of quantitative structure–activity relationship models of 5-hydroxytryptamine (2B) receptor ligands to

identify novel receptor binders and putative valvulopathic compounds among common drugs. *J. Med. Chem.* 53, 7573–7586

64 Gozalbes, R. *et al.* (2005) QSAR strategy and experimental validation for the development of a GPCR focused library. *QSAR Comb. Sci.* 24, 508–516

65 Lima, M.N.N. *et al.* (2018) QSAR-driven design and discovery of novel compounds with antiplasmodial and transmission blocking activities. *Front. Pharmacol.* 9, 146

66 Ballante, F. *et al.* (2017) Structural insights of SmKDAC8 inhibitors: targeting schistosoma epigenetics through a combined structure-based 3D QSAR, *in vitro* and synthesis strategy. *Bioorg. Med. Chem.* 25, 2105–2132

67 Durán-Iturbide, N.A. *et al.* (2020) In silico ADME/Tox profiling of natural products: a focus on BIOFACQUIM. *ACS Omega* 5, 16076–16084

68 Lyu, J. *et al.* (2019) Ultra-large library docking for discovering new chemotypes. *Nature* 566, 224–229

69 Gentile, F. *et al.* (2020) Deep docking: a deep learning platform for augmentation of structure based drug discovery. *ACS Cent. Sci.* 6, 939–949

70 Popova, M. *et al.* (2018) Deep reinforcement learning for *de novo* drug design. *Sci. Adv.* 4, eaap7885

71 Zhavoronkov, A. *et al.* (2019) Deep learning enables rapid identification of potent DDR1 kinase inhibitors. *Nat. Biotechnol.* 37, 1038–1040

72 Walters, W.P. and Murcko, M. (2020) Assessing the impact of generative ai on medicinal chemistry. *Nat. Biotechnol.* 38, 143–145

73 Bantscheff, M. *et al.* (2009) Revealing promiscuous drug–target interactions by chemical proteomics. *Drug Discov. Today* 14, 1021–1029

74 Van Norman, G.A. (2019) Phase II trials in drug development and adaptive trial design. *JACC Basic Transl. Sci.* 4, 428–437

75 Peters, J.U. (2013) Polypharmacology – foe or friend? *J. Med. Chem.* 56, 8955–8971

76 Zhang, W. *et al.* (2017) Computational multitarget drug design. *J Chem Inf Model* 57, 403–412

77 Bechter, O. and Schöffski, P. (2020) Make your best BET: the emerging role of BET inhibitor treatment in malignant tumors. *Pharmacol. Ther.* 208, 107479

78 Atkinson, S.J. *et al.* (2014) The structure based design of dual HDAC/BET inhibitors as novel epigenetic probes. *MedChemComm* 5, 342–351

79 Yuan, Z. *et al.* (2019) Development of a versatile DNMT and HDAC inhibitor C02S modulating multiple cancer hallmarks for breast cancer therapy. *Bioorg. Chem.* 87, 200–208

80 San José-Enériz, E. *et al.* (2017) Discovery of first-in-class reversible dual small molecule inhibitors against G9a and DNMTs in hematological malignancies. *Nat. Commun.* 8, 15424

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

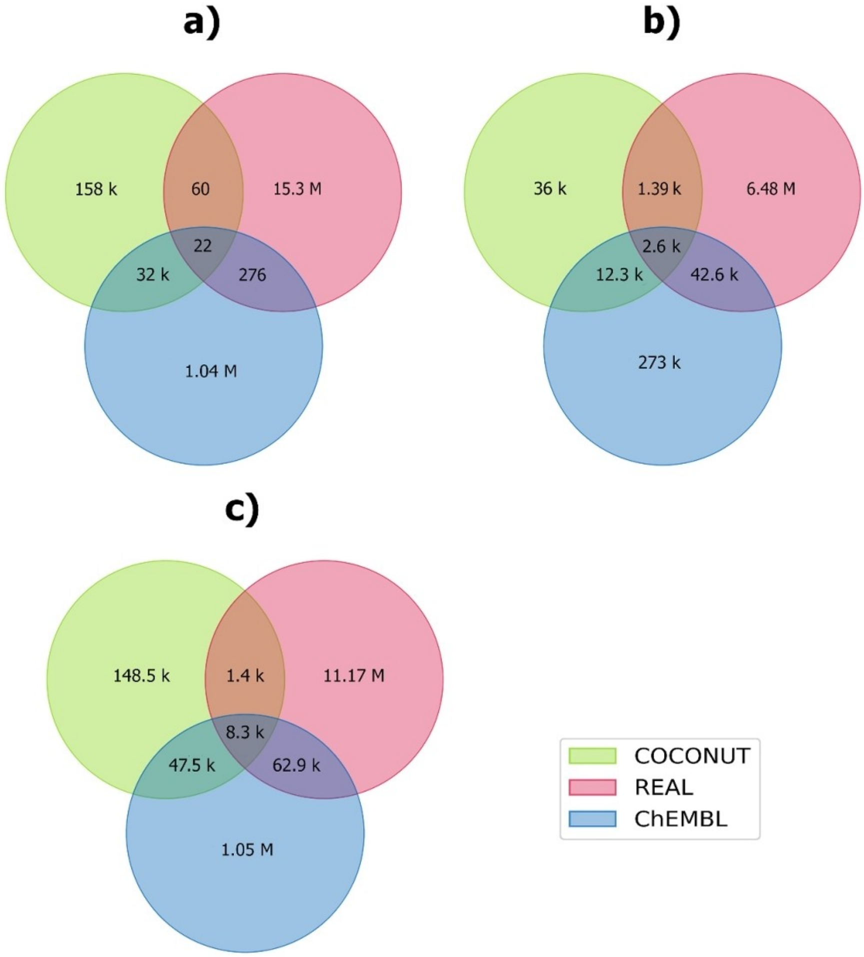


Figure 1. Unique and overlapping structures between COCONUT, ChEMBL and REAL data sets analyzed in this work (Table 1). Structural content was analyzed in terms of **a)** Compounds, **b)** Molecular scaffolds, and **c)** Fragments. The letter **k** represents thousands and the letter **M** represents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

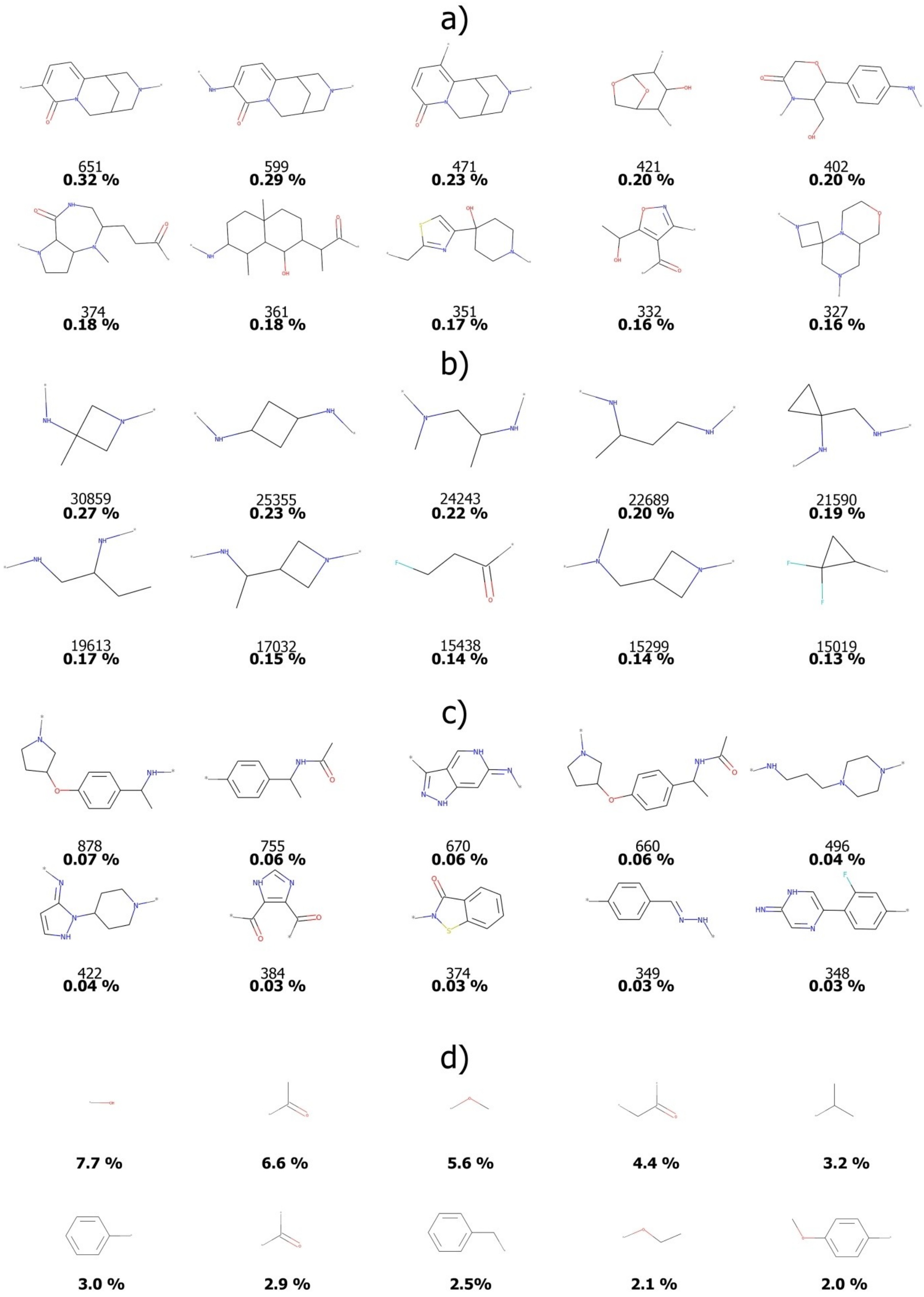


Figure 2. Ten most frequent unique fragments from **a)** COCONUT, **b)** Enamine-REAL, and **c)** ChEMBL.**d)** Ten most frequent common fragments in all three data sets. Occurrences in the data set are indicated in regular letter and percentage in bold.

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

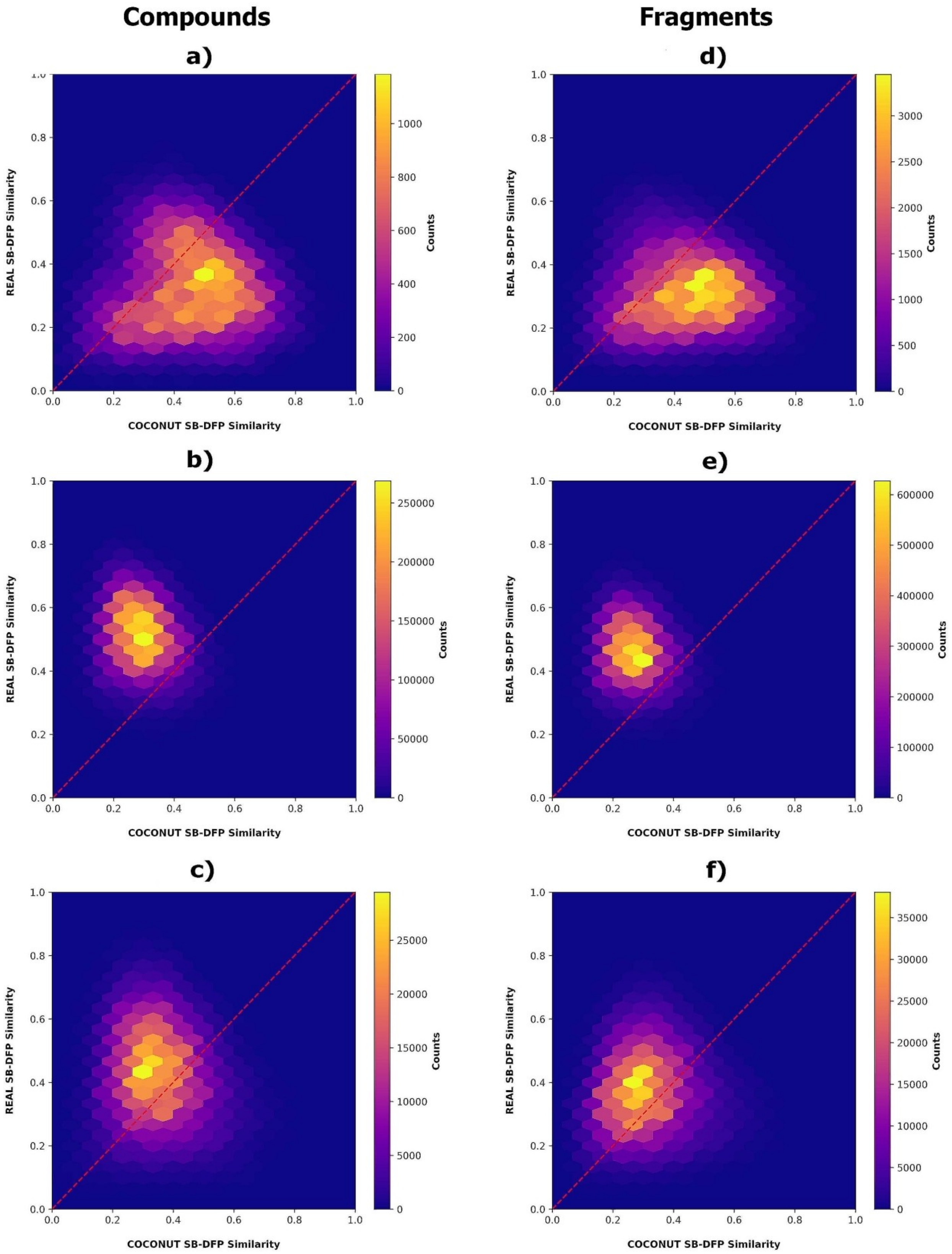


Figure 3. Visual representation of the chemical space for compounds and fragments of natural products, synthetics compounds, and biologically relevant compounds. The number of compounds is represented with a continuous color scale. Compounds data sets used: **a)** COCONUT, **b)** Enamine-REAL, **c)**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

RESEARCH ARTICLE

REVISED Functional group and diversity analysis of BIOFACQUIM: A Mexican natural product database [version 2; peer review: 3 approved]

Norberto Sánchez-Cruz, B. Angélica Pilón-Jiménez, José L. Medina-Franco

Department of Pharmacy, School of Chemistry, National Autonomous University of Mexico, Mexico City, Mexico City, 04510, Mexico

v2 **First published:** 10 Dec 2019, **8**(Chem Inf Sci):2071
https://doi.org/10.12688/f1000research.21540.1

Latest published: 08 Jun 2020, **8**(Chem Inf Sci):2071
https://doi.org/10.12688/f1000research.21540.2

Abstract

Background: Natural product databases are important in drug discovery and other research areas. An analysis of its structural content, as well as functional group occurrence, provides a useful overview, as well as a means of comparison with related databases. BIOFACQUIM is an emerging database of natural products characterized and isolated in Mexico. Herein, we discuss the results of a first systematic functional group analysis and global diversity of an updated version of BIOFACQUIM.
Methods: BIOFACQUIM was augmented through a literature search and data curation. A structural content analysis of the dataset was performed. This involved a functional group analysis with a novel algorithm to automatically identify all functional groups in a molecule and an assessment of the global diversity using consensus diversity plots. To this end, BIOFACQUIM was compared to two major and large databases: ChEMBL 25, and a herein assembled collection of natural products with 169,839 unique compounds.
Results: The structural content analysis showed that 15.7% of compounds and 11.6% of scaffolds present in the current version of BIOFACQUIM have not been reported in the other large reference datasets. It also gave a diversity increase in terms of scaffolds and molecular fingerprints regarding the previous version of the dataset, as well as a higher similarity to the assembled collection of natural products than to ChEMBL 25, in terms of diversity and frequent functional groups.
Conclusions: A total of 148 natural products were added to BIOFACQUIM, which meant a diversity increase in terms of scaffolds and fingerprints. Regardless of its relatively small size, there are a significant number of compounds and scaffolds that are not present in the reference datasets, showing that curated databases of natural products, such as BIOFACQUIM, can serve as a starting point to increase the biologically relevant chemical space.

Keywords

Consensus Diversity Plot, compound databases, data mining, diversity, natural products, functional groups, in silico

Open Peer Review

Reviewer Status ✓ ✓ ✓

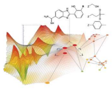
	Invited Reviewers		
	1	**2**	**3**
version 2 (revision) 08 Jun 2020	✓ report		
version 1 10 Dec 2019	? report	✓ report	✓ report

1 **Johannes Kirchmair**, University of Vienna, Vienna, Austria

2 **W. Patrick Walters**, Relay Therapeutics, Cambridge, USA

3 **Trong Tran**, University of the Sunshine Coast, Maroochydore, Australia

Any reports and responses or comments on the article can be found at the end of the article.



This article is included in the Chemical Information Science gateway.

This article is included in the Mathematical, Physical, and Computational Sciences collection.

Corresponding authors: Norberto Sánchez-Cruz (norberto.sc90@gmail.com), José L. Medina-Franco (medinajl@unam.mx)

Author roles: Sánchez-Cruz N: Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Software, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Pilón-Jiménez BA**: Investigation, Methodology, Writing – Original Draft Preparation, Writing – Review & Editing; **Medina-Franco JL**: Conceptualization, Funding Acquisition, Supervision, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This work was supported by a Consejo Nacional de Tecnología (CONACyT) scholarship number 335997 (NS-C). The work was also supported by the program NUATEI (Nuevas Alternativas para el Tratamiento de Enfermedades Infecciosas), Instituto de Ciencias Biomédicas, UNAM. The authors are grateful for the computational resources granted by Dirección General de Cómputo y de Tecnologías de Información y Comunicación (DGTIC), project grant LANCAD-UNAM-DGTIC-335 that allows to use the supercomputer Miztli at UNAM.
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: Sánchez-Cruz N, Pilón-Jiménez BA and Medina-Franco JL. **Functional group and diversity analysis of BIOFACQUIM: A Mexican natural product database [version 2; peer review: 3 approved]** F1000Research 2020, **8**(Chem Inf Sci):2071 https://doi.org/10.12688/f1000research.21540.2

First published: 10 Dec 2019, **8**(Chem Inf Sci):2071 https://doi.org/10.12688/f1000research.21540.1

REVISED **Amendments from Version 1**

We thank all three reviewers for the constructive comments and suggestions to improve the study and the manuscript. In the revised version, we repeated the analysis standardizing the chemical structures using MolVS. The standardization procedure is described in the revised "Methods, Databases and curation" section; the code was made available at GitHub. To visualize the chemical space of BIOFACQUIM and reference compounds databases (entire sets and drug-like subsets) we employed the recently published method TMAP (Tree Manifold Approximation and Projection) using Morgan fingerprints with radius 2. An analysis of the distribution of the fraction of carbon atoms that are sp3 hybridized was added as a new subsection "Molecular complexity". Further details of the BIOFACQUIM database were added (source of compounds and names of the peer-reviewed journals where the compound information was retrieved from). The English language of the entire manuscript was revised.

Any further responses from the reviewers can be found at the end of the article

Introduction

Natural product-based drug discovery continues to be an important part of drug discovery. Recently, the synergy between natural product research with molecular modeling and chemoinformatics is gaining importance, speeding up the drug discovery process[1,2]. As part of these synergistic efforts, curated databases of natural products have an important role as they are major tools for data mining, hypothesis generation, and starting points of virtual screening. There are several databases of natural products in the public domain as reviewed recently[3]. Our research group has reported initial efforts to assemble a database of natural products from Mexico called BIOFACQUIM[4]. As part of that work, scaffold content and chemical space diversity were examined. However, detailed functional group (FG) content analysis, which has been proven to be valuable to characterize compound databases[5], in particular from natural sources[6], has not been reported for BIOFACQUIM. One of the main reasons is that most of the currently available software employed for identification of functional groups rely on a predefined set of substructures, even when it has been established that one of the major features that discriminate natural products from synthetic compounds are their unique functional groups.

Herein, we report a functional group content analysis of an updated version of BIOFACQUIM. We employed a validated and novel algorithm that identifies all functional groups in a molecule. As part of the analysis and to compare the results of BIOFACQUIM we also discuss the functional group contents of other large and related databases in the public domain, namely ChEMBL 25[7] and a herein assembled collection of natural products (NPs) with 169,839 compounds.

Methods

Databases and data curation

As described elsewhere, the first version of BIOFACQUIM was developed as a proof-of-concept database applying several filters to include compounds[4]. Briefly, the database was focused on natural products published between 2000 and 2018 by research groups in a major Mexican institution in eight indexed journals: *Journal of Ethnopharmacology*, *Natural Products Research*, *Journal of Agricultural and Food Chemistry*, *Journal of Natural Products*, *Planta Medica*, *Phytochemistry, Natural Product Letters*, and *Molecules*. As additional criteria for inclusion of compounds and to increase the quality and reliability of the contents of the database, the procedure for the isolation, purification, and characterization of the natural product should have been described in the article. In this work, we expanded the contents of the BIOFACQUIM database to further explore the diversity of natural products from Mexico.

The second version of BIOFACQUIM was assembled using the same methodology described to develop the first version[4] extending the date of publication to 2019. To achieve the objective of being representative of Mexico, one additional criterion was considered, including only compounds collected in Mexico at any of its institutions (universities, research laboratories and research centers). For the new version of the database, the same procedure for the curation was performed[4], using Molecular Operating Environment (MOE) software, although this procedure can be performed using open source software, such as MolVS and RDKit. The updated and curated version of BIOFACQUIM contains 531 compounds.

Table 1 summarizes the information of BIOFACQUIM and other major compound databases used in this work as reference: ChEMBL 25 as a representative example of the biologically

Table 1. Compound databases analyzed in this work and summary statistics of their diversity.

Database	Size (compounds)	Median similarity (MACCS keys - 166 bits)	Median similarity (Morgan2 - 1024 bits)	Mean distance (PCP)	Scaffold diversity (AUC)	Scaffold diversity (F_{50})
BIOFACQUIM V1	403	0.457	0.123	3.648	0.725	0.165
BIOFACQUIM V2	503	0.446	0.119	3.319	0.710	0.171
Natural products	168,030	0.422	0.111	3.775	0.830	0.032
ChEMBL 25	1,667,509	0.382	0.117	2.187	0.809	0.057

PCP: physicochemical properties; AUC: area under the cyclic system retrieval curve.

tested chemical space with 1,667,509 unique compounds; and a collection of known natural products with a total of 168,030 molecules. The reference natural product collection was assembled from three general and publicly available natural products databases: the Universal Natural Products Database (UNPD)[8], the Natural Products Atlas[9] and Natural Products in PubChem Substance Database[10]. The data sets were curated using the same procedure. Briefly, 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. Finally, canonical simplified molecular-input line-entry system (SMILES) (ignoring stereochemistry information) were generated as molecular representation and duplicate structures in the context of each database were removed. The entire process was performed by using the functions Standardizer, LargestFragmentChoser, Uncharger, Reionizer and TautomerCanonicalizer implemented in the molecule validation and standardization tool MolVS for the open source cheminformatics toolkit RDKit. The code is available at GitHub (https://github.com/DIFACQUIM/IFG_General).

Databases overlap

Overlap of BIOFACQUIM with the databases selected as reference was assessed in terms of three different structural levels: compounds, scaffolds, and functional groups. Compound overlap was determined in terms of canonical SMILES. For scaffold comparison we use the definition proposed by Bemis and Murcko[11] as implemented in RDKit, while for functional group overlap we selected the recently published definition and implementation suggested by Ertl[5]. For each structural level, we identified the unique structures belonging to each dataset as well as those belonging to two or three of them.

Functional group analysis

For the functional group content analysis we selected the algorithm recently described by Ertl[5], which is able to identify all functional groups in a molecule based on an iterative marching through its atoms. In short, the proposed algorithm identifies all heteroatoms in a molecule, all atoms connected by multiple bonds as well as the atoms in oxirane, aziridine, and thiirane rings. Afterwards, all connected atoms are joined together to form a functional group. Single aromatic heteroatoms are retained only if they are connected to an additional aliphatic functionality. Finally, a generalization scheme is applied in which for a defined list of common FGs, information about the parent carbon is retained (e.g. to differentiate between alcohols and phenols) as well as hydrogen atoms (e.g. to differentiate between aldehydes and ketones). The method is fully described in 5. An open source version of this algorithm is available for Python (https://github.com/rdkit/rdkit/tree/master/Contrib/IFG); however, it does not cover the generalization scheme proposed originally. To this end and based on the code available, we implemented with RDKit a fragmentation approach considering keeping the parent carbon and hydrogen atoms proposed originally, were the remaining carbon atoms are replaced by dummy atoms. This implementation works over a SMILES string and returns a list with the canonical SMILES of the FGs identified in the molecule. The code is freely available at GitHub (https://github.com/DIFACQUIM/IFG_General). After determining the FGs content of the different datasets, we compare the proportion of the most frequent FGs at each library.

Complexity analysis

The fraction of carbon atoms that are sp^3 hybridized (F-sp3) is a common metric to quantify molecular complexity[12]. Higher values for this descriptor have been associated to an improved binding selectivity of compounds[13]. In order to compare the complexity of BIOFACQUIM with the data sets selected as reference, we computed the F-sp3, as implemented in RDKit, for all compounds in the three data sets and compared its distribution among libraries.

Chemical space visualization

In order to generate a visual representation of the chemical space covered by the analyzed databases, we selected a recently proposed method, named TMAP[14] (Tree Manifold Approximation and Projection). This method enables the visualization of up to millions of data points with high dimensionality as a two-dimensional tree and has shown to be better suited than t-distributed Stochastic Neighbor Embedding[15] (t-SNE) and Uniform Manifold Approximation and Projection[16] (UMAP) for the exploration of large datasets. TMAP consists in four phases: (I) the input data are indexed in an local sensitive hashing forest data structure, using l prefix trees and d hash functions in encoding the data, (II) an undirected weighted c-approximate k-nearest neighbor graph (c-k-NNG) is constructed from the indexed data points with the Jaccard distances between vertices used as edges weights, (III) a minimum spanning tree (MST) is constructed for the weighted c-k-NNG using Kruskal's algorithm and (IV) a layout for the resulting MST is constructed by using a spring-electrical model layout algorithm with multilevel multipole-based force approximation as provided by the modular C++ library, open graph drawing framework[17] (OGDF).

For the description of compounds, we selected Morgan fingerprints with radius 2 (Morgan2, 1024-bits) as implemented in RDKit. For the generation of the TMAP, the input data was encoded by 1024 hash functions and indexed using 64 prefix trees. The weighted c-k-NNG was built using the 5 nearest neighbors and a factor of 20 for the LSH forest query algorithm. For the layout generation a node size of 0.01 was selected while all the remaining parameters were set to default. These calculations were done using the TMAP python package and all the charts were generated using the matplotlib library[18].

Global diversity

The "global" or total diversity of the datasets was analyzed through the Consensus Diversity (CD) Plot[19]. A CD Plot is a two-dimensional representation of compound datasets based on four different and complementary diversity criteria: molecular fingerprints, molecular scaffolds, physicochemical properties (PCP), and size. Fingerprint-based diversity of each dataset is represented in the X-axis, while scaffold-based diversity is represented in the Y-axis, PCP-based diversity is represented as the filling of the data points using a continuous color scale and the size of the data set is represented with the size of the data points.

For this work, scaffold diversity was assessed as the area under the cyclic system retrieval curve and the fraction of chemotypes that covers 50% of the dataset (F_{50}). The median of the lower triangle from the pairwise similarity matrix computed as the Tanimoto coefficient of both MACCS keys (166-bits) fingerprint and Morgan2 (1024-bits), were used as molecular fingerprint-based diversity. For PCP-based diversity, six molecular properties of pharmaceutical interest were computed for each unique compound, being 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); PCP-based diversity was measured as the mean distance of the lower triangle of the pairwise distance matrix computed as the Euclidean distance of those PCPs scaled (mean 0 and unit variance). The number of compounds in each dataset was selected as measure of the size-based diversity. PCPs were calculated as implemented in RDKit and the CD Plot was constructed using R.

Results and discussion

Update of BIOFACQUIM

As described in the Methods section, the updated BIOFACQUIM database contains the chemical structure of 531 compounds, all collected from Mexico. As with the first version[4], each molecule is annotated with information of the chemical structure, the original source of the information (Digital Object Identifier, DOI, to reference paper), kingdom, genus, and species of the organism from which the natural product was isolated, place of collection (city and state), and activity value of the reported biological activity. From the original dataset containing 423 compounds, 40 were discarded since they were not collected in Mexico, which means an increase of 148 unique compounds compared to the previous release of the database. The sources of the 531 compounds are distributed as follows: 406 from plants, 97 from fungus, 15 from propolis and 13 from marine animals.

Database overlap

To assess the chemical space not covered by ChEMBL and NPs but by BIOFACQUIM, we characterized the structural content of the three datasets in terms of unique compounds, scaffolds and functional groups and determined the overlap among them. Figure 1 depicts Venn diagrams showing the overlap

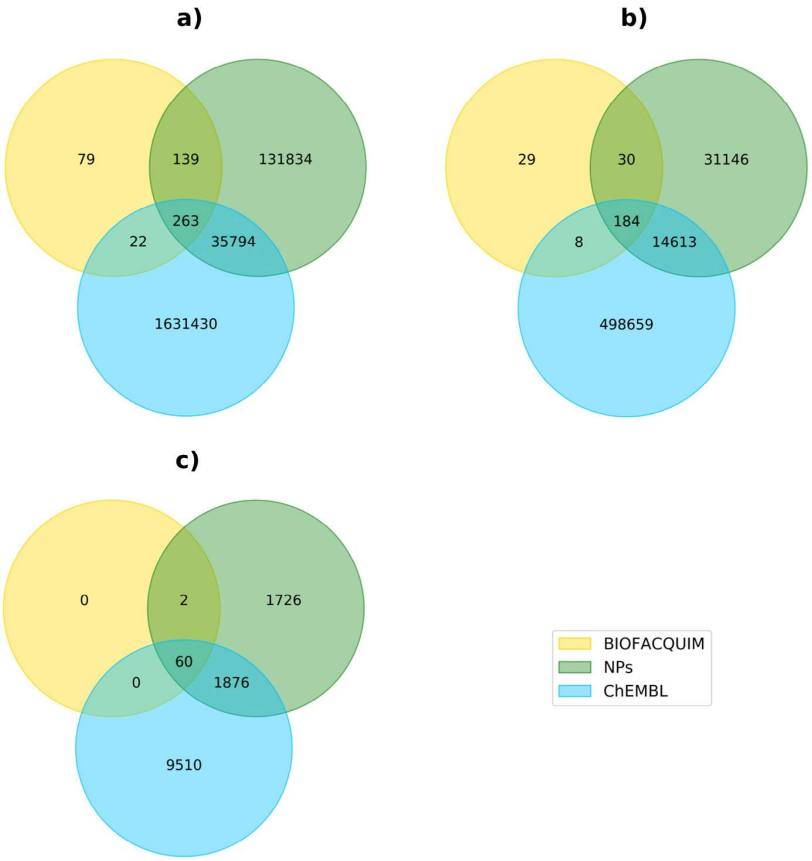


Figure 1. **Overlap between BIOFACQUIM, reference natural products (NPs) and ChEMBL.** Datasets content was analyzed in terms of (**a**) Compounds, (**b**) Scaffolds and (**c**) Functional Groups.

among those datasets. It should be noted that despite its small size in comparison with ChEMBL and NPs, 15.7% of compounds present in BIOFACQUIM have not been reported in those other major datasets, as well as 11.6% of its scaffolds. Figure 2 shows the structure of representative scaffolds identified in at least 2 compounds from BIOFACQUIM, all of them associated to compounds isolated from different plants. Scaffold **a** associated to hexasaccharides of convolvulinic and jalapinolic acids isolated from *Ipomoea purga*[20], scaffold **b** corresponding to glycosides of 4-phenylcoumarin isolated from *Exostema caribaeum*[21], scaffold **c** representing karwinaphthopyranones A2 and B3 isolated from *Karwinskia parvifolia*[22] and scaffold **d** associated to batatins X and XI isolated from *Ipomoea batatas*[23]. Another remarkable observation is the fact that most of the overlap of BIOFACQUIM with the other datasets involves NPs, either alone or in combination with ChEMBL, being 79.9%, 85.3%, and 100% for compounds, scaffolds and FGs, respectively.

Functional group analysis

A systematic analysis of functional groups was carried out over BIOFACQUIM and the two datasets selected as reference. 62, 3664, and 11446 functional groups were identified in BIOFACQUIM, NPs and ChEMBL datasets, respectively (the overlap between them is shown in Figure 1). From the total number of functional groups present in each dataset, only 12, 15, and 22 were present in at least 1% of the corresponding library (19.4%, 0.4% and 0.2%, respectively) while 30, 1879, and 5212 (48.4%, 51.3% and 45.5%, respectively) were singletons. This result is consistent with the typical power law observed in other databases[6]. The most frequent FGs present in BIOFACQUIM are oxygen-containing FGs, being the phenolic hydroxyl group (46.1%), followed by ether (41.4%), alcohol hydroxyl group (38.4%), alkene (28.6%) and ester (26.8%), which although in a different order, are the most frequent FGs in the herein assembled NPs collection and in other natural product libraries[6]. This is in contrast to ChEMBL in which only ether is part of the most frequent FGs while the rest of them are nitrogen containing FGs and halogens. The complete results of the FGs found of the datasets is included as *Extended data* (Supplementary File 1).

Comphlexity analysis

As described in the Methods section, molecular complexity of BIOFACQUIM, NPs and ChEMBL data sets were assessed through the F-sp3 of its compounds. Figure 3 depicts letter plots for the distribution of this descriptor among data sets. This shows that NPs is the more complex data set overall regarding to this metric, with a median of 0.60 and a long tail towards low values. In contrast, ChEMBL is the less complex data set, with a median F-sp3 of 0.31 and a long tail towards high values. BIOFACQUIM is in an intermediate position with a median of 0.42 and a more symmetrical distribution, which is consistent with its small size in comparison to the other sets and its major overlap with NPs.

Chemical space visualization

For visualization of the chemical space of the datasets compared in this work, we built a TMAP as described in the Methods section. This method allows the representation of large datasets as a two-dimensional tree. TMAP shows the relationships among subsets of data points and data points itself as branches and sub-branches, so similar compounds and clusters tend to be close in the final representation even if the tree edges are not included, for that reason all charts in this work do not show the tree edges. Figure 4 shows a visual representation of the chemical space of the three datasets analyzed in this work, including the whole datasets and drug-like subsets. Drug-like compounds for each subset were defined those complying with the Lipinski "rule of 5"[24] and Veber criteria[25] (AMW ≤ 500, -1 ≤ SlogP ≤ 5, HBA ≤ 10, HBD ≤ 5, RB ≤ 10 and TPSA ≤ 140). For this plot, in order to better illustrate the unique compounds present in BIOFACQUIM, all compounds belonging to more

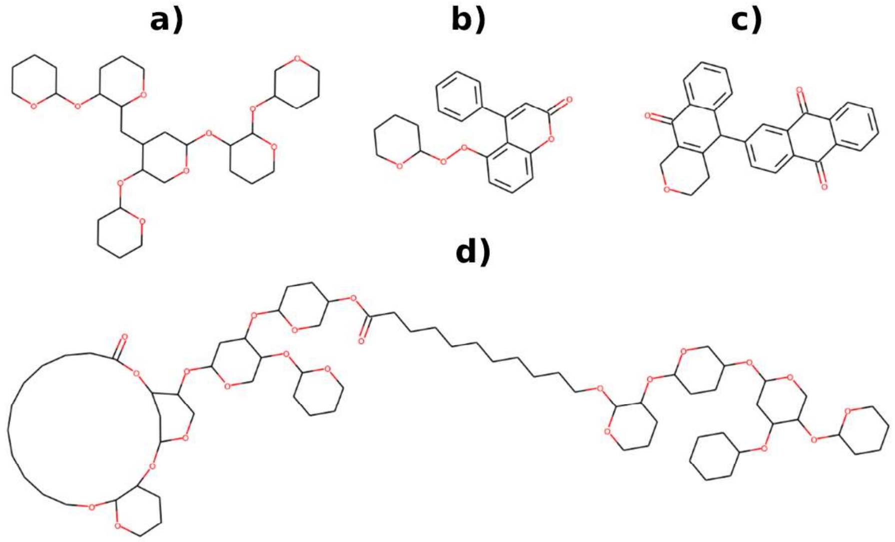


Figure 2. Representative unique scaffolds from BIOFACQUIM.

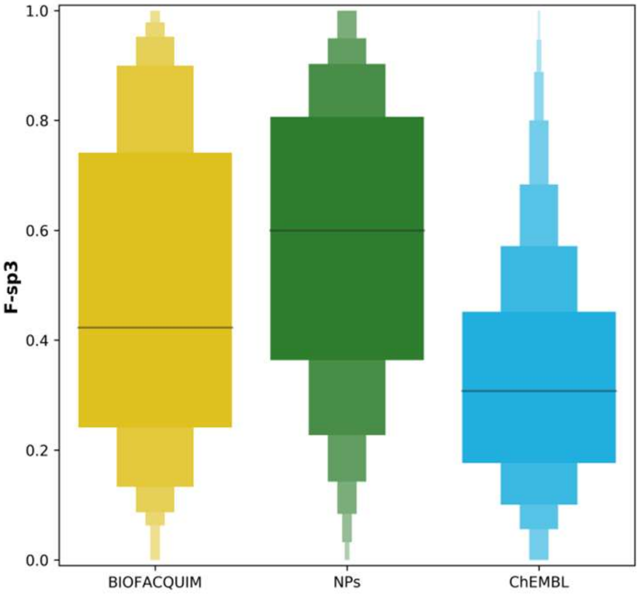


Figure 3. Distribution of F-sp3 for BIOFACQUIM, NPs, and ChEMBL.

than one library were assigned to a single one: ChEMBL if they belong to this dataset, NPs if they belong to this dataset but not to ChEMBL and BIOFACQUIM if they were unique for this library. Figure 4 shows that the chemical space covered by the analyzed datasets is practically defined by ChEMBL and highly focused in the upper right section of the plot, meaning that the biologically relevant space does not cover evenly the available chemical space. It is also shown that NPs cover, in a sparser manner, the same space as ChEMBL. Unique compounds from BIOFACQUIM on the other hand are distributed only in the less populated regions of the space, and even in the outer region of the plot, which implies the presence of few similar compounds in the other datasets. All these observations are equally applicable for the drug-like subsets from the original data sets, which represent 44.3%, 48.4% and 69.1% of BIOFACQUIM, NPs, and ChEMBL, respectively.

Global diversity

In order to compare the chemical diversity of the current version of BIOFACQUIM with the previous one and the two datasets selected as reference, we employed a CD Plot. Figure 5 shows the plot comparing the diversity of all datasets considering four different criteria: scaffolds in the *y*-axis, molecular fingerprints

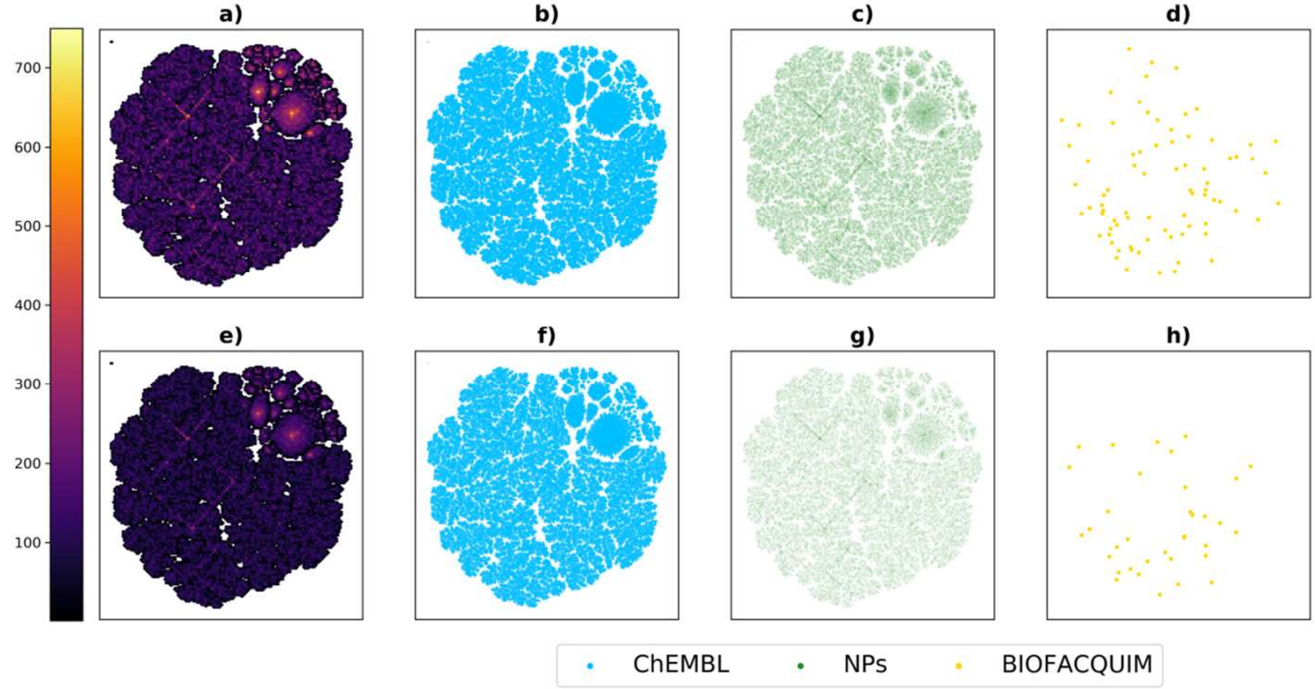


Figure 4. TMAP visualization of the chemical space covered by BIOFACQUIM. Comparison of BIOFACQUIM with two reference datasets. Panels **a**–**d** show all compounds in the datasets, panels **e**–**h** show drug-like compounds only. Panels **a** and **f** show the distribution of compounds from the three data sets among the chemical space in a continuous color scale.

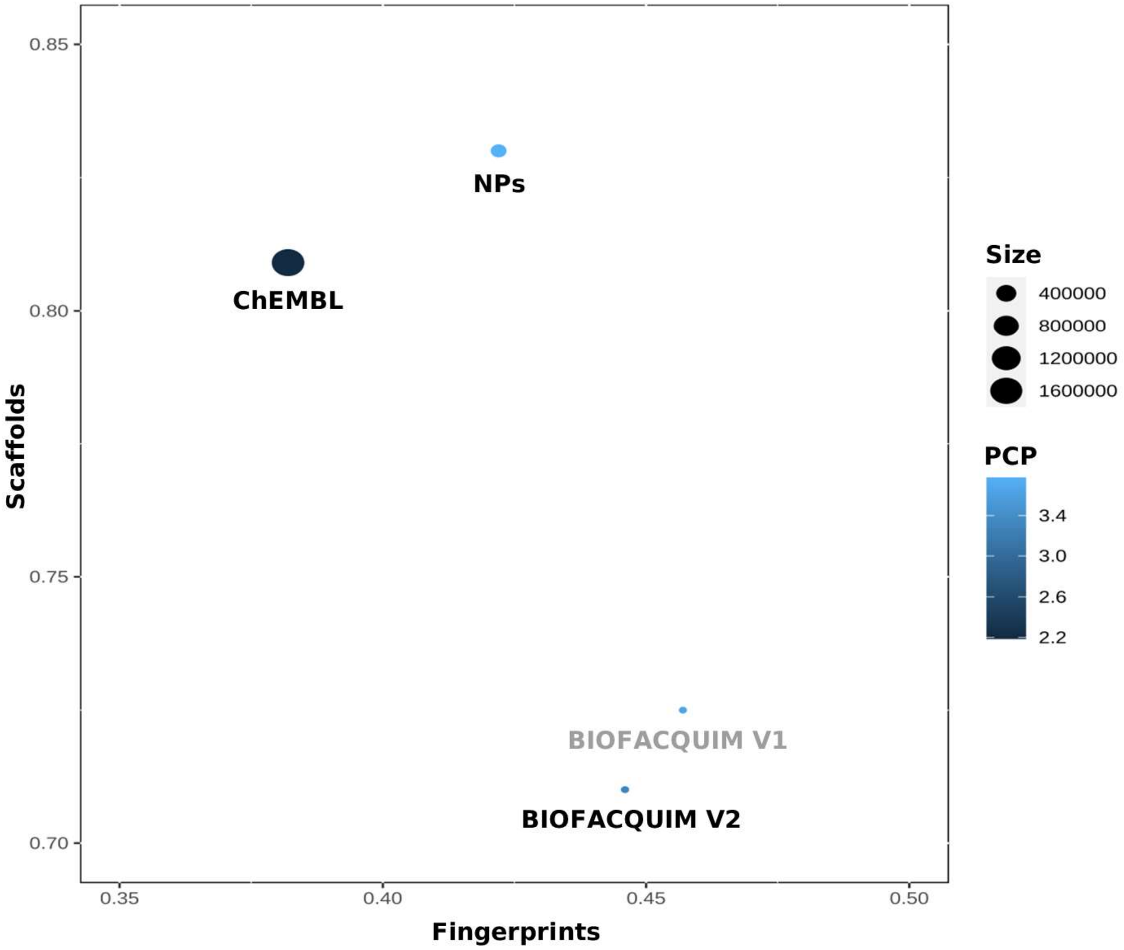


Figure 5. Consensus diversity plot of BIOFACQUIM.

in the *x*-axis, physicochemical properties as the filling of the data points in a continuous color scale, and number of compounds as the data points size. This comparison shows the relatively small size of BIOFACQUIM in comparison with the reference datasets. As compared to the previous release of BIOFACQUIM, the current version has increased its diversity in terms of scaffolds and fingerprints but decreased in terms of physicochemical properties. Also, it is shown that its diversity in terms of molecular fingerprints and physicochemical properties, although not the greatest ones of the three datasets, are closer to the ones for NPs, contrary to scaffolds, in which is the most diverse.

The molecular fingerprint diversity of each data set is represented on the *x*-axis and was defined as the median Tanimoto coefficient of MACCS keys (166-bits) fingerprint. The scaffold diversity of each database is represented on the *y*-axis and was defined as the area under the corresponding cyclic system retrieval curve. The diversity based on physicochemical properties (PCP) was defined as the mean euclidean distance of six scaled physicochemical properties (SlogP, TPSA, AMW, RB, HBD, and HBA) and is shown as the filling of the data points using a continuous color scale. The number of compounds is represented by the size of the data points.

Conclusions

The current version of BIOFACQUIM involved the addition of 148 natural products. This was reflected in a diversity increase based on both scaffolds and molecular fingerprints. It was shown that in terms of diversity, structural content overlap and complexity, BIOFACQUIM is more similar to the assembled set of natural products than to the set of biologically tested compounds. The herein reported chemoinformatic study revealed that 44.3% of the unique compounds contained in BIOFACQUIM are focused in the drug-like space in terms of physicochemical properties. Interestingly, despite the fact of its relative small size, there were identified a significant number of compounds and scaffolds (79 and 29, respectively) that were not present in the two large sets used as reference, showing that curated databases of natural products, such as BIOFACQUIM, can serve as a starting point for the study and increase of the biologically relevant chemical space.

Data availability

Underlying data

Figshare: BIOFACQUIM_V2. http://doi.org/10.6084/m9.figshare.11312702

This file contains the chemical structures of 531 compounds in SDF format, alongside ID number, compound name, simplified molecular input line entry system, literature reference, kingdom, genus, species, geographical location and biological activity.

Underlying data are available under the terms of the Creative Commons Zero "No rights reserved" data waiver (CC0 1.0 Public domain dedication).

Extended data

Figshare: Supporting information for "Functional group and diversity analysis of BIOFACQUIM: A Mexican natural product database". http://doi.org/10.6084/m9.figshare.11312735

This project contains the following extended data:

- Supplementary File 1. File with summary results of the functional group analysis.

Extended data are available under the terms of the Creative Commons Attribution 4.0 International License (CC BY 4.0).

Acknowledgments

We thank Roberto Hernández for his valuable contribution toward the update of BIOFACQUIM.

References

1. Kinghorn AD, Falk H, Gibbons S, *et al.*: eds. **Progress in the Chemistry of Organic Natural Products 110: Cheminformatics in Natural Product Research.** Cham: Springer International Publishing. 2019; **110**.
Publisher Full Text
2. Medina-Franco JL: **New Approaches for the Discovery of Pharmacologically-Active Natural Compounds.** *Biomolecules.* 2019; **9**(3): 115.
PubMed Abstract | Publisher Full Text | Free Full Text
3. Chen Y, Garcia de Lomana M, Friedrich NO, *et al.*: **Characterization of the Chemical Space of Known and Readily Obtainable Natural Products.** *J Chem Inf Model.* 2018; **58**(8): 1518–1532.
PubMed Abstract | Publisher Full Text
4. Pilón-Jiménez BA, Saldívar-González FI, Díaz-Eufracio BI, *et al.*: **BIOFACQUIM: A Mexican Compound Database of Natural Products.** *Biomolecules.* 2019; **9**(1): 31.
PubMed Abstract | Publisher Full Text | Free Full Text
5. Ertl P: **An Algorithm to Identify Functional Groups in Organic Molecules.** *J Cheminform.* 2017; **9**(1): 36.
PubMed Abstract | Publisher Full Text | Free Full Text
6. Ertl P, Schuhmann T: **A Systematic Cheminformatics Analysis of Functional Groups Occurring in Natural Products.** *J Nat Prod.* 2019; **82**(5): 1258–1263.
PubMed Abstract | Publisher Full Text
7. Gaulton A, Hersey A, Nowotka M, *et al.*: **The ChEMBL database in 2017.** *Nucleic Acids Res.* 2017; **45**(D1): D945–D954.
PubMed Abstract | Publisher Full Text | Free Full Text
8. Gu J, Gui Y, Chen L, *et al.*: **Use of Natural Products as Chemical Library for Drug Discovery and Network Pharmacology.** *PLoS One.* 2013; **8**(4): e62839.
PubMed Abstract | Publisher Full Text | Free Full Text
9. van Santen JA, Jacob G, Singh AL, *et al.*: **The Natural Products Atlas: An Open Access Knowledge Base for Microbial Natural Products Discovery.** *ACS Cent Sci.* 2019; **5**(11): 1824–1833.
PubMed Abstract | Publisher Full Text | Free Full Text
10. Ming H, Tiejun C, Yanli W, *et al.*: **Web Search and Data Mining of Natural Products and Their Bioactivities in PubChem.** *Sci China Chem.* 2013; **56**(10): 1424–1435.
PubMed Abstract | Publisher Full Text | Free Full Text
11. Bemis GW, Murcko MA: **The properties of known drugs. 1. Molecular frameworks.** *J Med Chem.* 1996; **39**(15): 2887–2893.
PubMed Abstract | Publisher Full Text
12. Lovering F, Bikker J, Humblet C: **Escape From Flatland: Increasing Saturation as an Approach to Improving Clinical Success.** *J Med Chem.* 2009; **52**(21): 6752–6756.
PubMed Abstract | Publisher Full Text
13. Clemons PA, Bodycombe NE, Carrinski HA, *et al.*: **Small Molecules of Different Origins Have Distinct Distributions of Structural Complexity That Correlate With Protein-Binding Profiles.** *Proc Natl Acad Sci.* 2010; **107**(44): 18787–18792.
PubMed Abstract | Publisher Full Text | Free Full Text
14. Probst D, Reymond JL: **Visualization of Very Large High-Dimensional Data Sets as Minimum Spanning Trees.** *J Cheminform.* 2020; **12**(1): 12.
Publisher Full Text
15. van der Maaten L, Hinton G: **Visualizing Data Using T-SNE.** *J Mach Learn Res.* 2008; **9**: 2579–2605.
Reference Source
16. McInnes L, Healy J, Melville J: **UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction**. 2018; arXiv:1802.03426v2.
Reference Source
17. Chimani M, Gutwenger C, Junger M, *et al.*: **The Open Graph Drawing Framework (OGDF).** *Handb Graph Draw Vis.* 2013; 543–570.
Reference Source
18. Hunter JD: **Matplotlib: A 2D Graphics Environment.** *Comput Sci Eng.* 2007; **9**(3): 90–95.
Publisher Full Text
19. González-Medina M, Prieto-Martínez FD, Owen JR, *et al.*: **Consensus Diversity Plots: a Global Diversity Analysis of Chemical Libraries.** *J Cheminform.* 2016; **8**: 63.
PubMed Abstract | Publisher Full Text | Free Full Text
20. Pereda-Miranda R, Fragoso-Serrano M, Escalante-Sanchez E, *et al.*: **Profiling of the Resin Glycoside Content of Mexican Jalap Roots with Purgative Activity.** *J Nat Prod.* 2006; **69**(10): 1460–1466.
PubMed Abstract | Publisher Full Text
21. Pérez-Vásquez A, Castillejos-Ramírez E, Cristians S, *et al.*: **Development of a UHPLC-PDA Method for the Simultaneous Quantification of 4-phenylcoumarins and Chlorogenic Acid in *Exostema Caribaeum* Stem Bark.** *J Nat Prod.* 2014; **77**(3): 516–520.
PubMed Abstract | Publisher Full Text
22. Rojas-Flores C, Rios MY, López-Marure R, *et al.*: **Karwinaphthopyranones From the Fruits of *Karwinskia Parvifolia* and Their Cytotoxic Activities.** *J Nat Prod.* 2014; **77**(11): 2404–2409.
PubMed Abstract | Publisher Full Text
23. Rosas-Ramírez D, Pereda-Miranda R: **Batatins VIII-XI, Glycolipid Ester-Type Dimers From *Ipomoea Batatas*.** *J Nat Prod.* 2015; **78**(1): 26–33.
PubMed Abstract | Publisher Full Text
24. Lipinski CA, Lombardo F, Dominy BW, *et al.*: **Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings.** *Adv Drug Deliv Rev.* 2001; **46**(1–3): 3–26.
PubMed Abstract | Publisher Full Text
25. Veber DF, Johnson SR, Cheng HY, *et al.*: **Molecular Properties That Influence the Oral Bioavailability of Drug Candidates.** *J Med Chem.* 2002; **45**(12): 2615–2623.
PubMed Abstract | Publisher Full Text

Sánchez-Cruz and Medina-Franco
J Cheminform (2018) 10:55
https://doi.org/10.1186/s13321-018-0311-x

Journal of Cheminformatics

METHODOLOGY

Open Access

Statistical-based database fingerprint: chemical space dependent representation of compound databases

Norberto Sánchez-Cruz* and José L. Medina-Franco*

Abstract

Background: Simplified representation of compound databases has several applications in cheminformatics. Herein, we introduce an alternative and general method to build single fingerprint representations of compound databases. The approach is inspired on the previously published modal fingerprints that are aimed to capture the most significant bits of a fingerprint representation for a compound data set. The novelty of the herein proposed statistical-based database fingerprint (SB-DFP) is that it is generated based on binomial proportions comparisons taking as reference the distribution of "1" bits on a large representative set of the chemical space.

Results: To illustrate the Method, SB-DFPs were constructed for 28 epigenetic target data sets retrieved from a recently published epigenomics database of interest in probe and drug discovery. For each target data set, the SB-DFPs were built based on two representative fingerprints of different design using as reference a data set with more than 15 million compounds from ZINC. The application of SB-DFP was illustrated and compared to other methods through association relationships of the 28 epigenetic data sets and similarity searching. It was found that SB-DFPs captured overall, the common features between data sets and the distinct features of each set. In similarity searching SB-DFP equaled or outperformed other approaches for at least 20 out of the 28 sets.

Conclusions: SB-DFP is a general approach based on binomial proportion comparisons to represent a compound data set with a single fingerprint. SB-DFP can be developed, at least in principle, based on any fingerprint and reference data set. SB-DFP is a good alternative for exploration of relationships between targets through its associated compound data sets and performing similarity searching.

Keywords: Chemical space, Epi-informatics, Molecular fingerprints, Representation, Similarity searching

Background

Molecular fingerprints are bit strings representations of chemical structures in which each position indicates the presence (1) or absence (0) of chemical features as defined in the design of the fingerprint. There are several types of molecular fingerprints described elsewhere [1, 2]. Such representations are broadly employed for the assessment of chemical space coverage, molecular diversity and similarity searching [1–3]. With the constant increasing size of chemical databases, such studies have become more computationally demanding, leading to the need of generating simplified representations of compound databases to optimize storage and calculation speed. To this end, many of the approaches that have been proposed generate a single fingerprint trying to capture the common chemical features presents in all compounds in a database (or at least in most of them). The first strategy dates back to 1996, when Shemetulskis et al. [4] employed the Daylight Chemical Information Systems, Inc. molecular fingerprint to build the so-called modal fingerprint, which contains the common bits found in

*Correspondence: norberto.sc90@gmail.com; medinajl@unam.mx
Department of Pharmacy, School of Chemistry, Universidad Nacional Autónoma de México, Avenida Universidad 3000, 04510 Mexico City, Mexico

This work is dedicated to Dr. Gerald M. Maggiora on the occasion of his 80th Birthday.

the molecular fingerprints in a given compound data set. In the modal fingerprint, the degree to which bits have to be in common in the data set in order to be set as "1" is determined by a user-defined threshold value, which ranges from 50 to 100%, being 50% the best performing threshold in different studies. Since 1996 the algorithm has been extended to different molecular fingerprints and a number of studies have shown its application in similarity searching [5, 6] and for the quantification of intra- and inter-database diversity [7]. In parallel, several modifications to this concept have been developed, mostly aiming to enhance its performance on similarity searches at the expense of increasing the complexity to implement the approach. Such approaches include bit scaling [8–10], bit silencing [11] and the determination of the best feature combinations [12]. In different publications the term "modal fingerprint" has been used to refer to distinct approaches. To avoid confusions, herein we refer as "database fingerprint (DFP)" to the modal fingerprint constructed using 50% as the predefined threshold.

In this work, we present the statistical-based database fingerprint (SB-DFP) as a novel and general approach to generate a compound database fingerprint based on binomial proportion comparisons. In this paper, we illustrate the application of SB-DFP in comparing target-associated compound data sets and performing similarity searching. As a case study, and to further advance the emerging field of epi-informatics [13], the SB-DFPs were applied to a recently published epigenomics database with potential therapeutic significance.

Methods

Concept and construction of SB-DFP

As commented on the Background, in a "classic" DFP representation, to set a bit "1" requires that such bit is present in at least 50% of all the molecules in the input data set. The basic idea of such threshold is to extract common bits in at least half of the input data set. However, the underlying hypothesis assumes that the probability of presence of a feature (bit) in a molecular representation is 50% for each of them, so all bits are compared against such probability.

SB-DFP is based on the basic hypothesis that the probability of presence of a feature (bit) in a molecular representation is not equal for each bit. Instead, it is determined by the availability of such feature in a reference set e.g. the "known chemical space" (or a reasonable approximation) and such availability has to be determined. Once the frequency occurrence of each bit in a molecular representation is determined for both, namely the reference set and the data set of study, the SB-DFP is constructed by comparing the frequency occurrence of each bit between both sets. Thus, a bit is set to "1" only if the frequency in the target data set is statistically higher than the reference. Figure 1 depicts a schematic comparison between a classic DFP (reminiscent of the modal fingerprint, vide supra) and the SB-DFP, respectively. In this scheme the database fingerprint is illustrated for a short hypothetical fingerprint representation with 20-bit positions.

It should be noted that the SB-DFP representation for a given data set requires three main features (Fig. 1b): (1) a reference set, (2) a molecular fingerprint representation and 3) a statistical method to do the binomial proportion comparisons. The chosen features for this work are described below, although SB-DFP can be developed with different fingerprints, reference sets, and statistical methods.

Compound data sets

As a case study we generated SB-DFPs for a recently published epigenomics database [14]. The set of targets used as a test case in this work were selected based on their relevance in probe and epigenetic drug discovery that have attracted the attention to perform virtual screening [15, 16]. However, the SB-DFP is general and could be used for other targets. The epigenomics database used in this study contains compounds associations against 60 epigenetic targets. For our analysis, we selected the information for 28 targets for which there was at least 50 reported compounds with a potency of 10 µM or better. Table 1 summarizes the targets considered in this work that included bromodomain-containing proteins (BRD2, BRD3 and BRD4), histone acetyltransferases (CREBBP and EP300), DNA methyltransferase (DNMT1), histone lysine methyltransferase (EHMT2), histone deacetylases (HDAC1-HDAC11), lysine acetyltransferase (KAT2B), lysine demethylases (KDM1A and KDM4C), histone methyl-lysine binding proteins (L3MBTL1 and L3MBTL3), mitogen-activated protein kinase (MAP3K7), O-GlcNAcase (MGEA5), nuclear receptor coactivators with histone acetyltransferase activity (NCOA1 and NCOA3), and protein arginine methyltransferase (PRMT1). Table 1 also includes the number of compounds in each set (350 compounds on average with a maximum of 2740 for HDAC1). Note that SB-DFP could be applied to other data sets with larger number of compounds and their performance in, for instance, virtual screening, would need to be assessed in a case-by-case basis. It might be anticipated that the performance could be target-dependent as it happens in other virtual screening approaches.

Reference set

In this study, the All Clean subset from the ZINC12 database [17], with 16,403,844 unique compounds, was

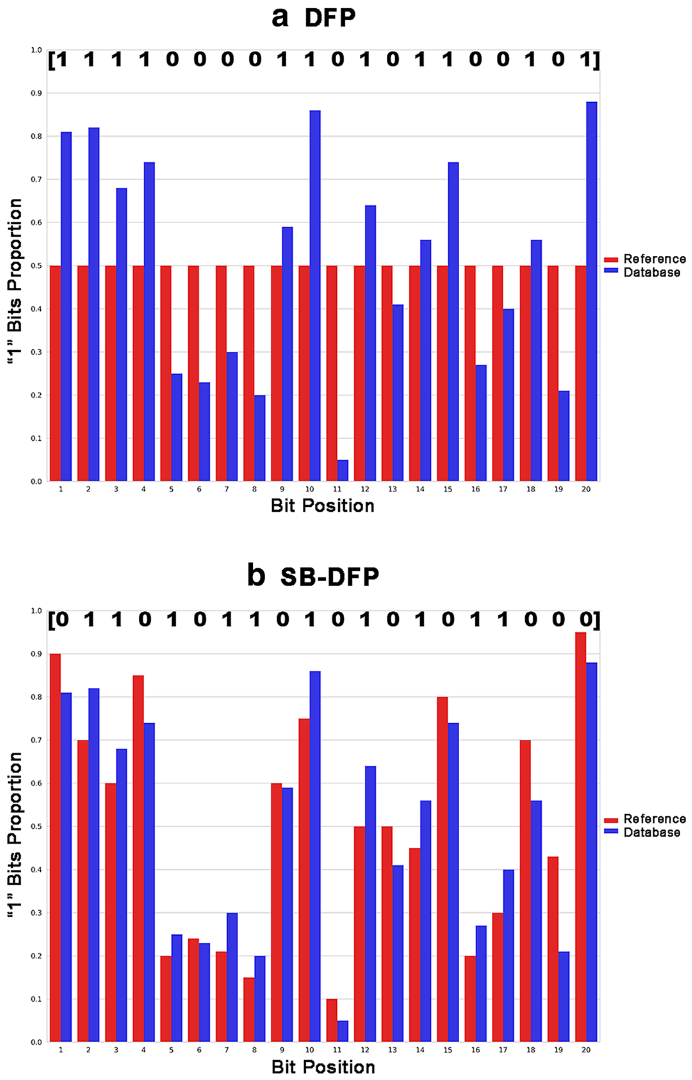


Fig. 1 Schematic representation of single fingerprints for a compound database and an hypothetical 20-bit fingerprint. The upper part of charts shows the binary representation of the generated single fingerprint: **a** database fingerprint (DFP) and **b** statistical-based database fingerprint (SB-DFP)

selected as starting point to build the reference set for SB-DFP calculations. We removed 21 compounds that could not be processed by the RDKit module for Python [18] and also 154 compounds present in the epigenomics database. The remaining molecules were randomly divided in two groups: one group with 1,000,000 compounds to be used as decoys in similarity searching (vide infra) and the second group with the remaining 15,403,690 molecules to be used as reference for SB-DFP calculations. We employed such database with more than 15 million compounds as a representative sample of the currently known chemical space of small molecules available in ZINC. We emphasize that SB-DFP could be implemented using other reference data sets.

Fingerprints

We selected two fingerprints to illustrate the applicability of the concept of SB-DFP: Molecular ACCess System (MACCS) keys (166-bit) [19] as a "low resolution" dictionary fingerprint, and Extended Connectivity Fingerprint diameter 4 (ECFP4) as a "high resolution" representation [20] in its folded version of 2048 bits. MACCS keys and ECFP4 were generated with RDKit.

Table 1 Selected datasets from the epigenomic database

Dataset	Number of compounds	Intra-set similarity median (Tc)		Average "1" bits		Number of "1" bits in DFP		Number of "1" bits in SB-DFP	
		MACCS[a]	ECFP4[b]	MACCS[a]	ECFP4[b]	MACCS[a]	ECFP4[b]	MACCS[a]	ECFP4[b]
BRD2	234	0.569	0.152	56.0	54.3	53	27	67	229
BRD3	246	0.573	0.153	56.6	54.6	53	26	73	231
BRD4	477	0.486	0.133	55.9	52.8	47	14	71	333
CREBBP	105	0.694	0.276	56.1	53.9	52	36	50	185
DNMT1	127	0.403	0.115	55.4	51.7	50	13	62	281
EHMT2	61	0.636	0.228	62.4	55.7	62	41	56	167
EP300	57	0.425	0.106	58.2	55.7	53	11	56	285
HDAC10	190	0.514	0.165	53.2	50.6	50	17	46	272
HDAC11	137	0.494	0.156	51.2	50.8	48	16	42	229
HDAC1	2740	0.453	0.149	53.2	51.4	51	15	63	499
HDAC2	767	0.447	0.149	50.3	48.4	46	13	53	336
HDAC3	669	0.474	0.147	52.6	50.3	49	13	54	356
HDAC4	452	0.427	0.135	50.4	46.4	42	10	49	248
HDAC5	112	0.455	0.153	47.3	44.1	39	13	26	176
HDAC6	1374	0.474	0.149	54.3	49.8	48	13	62	415
HDAC7	112	0.489	0.165	50.4	45.8	43	12	28	197
HDAC8	864	0.500	0.153	54.9	51.2	50	12	52	398
HDAC9	102	0.494	0.169	52.6	47.4	46	13	29	190
KAT2B	55	0.583	0.179	50.8	37.3	46	13	44	99
KDM1A	241	0.380	0.143	44.8	46.2	31	21	31	216
KDM4C	88	0.359	0.101	48.8	40.3	41	10	38	158
L3MBTL1	50	0.804	0.551	42.2	36.8	37	27	37	56
L3MBTL3	89	0.731	0.404	40.4	36.6	37	26	35	83
MAP3K7	96	0.539	0.137	57.1	60.5	59	35	45	190
MGEA5	67	0.683	0.316	54.2	39.6	48	19	42	126
NCOA1	51	0.350	0.105	45.5	43.3	34	11	18	132
NCOA3	157	0.368	0.109	47.7	44.6	39	10	26	166
PRMT1	61	0.395	0.076	53.0	53.5	41	9	40	239
Average	350	0.507	0.178	52	48	46	18	46	232

[a] MACCS keys 166-bit

[b] ECFP4 2048-bit

Binomial proportion comparisons

To perform the binomial proportion comparisons we employed a Z-test, as implemented in the statsmodels [21] module for Python. As can be found elsewhere [22], the proportion comparison relies on the calculation of a test statistic (called Z_{test}) defined as:

$$Z_{test} = \frac{p_t - p_r}{\sqrt{\boldsymbol{P}(1-\boldsymbol{P})\left(\frac{1}{n_t}+\frac{1}{n_r}\right)}}$$

where p_t and p_r are the proportions in which a given bit appears as "1" in the target and reference data sets for a total of n_t and n_r observations, respectively. $\boldsymbol{P}$ is the estimated true proportion of "1" bits considering both sample observations and it is calculated as:

$$\boldsymbol{P} = \frac{n_t p_t + n_r p_r}{n_t + n_r}$$

With the Z_{test} calculated and through the standard Normal distribution, the exact probability than the observed difference between proportion is due to random variation can be determined (the p value). So that the proportion difference is statistically significative if the p value is lower than the associated to the confidence level selected a priori. For example, for the bit 100 in MACCS fingerprint, the bit "1" occurrence in the reference set is

10,892,579 from 15,403,690 observations ($p_r = 0.707$). By selecting a confidence level of 99% (p value < 0.01) and doing the calculations one gets that for a target data set of 350 compounds, the bit occurrence must be equal or greater than 268 ($p_t = 0.766$, p value = 0.008) to be set as an “1” bit in the SB-DFP representation even when for a bit occurrence of 248 the proportion seems to be larger ($p_t = 0.708$, p value = 0.476). This example illustrates that a greater proportion of “1” in a given bit for the target data set in comparison to the reference data set does not necessarily implies that such bit will be set as “1” in the SB-DFP. In other words, the proportion difference must be “big enough”.

For this work we choose a confidence level of 99% (p value < 0.01) based on the average AUC values obtained from similarity searching for ECFP4 and MACCS keys at five different confidence levels (vide infra). For the sets of targets and the fingerprints explored, the best performing method is the one with a confidence level of 99% (Additional file 1: Table S1) and all further calculations and discussion are based on such method. Of note, other p values could be chosen for other targets and/or other fingerprints.

SB-DFP to study inter-data set relationships

To evaluate the performance of SB-DFP to capture the differences between data sets we calculated both, the classic DFP and the SB-DFP for each of the 28 targets. Both database fingerprints were constructed based on ECFP4 and MACCS keys fingerprints. Using the Tanimoto coefficient [23] and for each molecular fingerprint, we constructed the similarity matrices between epigenetic targets with three methodologies to calculate the similarity between pairs of targets: the median similarity between all-compound comparisons (ACC) in the data sets, the similarity between DFPs, and the similarity between SB-DFPs. This led to a total of six representations herein referred as ACC/MACCS, ACC/ECFP4, DFP/MACCS, DFP/ECFP4, SB-DFP/MACCS and SB-DFP/ECFP4. The range of similarity values for each representation was taken as a measure of its resolution. All six similarity matrices were transformed to their corresponding distance matrices based on the relationship (distance = 1 − similarity). The distance matrices were used as basis for hierarchical clustering with complete linkage to analyze the ability of the representations to recover the known relationships between epigenetic targets based on its sequence identity. Such ability was assessed by calculating the Adjusted Rand Index (ARI) of each clustering [24] at a level of 10 clusters. The ARI measures the similarity between a given clustering and a ground truth: an ARI value of 1 indicates that the clustering recovers the original groups and an ARI value of 0 indicates random assignations. As ground truth, we used the hierarchical clustering with complete linkage obtained from the distance form of the sequence identity matrix (shown as Additional file 1: Table S10) as obtained from the alignment with Clustal Omega [25] with default parameters for the 28 targets studied. Sequences for all targets were taken from the Universal Protein Knowledgebase (UniProt) [26]. In addition, the number of “1” bits present in each representation was calculated as an approach of the amount of information contained in each one.

SB-DFP as query for similarity searching

Previous studies have shown that using single fingerprint representation of compound databases as query yield better results in similarity searching than fingerprint representations of single compounds [5, 6]. However, when single fingerprint representations are compared with methods that use information for multiple compound in a database, such as k-nearest neighbors (k-NN) and binary kernel discrimination, the single fingerprint searches are outperformed [5]. In this work, we tested the performance of SB-DFP in similarity searching as compared to the classic DFP and 1-NN search strategies for both MACCS keys and ECFP4 fingerprints, methods such as binary kernel discrimination were not compared in this work given its reported lack of efficiency [5]. The Tanimoto coefficient was used as similarity measure, although other similarity metrics could be explored. For SB-DFP, five different confidence levels were tested for binomial proportion comparisons, here we report only the best performing one (99%), the rest are summarized in Additional file 1: Table S1.

Using an approach similar to the one reported by Heikamp et al. [27], from each of the 28 epigenetic targets, 100 random sets of 10 active compounds each were randomly selected and used as query. In each case, all remaining active compounds were added as active database of compounds (ADCs) to a database containing one million compounds randomly selected from the ZINC All Clean subset (vide supra), called the search set. For the searches involving DFP and SB-DFP, the 10 compounds used as query were employed to build the corresponding single fingerprint, which was compared against all compounds in the search set, leading directly to a single similarity value per compound. On the other hand, for 1-NN, each of the compounds in the search set was compared to the 10 compounds used as query, leading to 10 similarity values per compound, from which the highest value was taken. For each similarity search, the compound recovery rates (RR) were calculated in a target-specific selection over the number of available ADCs as a measure of early enrichment. Receiver operating characteristic (ROC) curves and ROC area under the curve (AUC) values were also computed.

Results and discussion

Bit proportions in the reference set

As detailed in the Methods section, 15,403,690 compounds from the ZINC All Clean subset were taken as a representative sample of the currently known chemical space of small molecules. For the complete data set, the frequency of each bit was calculated for ECFP4 and MACCS keys. The results are summarized in Additional file 1: Tables S2 and S3. Of note, only 43 out of 166 bits for MACCS keys and 12 out of 2048 bits for ECFP4 have frequencies over 0.5. This means that 43 and 12 bits of MACCS keys and ECFP4, respectively, are the most likely to appear in the DFP representation of any data set. Such bias is avoided in SB-DFP.

Compound data sets

For the 28 data sets studied in this work a total of six representations were generated for each set: the fingerprints for each compound, the single DFP, and SB-DFP, all based on ECFP4 and MACCS keys, respectively. Of note, the data sets representations based on DFP and SB-DFP have the advantage over "all-compounds" representation in that the speed of calculation is *NxM* times faster than doing pairwise comparisons with all compounds in a set (with *N* and *M* being the number of compounds in two data sets).

The median of the intra-set similarity for all compounds in each data set was computed with MACCS keys and ECFP4 and the results are summarized in Table 1. Overall, all 28 sets have structural diverse compounds with, for instance, maximum median MACCS keys similarity of 0.694 (average of 0.507) and maximum median ECFP4 similarity of 0.551 (average of 0.178).

Table 1 also reports the average number of "1" bits for all compounds, as well as the number of "1" bits in the DFP and SB-DFP, respectively. For both MACCS keys and ECFP4 fingerprints, DFP representation has, on average, number of "1" bits (46 and 18, respectively) lower than all-compounds representation (52 and 48, respectively) but higher than the number of bits with occurrence frequencies over 0.5 in the reference set (vide supra). As expected, DFP contains less information than the complete data set. However, DFP captures more features in the data set than expected according to the occurrence frequencies in the reference data set.

DFP/MACCS and SB-DFP/MACCS capture similar amount of information with an average number of "1" bits of 46. However, as shown in Table 1, there is a dramatic increase in the number of "1" bits for SB-DFP/ECFP4 as compared to DFP/ECFP4 (232 vs. 18). These results indicate that for the 28 data sets considered in this work, SB-DFP/ECFP4 captures a higher amount of specific structural features of the compounds.

Similarity matrices

The similarity matrices between epigenetic targets were calculated with three different approaches to calculate the similarity between pairs of targets: the median similarity of the all pairwise comparisons (e.g., all-compound comparisons) in the data sets (ACC), the similarity between their DFPs, and the similarity between their SB-DFPs, all based on MACCS keys and ECFP4 using the Tanimoto coefficient. As described in the Methods section, these representations are referred in this work as ACC/MACCS, ACC/ECFP4, DFP/MACCS, DFP/ECFP4, SB-DFP/MACCS, and SB-DFP/ECFP4. The six matrices are shown in Additional file 1: Tables S4–S9. Table 2 summarizes the maximum, minimum, average and range of Tanimoto similarity values for each similarity matrix. By using the median similarity between ACC in the data sets, the ranges are the smallest for MACCS keys and

Table 2 Range of Tanimoto similarity values in similarity matrices

Representation	MACCS keys (166-bit)				ECFP4 (2048-bit)			
	Minimum	Average	Maximum	Range	Minimum	Average	Maximum	Range
All compounds[a]	0.293	0.407	0.804	0.511	0.059	0.114	0.553	0.494
DFP	0.254	0.540	1.000	0.746	0.070	0.408	1.000	0.930
SB-DFP	0.050	0.342	1.000	0.950	0.011	0.185	1.000	0.989

[a] It should be noted that the comparisons involving the self-similarity of data sets does not reach a value of 1 and in some cases such self-similarity does not correspond to the highest value in the matrix row, that could be misinterpreted as the existence of pairs of databases more similar to each other than to themselves, which makes no sense. The matrices constructed by using DFP or SB-DFP do not present such problem, since when dealing with unique comparisons, a maximum of 1 is guaranteed for the diagonal of the matrix

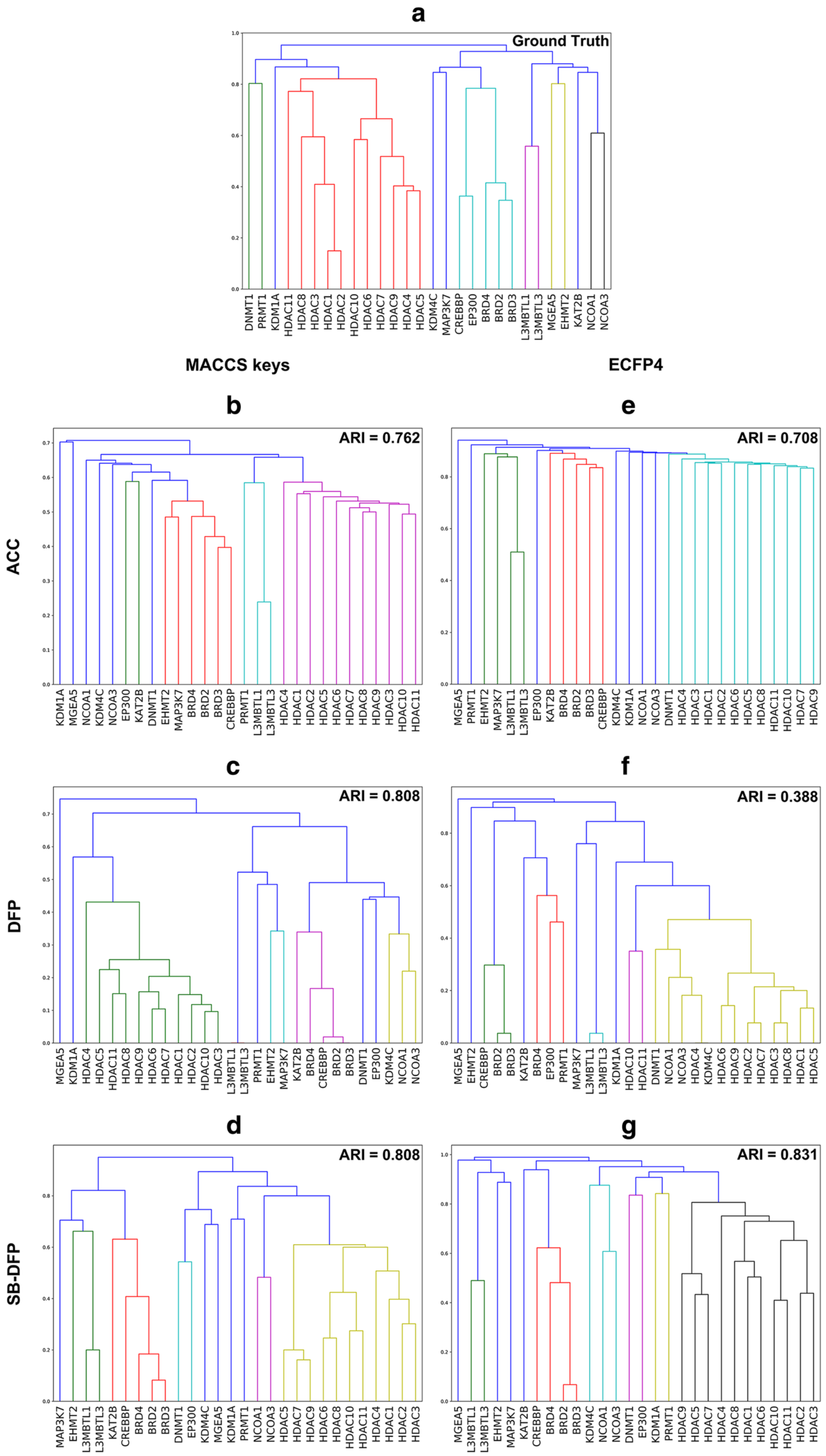


Fig. 2 Dendograms for hierarchical clustering of targets computed with different approaches based in two molecular fingerprints, MACCS keys and ECFP4. **a** The ground truth; **b**, **e** all-compound comparisons (ACC); **c**, **f** database fingerprint (DFP); **d**, **g** statistical-based database fingerprint (SB-DFP). The Adjusted Rand Index (ARI) of each clustering is indicated in each panel. See main text for details

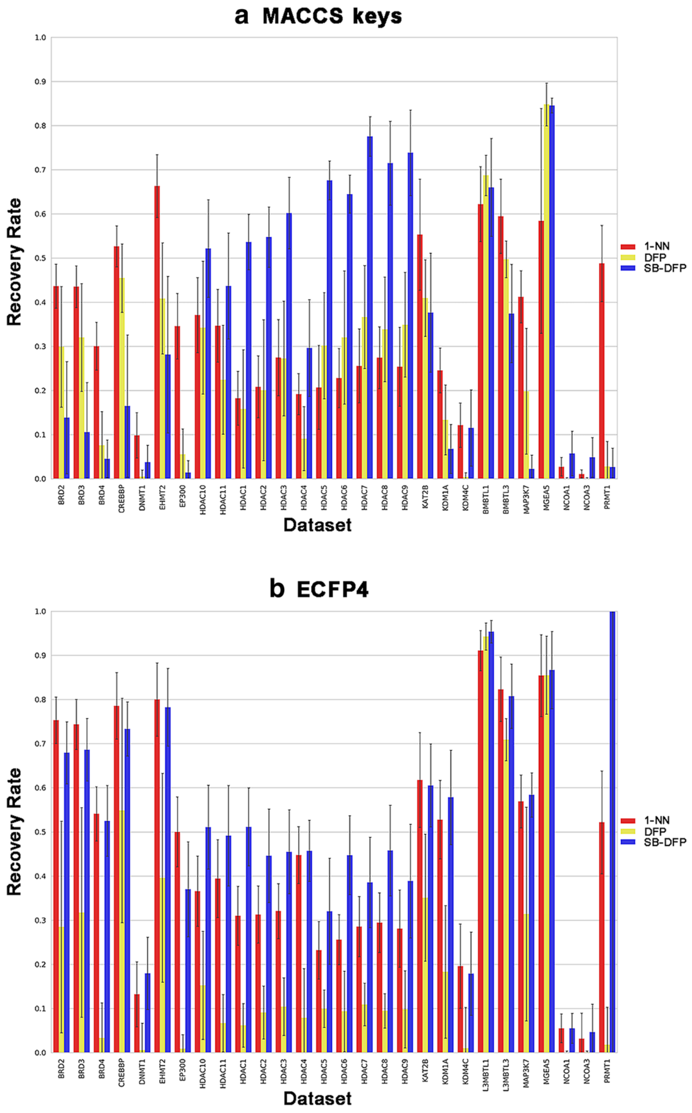


Fig. 3 Early enrichment performance of similarity searches. Average recovery rates (selection set size equal to the number of ADCs) for three search strategies over 28 epigenetic data sets are reported in a histogram representation for **a** MACCS keys and **b** ECFP4. Standard deviations are displayed as error bars

ECFP4 (i.e., 0.51 and 0.49, respectively). Table 2 shows that the similarity matrices constructed using SB-DFP present a broader range of values (0.950 and 0.989) than those constructed using DFP (0.746 and 0.930).

The SB-DFP matrices also have lower average similarities between data sets than the DFP matrices (0.540 vs. 0.342 for MACCS keys and 0.408 vs. 0.185 for ECFP4, respectively). Based on these results, the representation that captures better the differences between data sets is SB-DFP/ECFP4. This result agrees with the relative "higher resolution" of SB-DFP/ECFP4 i.e., higher number of "1" bits discussed above (Table 1).

SB-DFP to study inter-data set relationship

Figure 2 shows the dendrograms for each hierarchical clustering obtained with the corresponding distance matrices (vide supra). Analyzing the differences between data sets is not a trivial task and it is not straightforward evaluating the performance of a structural representation. In this work, we assessed the ability of the six representations listed above to recover the known relationships between epigenetic targets based on its sequence identity, using as metric the ARI at a level of 10 clusters and as ground truth the hierarchical clustering obtained from the distance form of the sequence

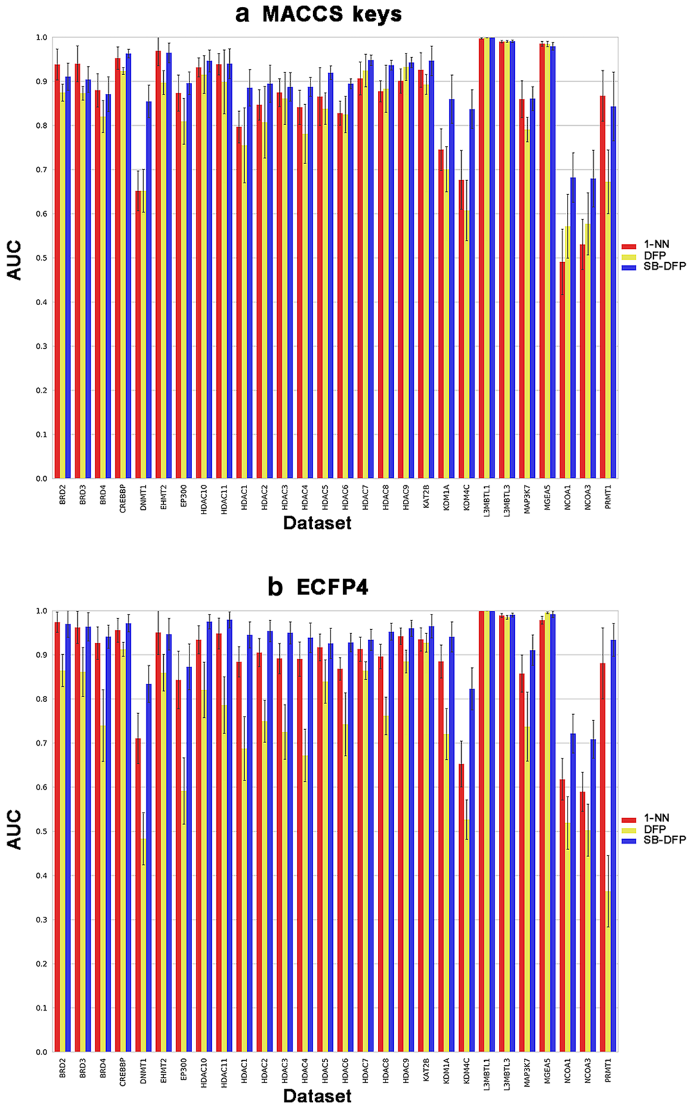


Fig. 4 General performance of similarity searches. Average AUCs for three search strategies over 28 epigenetic data sets are reported in a histogram representation for **a** MACCS keys and **b** ECFP4. Standard deviations are displayed as error bars

identity matrix (vide supra). The level of ten clusters was selected as ground truth given its recovery of four groups of epigenetic targets with known relationships: group 1 containing BRDs 2–4, CREBBP and EP300; group 2 containing HDACs 1–11; group 3 including L3MBTLs 1 and 3; and group 4 consisting of NCOAs 1 and 3. According to the results, the best performing methods were those based on the SB-DFP, with ARI values of 0.831 for SB-DP/ECFP4 and 0.808 for SB-DFP/MACCS. Methods based on ACC had worst but similar performances for both fingerprints with ARI values of 0.762 and 0.708 for ACC/MACCS and ACC/ECFP4 respectively. Finally, methods based on DFP had contrasting performances, being DFP/MACCS tied as the second best method with an ARI value of 0.808 and DFP/ECFP4 the worst of them with an ARI value of 0.388.

SB-DFP as template for similarity searching

All 28 epigenetic data sets were subjected to systematic fingerprint search calculations. To obtain statistically relevant data, from each data set, 100 compound reference sets of 10 compounds were randomly selected and used as query in six different representations: the fingerprints for each compound (1-NN), the DFP and the SB-DFP, the three of them based on ECFP4 and MACCS keys. For the

Table 3 Average recovery rates

Dataset	MACCS keys (166-bit)			ECFP4 (2048-bit)		
	1-NN	DFP	SB-DFP	1-NN	DFP	SB-DFP
BRD2	43.7 (5.0)	29.9 (13.7)	13.8 (12.8)	**75.4 (5.2)**	28.4 (24.2)	68.0 (7.1)
BRD3	43.5 (4.8)	32.0 (12.3)	10.6 (11.3)	**74.4 (5.7)**	31.9 (23.8)	68.7 (7.1)
BRD4	30.0 (5.4)	7.6 (7.7)	4.5 (4.3)	**54.1 (6.2)**	2.7 (4.7)	52.6 (8.1)
CREBBP	52.7 (4.7)	45.5 (7.8)	16.5 (16.2)	**79.0 (5.4)**	55.6 (25.0)	73.7 (4.2)
DNMT1	9.9 (5.2)	0.5 (1.5)	3.8 (3.9)	12.9 (5.7)	0.0 (0.0)	**17.7 (7.1)**
EHMT2	66.3 (7.1)	40.9 (12.6)	28.1 (17.8)	**80.1 (8.0)**	40.2 (23.5)	78.4 (8.3)
EP300	34.6 (7.5)	5.5 (5.8)	1.4 (2.7)	**50.2 (7.7)**	0.7 (2.8)	37.0 (10.8)
HDAC10	37.1 (8.6)	34.2 (15.1)	**52.2 (11.1)**	36.5 (8.0)	15.4 (12.3)	**51.1 (9.5)**
HDAC11	34.7 (8.3)	22.5 (12.4)	43.7 (12.1)	39.6 (8.8)	6.6 (6.4)	**49.3 (11.3)**
HDAC1	18.2 (6.1)	15.8 (13.5)	**53.7 (6.3)**	30.9 (6.7)	6.3 (5.1)	51.1 (9.0)
HDAC2	20.9 (7.0)	20.1 (16.1)	**54.8 (6.9)**	31.3 (6.5)	9.1 (6.0)	44.7 (10.6)
HDAC3	27.5 (8.7)	27.3 (13.1)	**60.2 (8.1)**	32.0 (6.2)	10.4 (6.6)	45.4 (9.6)
HDAC4	19.2 (4.7)	9.1 (7.3)	29.6 (11.0)	**44.9 (6.2)**	7.9 (11.2)	**45.8 (7.0)**
HDAC5	20.7 (9.6)	30.2 (12.1)	**67.6 (4.4)**	23.1 (6.4)	10.0 (4.3)	32.0 (12.1)
HDAC6	22.8 (6.7)	32.0 (15.1)	**64.5 (4.3)**	25.7 (5.8)	9.3 (9.1)	44.6 (9.0)
HDAC7	25.6 (8.4)	36.6 (11.7)	**77.6 (4.5)**	28.4 (6.7)	11.0 (4.9)	38.6 (10.4)
HDAC8	27.4 (7.0)	33.9 (11.9)	**71.5 (9.5)**	29.6 (6.9)	9.5 (3.9)	46.2 (9.8)
HDAC9	25.4 (9.0)	34.9 (11.9)	**73.9 (9.7)**	27.7 (7.5)	9.6 (8.7)	38.4 (13.0)
KAT2B	55.3 (12.7)	41.0 (8.7)	37.6 (13.5)	**61.8 (10.8)**	35.3 (14.1)	**60.4 (9.4)**
KDM1A	24.6 (5.1)	13.3 (8.0)	6.8 (5.6)	53.3 (8.4)	18.3 (15.1)	**58.4 (10.0)**
KDM4C	12.2 (5.1)	0.4 (1.0)	11.5 (8.7)	**18.9 (6.4)**	0.1 (0.3)	17.1 (5.8)
L3MBTL1	62.2 (8.5)	68.8 (4.6)	66.0 (11.1)	91.1 (4.6)	94.5 (1.8)	**95.5 (2.3)**
L3MBTL3	59.5 (8.5)	49.7 (4.2)	37.4 (11.2)	**82.8 (6.6)**	71.1 (4.5)	81.1 (6.8)
MAP3K7	41.2 (6.0)	19.8 (14.3)	2.2 (3.1)	56.6 (5.2)	31.1 (23.8)	**58.0 (4.0)**
MGEA5	58.5 (25.6)	84.8 (4.9)	84.6 (1.7)	86.3 (3.5)	86.4 (2.0)	**87.6 (2.2)**
NCOA1	2.7 (2.1)	0.0 (0.2)	**5.7 (5.1)**	**5.5 (3.3)**	0.1 (0.3)	**5.5 (3.4)**
NCOA3	1.1 (0.9)	0.1 (0.2)	**4.9 (4.5)**	2.6 (1.4)	0.1 (0.3)	**4.1 (2.5)**
PRMT1	48.8 (8.7)	2.8 (5.7)	2.7 (4.3)	52.8 (10.5)	1.0 (3.8)	**55.3 (12.1)**
Average	**33.1 (19.6)**	**26.4 (22.8)**	**35.3 (29.1)**	**46.0 (25.9)**	**21.5 (28.4)**	**50.2 (23.8)**

The best performing methods for each dataset are shown in bold. If there were no significative difference between two or more methods, all of them are marked. Standard deviations are shown in parentheses

six search strategies, Figs. 3 and 4 show the results of the RR and AUC, respectively. In terms of early enrichment, by using MACCS keys as molecular representation, the SB-DFP approach outperformed the other methods with an average RR of 35.3%, followed by 1-NN (33.1%) and DFP (26.4%). Similar trends were obtained using ECFP4, being the average RRs 50.2%, 46% and 21.5 for SB-DFP, 1-NN, and DFP respectively. Regarding to the global performance, the tendency was identical. The best performing method in both cases was SB-DFP, for MACCS keys with an average AUC of 0.898, followed by 1-NN and DFP with average AUCs of 0.853 and 0.824 respectively and for ECFP4 with average AUCs of 0.926, 0.882 and 0.755 for SB-DFP, 1-NN and DFP respectively. These results revealed the anticipated differences between high- and low-resolution fingerprints, since ECFP4 achieved higher RRs and AUCs for 1-NN searches, while for the single fingerprint searches the higher values corresponded to the most populated representations in terms of number of bits "1" (MACCS keys for DFP and ECFP4 for SB-DFP).

The results also illustrated the general data set-dependence of the similarity searching performance and the good success rates achieved for 2D fingerprint methods, since the best performing search strategy for each data set obtained an average RR of at least 50% in 22 of 28 cases, and an average AUC larger than 0.7 in all of them. By analyzing the individual performances, according to RRs (Table 3), SB-DFP was the best method for 17 cases, from which eight were based on MACCS keys, seven

Table 4 Average areas under ROC curves

Dataset	MACCS keys (166-bit)			ECFP4		
	1-NN	DFP	SB-DFP	1-NN	DFP	SB-DFP
BRD2	0.938 (0.035)	0.875 (0.019)	0.911 (0.031)	**0.974 (0.023)**	0.865 (0.037)	0.970 (0.030)
BRD3	0.940 (0.041)	0.873 (0.015)	0.905 (0.029)	**0.962 (0.037)**	0.861 (0.056)	**0.964 (0.032)**
BRD4	0.880 (0.038)	0.821 (0.036)	0.871 (0.040)	0.927 (0.037)	0.740 (0.082)	**0.941 (0.026)**
CREBBP	0.953 (0.025)	0.924 (0.008)	0.963 (0.009)	0.956 (0.027)	0.913 (0.016)	**0.972 (0.020)**
DNMT1	0.652 (0.045)	0.652 (0.049)	**0.855 (0.037)**	0.711 (0.058)	0.484 (0.060)	0.834 (0.042)
EHMT2	**0.969 (0.033)**	0.897 (0.027)	0.965 (0.023)	0.951 (0.050)	0.860 (0.042)	0.947 (0.036)
EP300	0.874 (0.041)	0.810 (0.052)	**0.896 (0.026)**	0.843 (0.066)	0.592 (0.076)	0.873 (0.052)
HDAC10	0.932 (0.022)	0.916 (0.043)	0.946 (0.025)	0.934 (0.032)	0.821 (0.063)	**0.975 (0.016)**
HDAC11	0.939 (0.024)	0.899 (0.073)	0.940 (0.034)	0.948 (0.035)	0.786 (0.065)	**0.979 (0.018)**
HDAC1	0.797 (0.036)	0.755 (0.085)	0.886 (0.041)	0.884 (0.035)	0.688 (0.073)	**0.945 (0.030)**
HDAC2	0.847 (0.035)	0.808 (0.081)	0.895 (0.042)	0.905 (0.032)	0.750 (0.048)	**0.954 (0.024)**
HDAC3	0.875 (0.032)	0.862 (0.059)	0.888 (0.032)	0.892 (0.035)	0.725 (0.062)	**0.950 (0.025)**
HDAC4	0.841 (0.039)	0.781 (0.067)	0.888 (0.021)	0.890 (0.039)	0.672 (0.060)	**0.939 (0.034)**
HDAC5	0.866 (0.066)	0.838 (0.036)	**0.920 (0.016)**	0.917 (0.030)	0.840 (0.049)	**0.926 (0.035)**
HDAC6	0.828 (0.028)	0.825 (0.042)	0.895 (0.011)	0.868 (0.026)	0.743 (0.072)	**0.928 (0.021)**
HDAC7	0.907 (0.037)	0.925 (0.037)	**0.948 (0.012)**	0.913 (0.027)	0.864 (0.020)	0.934 (0.024)
HDAC8	0.878 (0.024)	0.883 (0.054)	0.937 (0.011)	0.896 (0.028)	0.762 (0.043)	**0.953 (0.019)**
HDAC9	0.901 (0.028)	0.933 (0.031)	0.943 (0.012)	0.942 (0.019)	0.885 (0.026)	**0.960 (0.018)**
KAT2B	0.926 (0.039)	0.893 (0.022)	0.947 (0.033)	0.935 (0.027)	0.928 (0.022)	**0.965 (0.027)**
KDM1A	0.745 (0.048)	0.701 (0.051)	0.860 (0.055)	0.885 (0.038)	0.721 (0.058)	**0.941 (0.034)**
KDM4C	0.677 (0.067)	0.608 (0.069)	**0.837 (0.044)**	0.653 (0.052)	0.527 (0.045)	0.823 (0.048)
L3MBTL1	0.997 (0.001)	0.999 (0.000)	**1.000 (0.000)**	1.000 (0.000)	1.000 (0.000)	**1.000 (0.000)**
L3MBTL3	0.990 (0.003)	**0.991 (0.002)**	**0.991 (0.003)**	0.989 (0.005)	0.985 (0.004)	**0.990 (0.005)**
MAP3K7	0.860 (0.042)	0.791 (0.028)	0.861 (0.027)	0.858 (0.042)	0.738 (0.079)	**0.911 (0.035)**
MGEA5	0.985 (0.005)	0.985 (0.006)	0.979 (0.009)	0.979 (0.009)	**0.996 (0.002)**	0.992 (0.007)
NCOA1	0.491 (0.074)	0.572 (0.073)	0.682 (0.056)	0.618 (0.047)	0.519 (0.060)	**0.722 (0.044)**
NCOA3	0.530 (0.057)	0.577 (0.071)	0.680 (0.064)	0.590 (0.045)	0.503 (0.059)	**0.709 (0.043)**
PRMT1	0.867 (0.058)	0.673 (0.072)	0.843 (0.078)	0.881 (0.081)	0.365 (0.081)	**0.934 (0.037)**
Average	**0.853 (0.132)**	**0.824 (0.129)**	**0.898 (0.082)**	**0.882 (0.113)**	**0.755 (0.171)**	**0.926 (0.077)**

The best performing methods for each dataset are shown in bold. If there were no significative difference between two or more methods, all of them are marked. Standard deviations are shown in parentheses

based on ECFP4 and two without significative difference between molecular fingerprints. The second best method was 1-NN with eight favorable cases by using ECFP4. For three data sets there was not significative difference between SB-DFP and 1-NN (Fig. 3). Additionally, the DFP representation was not the best performing method for any of the data sets studied.

According to the AUCs values (Table 4), the best performing method for 23 data sets was SB-DFP, from which four were based on MACCS keys, 17 based on ECFP4 and two without significative difference between fingerprints. The overall second-best approach was 1-NN with better predictions for two data sets (one for each molecular fingerprint). In general, DFP had lower AUCs values as compared to the other two search methods (Table 4).

Remarkably, the search method based on SB-DFP could be applied in at least 20 out of the 28 data sets studied leading to the best RRs, with the additional advantage over 1-NN that the speed of calculation is N times faster (with N being the number of compounds used as query). This fact is because the number of comparisons needed for the screening is always equal to the number of compounds in the screened database in contrast to 1-NN, where this number scale with the number of compounds used as query.

Conclusions and perspectives

Here we presented the statistical-based database fingerprint (SB-DFP) as a novel general approach to generate single fingerprints of compound databases based on binomial proportion comparisons. In this work we shown its implementation for two molecular fingerprints (e.g., ECFP4 and MACCS keys) and one specific reference set (e.g., ZINC). However, the applicability of SB-DFP can be extended to any binary fingerprint and to other reference sets. Using as a case study a recently published set of 28 epigenetic compound sets with therapeutic relevance, we illustrate the application of SB-DFP to capture the inter-data sets relationships and to perform similarity searching. For the data sets explored in this work the largest set has 2740 compounds (as deposited in ChEMBL) but SB-DFP could be applied to other larger compound data with relevance in drug or probe discovery. Despite the fact that no quantitative analysis was performed in terms of speed of calculation, it is clear that single fingerprint approaches to represent compound databases are faster because they depend on single rather than multiple comparisons.

Two major perspectives of the SB-DFP approach are application in high throughput virtual screening and target identification. To these ends, studies involving different molecular fingerprints, target-associated compound sets and reference data sets would be required, as well as exhaustive validations of their performance. Part of this work in ongoing and will be reported in due course.

Additional file

Additional file 1: Table S1. Average similarity searching performances for SB-DFP constructed at different confidence levels. **Table S2.** "1" bits count for 15,403,690 compounds taken from ZINC using MACCS keys. **Table S3.** "1" bits count for 15,403,690 compounds taken from ZINC using ECFP4. **Table S4.** Similarity matrix of compound data sets computed as the median Tanimoto coefficient between its compounds using MACCS keys. **Table S5.** Similarity matrix of compound data sets computed as Tanimoto coefficient between its DFP based on MACCS keys. **Table S6.** Similarity matrix of compound data sets computed as Tanimoto coefficient between its SB-DFP based on MACCS keys. **Table S7.** Similarity matrix of compound data sets computed as the median Tanimoto coefficient between its compounds using ECFP4. **Table S8.** Similarity matrix of compound data sets computed as Tanimoto coefficient between its DFP based on ECFP4. **Table S9.** Similarity matrix of compound data sets computed as Tanimoto coefficient between its SB-DFP based on ECFP4. **Table S10.** Sequence identity matrix of targets computed from Clustal Omega alignments.

Abbreviations

ACC: all compound comparisons; ADC: active database of compounds; ARI: Adjusted Rand Index; AUC: area under the curve; DFP: database fingerprint; ECFP4: extended connectivity fingerprint of diameter four; k-NN: k-nearest neighbors; MACCS: molecular access system; ROC: receiver operating characteristic; RR: recovery rate; SB-DFP: statistical-based database fingerprint; UniProt: Universal Protein Knowledgebase.

Authors' contributions

All authors designed the study. NS-C performed the calculations. All authors wrote read and approved the final manuscript.

Acknowledgements

This work was funded by the *Programa de Apoyo a Proyectos de Investigación e Innovación Tecnológica* (PAPIIT) IA203718, Facultad de Química, UNAM. NS-C is thankful to CONACyT for the granted scholarship number 335997.

Competing interests

The authors declare that they have no competing interests.

Availability of data and materials

All datasets reported in the current study as well as the code implemented for calculations of DFP and SB-DFP are available in the GitHub repository: https://github.com/DIFACQUIM/SB-DFP

Funding

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

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 8 September 2018 Accepted: 14 November 2018
Published online: 22 November 2018

References

1. Cereto-Massagué A, Ojeda MJ, Valls C et al (2015) Molecular fingerprint similarity search in virtual screening. Methods 71:58–63. https://doi.org/10.1016/j.ymeth.2014.08.005
2. Muegge I, Mukherjee P (2016) An overview of molecular fingerprint similarity search in virtual screening. Expert Opin Drug Discov 11:137–148. https://doi.org/10.1517/17460441.2016.1117070
3. Heikamp K, Bajorath J (2012) Fingerprint design and engineering strategies: rationalizing and improving similarity search performance. Future Med Chem 4:1945–1959. https://doi.org/10.4155/fmc.12.126
4. Shemetulskis NE, Weininger D, Blankley CJ et al (1996) Stigmata: an algorithm to determine structural commonalities in diverse datasets. J Chem Inf Comput Sci 36:862–871. https://doi.org/10.1021/ci950169+
5. Hert J, Willett P, Wilton DJ et al (2004) Comparison of fingerprint-based methods for virtual screening using multiple bioactive reference structures. J Chem Inf Comput Sci 44:1177–1185. https://doi.org/10.1021/ci034231b
6. Duan J, Dixon SL, Lowrie JF, Sherman W (2010) Analysis and comparison of 2D fingerprints: insights into database screening performance using eight fingerprint methods. J Mol Graph Model 29:157–170. https://doi.org/10.1016/j.jmgm.2010.05.008
7. Fernández-De Gortari E, García-Jacas CR, Martinez-Mayorga K, Medina-Franco JL (2017) Database fingerprint (DFP): an approach to represent molecular databases. J Cheminform 9:1–9. https://doi.org/10.1186/s13321-017-0195-1
8. Xue L, Stahura FL, Godden JW, Bajorath J (2001) Fingerprint scaling increases the probability of identifying molecules with similar activity in virtual screening calculations. J Chem Inf Comput Sci 41:746–753. https://doi.org/10.1021/ci000311t
9. Xue L, Godden JW, Stahura FL, Bajorath J (2003) Profile scaling increases the similarity search performance of molecular fingerprints containing numerical descriptors and structural keys. J Chem Inf Comput Sci 43:1218–1225. https://doi.org/10.1021/ci030287u
10. Xue L, Stahura FL, Bajorath J (2004) Similarity search profiling reveals effects of fingerprint scaling in virtual screening. J Chem Inf Comput Sci 44:2032–2039. https://doi.org/10.1021/ci0400819
11. Wang Y, Bajorath J (2008) Bit silencing in fingerprints enables the derivation of compound class-directed similarity metrics. J Chem Inf Model 48:1754–1759. https://doi.org/10.1021/ci8002045
12. Lounkine E, Hu Y, Batista J, Bajorath J (2009) Relevance of feature combinations for similarity searching using general or activity

class-directed molecular fingerprints. J Chem Inf Model 49:561–570. https://doi.org/10.1021/ci800377n

13. Medina-Franco JL (2016) Epi-informatics: discovery and development of small molecule epigenetic drugs and probes. Academic Press, Cambridge. https://doi.org/10.1016/C2014-0-03789-6
14. Naveja JJ, Medina-Franco JL (2017) Insights from pharmacological similarity of epigenetic targets in epipolypharmacology. Drug Discov Today 23:141–150. https://doi.org/10.1016/j.drudis.2017.10.006
15. Lu W, Zhang R, Jiang H et al (2018) Computer-aided drug design in epigenetics. Front Chem 6:57. https://doi.org/10.3389/fchem.2018.00057
16. Prieto-Martinez FD, Medina-Franco JL (2018) Charting the Bromodomain BRD4: towards the identification of novel inhibitors with molecular similarity and receptor mapping. Lett Drug Des Discov 15:1002–1011. https://doi.org/10.2174/1570180814666171121145731
17. Irwin JJ, Sterling T, Mysinger MM et al (2012) ZINC: a free tool to discover chemistry for biology. J Chem Inf Model 52:1757–1768. https://doi.org/10.1021/ci3001277
18. RDKit: open-source cheminformatics. http://www.rdkit.org. Accessed Nov 2018.
19. 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. https://doi.org/10.1021/ci010132r
20. Rogers D, Hahn M (2010) Extended-connectivity fingerprints. J Chem Inf Model 50:742–754. https://doi.org/10.1021/ci100050t
21. Seabold S, Perktold J (2010) Statsmodels: econometric and statistical modeling with python. In: Proceedings of 9th python in science conference, pp 57–61
22. LeBlanc D (2004) Statistics: concepts and applications for science. Jones & Bartlett Publishers, Sudbury
23. Bajusz D, Rácz A, Héberger K (2015) Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? J Cheminform 7:20. https://doi.org/10.1186/s13321-015-0069-3
24. Hubert L, Arabie P (1985) Comparing partitions. J Classif 2:193–218. https://doi.org/10.1007/BF01908075
25. Sievers F, Wilm A, Dineen D et al (2014) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol 7:539. https://doi.org/10.1038/msb.2011.75
26. The UniProt Consortium (2017) UniProt: the universal protein knowledgebase. Nucl Acids Res 45:D158–D169. https://doi.org/10.1093/nar/gkw1099
27. Heikamp K, Bajorath J (2011) Large-scale similarity search profiling of ChEMBL compound data sets. J Chem Inf Model 51:1831–1839. https://doi.org/10.1021/ci200199u

RSC Advances

REVIEW

View Article Online
View Journal | View Issue

Cite this: *RSC Adv.*, 2017, 7, 54153

Open chemoinformatic resources to explore the structure, properties and chemical space of molecules

Mariana González-Medina,[a] J. Jesús Naveja,[ab] Norberto Sánchez-Cruz[a] and José L. Medina-Franco*[a]

Received 27th October 2017
Accepted 21st November 2017

DOI: 10.1039/c7ra11831g

rsc.li/rsc-advances

New technologies are shaping the way drug discovery data is analyzed and shared. Open data initiatives and web servers are assisting the analysis of the large amounts of data that we are now able to produce. The final goal is to accelerate the process of moving from new data to useful information that could lead to treatments for human diseases. This review discusses open chemoinformatic resources to analyze the diversity and coverage of the chemical space of screening libraries and to explore structure–activity relationships of screening data sets. Free resources to implement workflows and representative web-based applications are emphasized. Future directions in this field are also discussed.

1. Introduction

During the past few years, there has been an important increase in open data initiatives to promote the availability of free research-based tools and information.[1] While there is still some resistance to open data in some chemistry and drug discovery fields, the availability of information has been a necessity for other research fields such as genomics, proteomics and bioinformatics. The Human Genome Project was paramount to the open-source movement in proteomics and genomics, demonstrating that a global community can be more successful and efficient in analyzing data than a single individual can.[2]

Computer-aided drug discovery has a large impact for the pharmaceutical industry by helping during the drug development process to reduce time and costs, in order to achieve a desired result. However, researchers from the pharmaceutical and medicinal chemistry fields often lack training on informatics. The creation of free and easy to use chemoinformatic tools for drug development will help investigators avoid having to spend time acquiring programming and development skills, in the already complex and multidisciplinary field of drug discovery. At the same time, the resources will assist research teams to focus on solving problems that are specific to their fields of expertise. In this context, chemoinformatics has an important role helping to mine the chemical space of the almost infinite number of organic drug-like molecules available for drug discovery. The outcome allows researchers to find connections between biological activities, ligands and proteins.[3]

Herein we review representative chemoinformatic tools essential to explore the structure, chemical space and properties of molecules. The review is focused on recent and representative free web-based applications. We also discuss KNIME as an open resource broadly used in chemoinformatics for automatization of data analysis. The review is organized in eight major sections. After this introduction, open sources of chemical biology data are discussed. Section 3 discusses online servers for the generation of molecular properties, diversity analysis, and visualization of the chemical space. The next section focuses on web-based application to predict ADME and toxicity properties, which are essential in drug discovery programs. Section 5 presents online applications to analyze structure–activity relationships (SAR) and structure–multiple activity relationships (SmAR). The section after that discusses web-servers aim to assist drug discovery and development efforts focused on a particular disease or target family. Section 7 covers open resources to implement workflows for data analysis. In contrast to most web-based applications discussed in Sections 3–6, the workflows presented in Section 7 can be highly customizable by the user. The last section presents Conclusions and future directions.

2. Open chemical biology data

Essential to medicinal chemistry and drug discovery is the availability to generate and retrieve relevant experimental data of screened compounds. Relevant experimental data implies curated information with enough quality for later SAR analysis. There is a large and still growing amount of molecules with bioactivity data available for the public domain, which is summarized in Table 1.

[a] Department of Pharmacy, School of Chemistry, Universidad Nacional Autónoma de México, Avenida Universidad 3000, Mexico City 04510, Mexico. E-mail: medinajl@unam.mx; jose.medina.franco@gmail.com; Tel: +52-55-5622-3899 ext. 44458

[b] PECEM, School of Medicine, Universidad Nacional Autónoma de México, Avenida Universidad 3000, Mexico City 04510, Mexico

Table 1 Open chemical biology data sets

Database	Data	General information	Ref.
ChEMBL	In total, there are >1.6 million distinct compound structures, with 14 million activity values from >1.2 million assays. These assays are mapped to ∼11 000 targets, including 9052 proteins	ChEMBL is an open large-scale bioactivity database. It contains data from the medicinal chemistry literature, deposited data sets from neglected disease screening, crop protection data, drug metabolism and disposition data, bioactivity data from patents, the annotation of assays and targets using ontologies, the inclusion of targets and indications for clinical candidates, addition of metabolic pathways for drugs and calculation of structural alerts	4
PubChem	It contains the information of 92 058 388 compounds; 1 252 809 bioassays; 2 395 818 tested compounds; 170 RNAi bioactivities; 233 516 687 bioactivities; 10341 protein targets; 22 104 gene targets	PubChem is a public chemical information repository in the National Center for Biotechnology Information. It provides information on the biological activities of small molecules. PubChem is organized as three linked databases within the NCBI's Entrez information retrieval system. These are PubChem substance, PubChem compound, and PubChem BioAssay. PubChem also provides a fast chemical similarity search tool	5,6
Binding Database	It holds about 1.1 million measured protein-small molecule affinities, involving about 490 000 small molecules and several thousand proteins	Binding DB is a publicly accessible database of experimental protein-small molecule interaction data primarily from scientific articles and US patents	7
CARLSBAD	The 2012 release of CARLSBAD contains 439 985 unique chemical structures, mapped onto 1 420 889 unique bioactivities	The CARLSBAD database has been developed as an integrated resource, focused on high-quality subsets from several bioactivity databases, which are aggregated and presented in a uniform manner, suitable for the study of the relationships between small molecules and targets	8
ExCAPE-DB	In total there are 998 131 unique compounds and 70 850 163 structure–activity relationship (SAR) data points covering 1667 targets	ExCAPE-DB is a large public chemogenomics dataset based on the PubChem and ChEMBL databases. Large scale standardization (including tautomerization) of chemical structures was performed using open source chemoinformatics software	9
BRENDA	BRENDA is the main collection of enzyme functional data available to the scientific community	Currently BRENDA contains manually curated data for 82 568 enzymes and 7.2 million enzyme sequences from UniProt	10
DrugCentral	Over 14 000 numeric values are captured covering 2190 human and non-human targets for 1792 unique active pharmaceutical ingredients	DrugCentral is a comprehensive drug information resource for FDA drugs and drugs approved outside US. The resources can be searched using drug, target, disease, and pharmacologic action terms	11
Probes & drugs portal	It contains 31 182 compounds, 4727 targets, and 114 825 bioactivities	The probes & drugs portal is a public resource joining together focused libraries of bioactive compounds (probes, drugs, specific inhibitor sets, *etc.*) with commercially available screening libraries	12
DrugBank	It contains 9591 drug entries including 2037 FDA-approved small molecule drugs, 241 FDA-approved biotech (protein/peptide) drugs, 96 nutraceuticals and over 6000 experimental drugs. Additionally, 4661 non-redundant protein sequences are linked to these drug entries	The DrugBank database is a unique bioinformatics and chemoinformatics resource that combines detailed drug (*i.e.* chemical, pharmacological and pharmaceutical) data with comprehensive drug target (*i.e.* sequence, structure, and pathway) information	13
repoDB	repoDB spans 1571 drugs and 2051 United Medical Language System (UMLS) indications disease concepts, accounting for 6677 approved and 4123 failed drug-indication pairs	repoDB contains a standard set of drug repositioning successes and failures that can be used to fairly and reproducibly benchmark computational repositioning methods. repoDB data was extracted from DrugCentral and ClinicalTrials.gov	14
PharmGKB	It has over 5000 genetic variants annotations, with over 900 genes related to drugs and over 600 drugs related to genes	PharmGKB captures pharmacogenomic relationships in a structured format so that it can be searched, interrelated, and displayed according to the researchers' interests. The knowledge base is valuable both to the researcher who is interested in a specific single nucleotide polymorphism and its influence on a particular drug treatment and to the researcher interested in a disease or drug and looking for candidate genes which may affect disease progression or drug response	15

Of note, although the availability of this data is important to build new models and make *in silico* predictions, the data and content in these databases is rather heterogeneous.

Perhaps the most common and widely used databases are ChEMBL, which contains 1.6 million distinct compounds and 14 million activity values,[4] PubChem[5,6] with more than 93 million compounds and more than 233 million bioactivities, and Binding Database with 490k small molecules and 1.1 million measured protein-small molecule affinities.[7]

Other resources are CARLSBAD, a bioactivity database with 435 343 compounds and 932 852 bioactivities. The advantage of CARLSBAD is that only one activity value of a given type (Ki, EC_{50}, *etc.*) is stored for a given structure–target pair.[8] ExCAPE-DB is a comprehensive chemogenomics dataset with 998 131 compounds and 70 850 163 biological activity data.[9] BRENDA is an enzyme information system of enzyme and enzyme–ligand information obtained from different sources; functional and structural data of more than 190 000 enzyme ligands are stored within this system.[10] The knowledge on bioactivity could help to identify potential targets for a specific molecule.

DrugCentral is a database that integrates structure, bioactivity, regulatory, pharmacologic actions and indications for active pharmaceutical ingredients approved by FDA and other regulatory agencies.[11] The probes and drugs portal is a public resource putting together focused libraries of bioactive compounds (877 probes and 12 190 drugs) with commercially available screening libraries. The rationale behind it is to reflect the current state of bioactive compound space and to enable its exploration from different points of view.[12] Finding new uses for old drugs could be economically advantageous, therefore the development of databases like DrugCentral and probes and drugs will be beneficial for polypharmacology.[16]

3. Online servers for exploring chemical space

The concept of chemical space can be understood in a simplistic manner as the number of possible molecules to be considered when searching for new drugs, the knowledge and understanding of this space is of great relevance in drug discovery, several approaches used for its analysis have been reported extensively for many authors.[17–19] The chemical space can be divided in two main groups: the known chemical space, that considers the organic molecules reported thus so far, which are mostly covered by the resources discussed in the previous section, and the unknown chemical space, larger by tens of orders of magnitude compared to the first group and refers to molecules that have been never synthesized yet. Several advances and applications on the enumeration of those virtual molecules are discussed in other works.[20,21]

One of the central points to the concept of chemical space is molecular representation *i.e.*, the set of descriptors used to define the space of the chemicals that will be analyzed. A second major point is the visual representation and mining of that space, *e.g.*, analysis of the diversity and coverage. Those aspects are important to consider when dealing with the analysis and interpretation of data, because distinct approaches may lead to representations that in most cases are not comparable to each other and the best one is usually defined by the he nature of the data analyzed. Web servers to explore chemical space usually incorporate one or more of the following operations: calculation of descriptors, visualization, and diversity analysis. Table 2 summarizes recent online servers for generating and mining the chemical space of compound databases using different approaches. Representative servers are further commented in this section.

ChemMine is an online portal with five main application domains: compounds visualization, similarity quantification, a search toolbox to retrieve similar compounds from PubChem, clustering, data visualization and molecular properties calculation.[22]

ChemBioServer is a free-web based tool that can aid researchers on compound filtering and clustering. Compounds that survive the filtering process can be visualized using molecular properties and principal component analysis.[23]

ChemDes is a free web-based platform for the calculation of molecular descriptors and fingerprints. It contains more than 3679 molecular descriptors that are divided into 61 logical blocks. In addition, ChemDes provides 59 types of molecular fingerprint systems.[26]

BioTriangle can calculate a large number of molecular descriptors of individual molecules, structural and physicochemical features of proteins and peptides from their amino acid sequences, and composition and physicochemical features of DNAs/RNAs from their primary sequences.[25]

FAF-Drugs3, now FAF-Drugs4, is a web server that applies an enhanced structure curation procedure that filters compounds based on physicochemical properties, ADMET rules and generally unwanted molecules also known as pan assay interference compounds (PAINS).[24] This server can be used to generate and analyze ADMET-relevant chemical spaces.[19]

The visualization of the chemical space of molecular databases has been proved to be relevant to measure molecular diversity and biological properties. webMolCS is a web-based interface to visualize sets of user-defined molecules in 3D chemical spaces, using different molecular fingerprints and selecting subsets.[27]

The visualization of the chemical space can offer a good idea on how diverse the datasets are, however, since the diversity criteria depends on the molecular representation employed, a tool to compute different diversity metrics would be useful to researchers with different backgrounds. Platform for Unified Molecular Analysis (PUMA) is a web server developed to visualize the chemical space and measure the molecular properties and structural diversity of datasets.

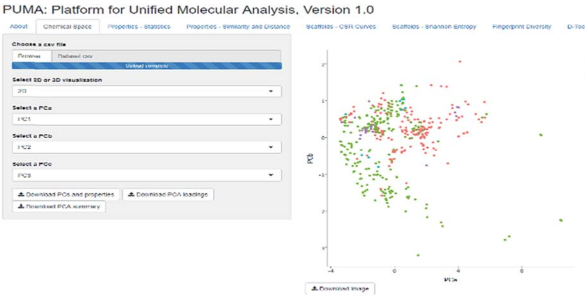
PUMA addresses the issue of the dependence of chemical space on structure representation. In this server the user can analyze a user-supplied data set using molecular scaffolds, properties of pharmaceutical relevance and fingerprints of different design. Fig. 1 illustrates a screenshot of the server PUMA. The figure exemplifies the analysis done with the chemical space tab available in the main top menu of the application.

Table 2 Recent online tools developed for mining chemical and target spaces

Tool	Primary use	Functions	Implementation	Ref.
ChemMine	Set of chemoinformatics and data mining tools	Compounds visualization, similarity quantification, a search toolbox to retrieve similar compounds from PubChem, clustering and data visualization and molecular properties calculation	The server integrates over 30 chemoinformatics and data mining tools, being ChemMineR, an R package that integrates Open Babel and JOELib functionalities, one of the most important. The web interface was written in Python using Django web framework	22
ChemBioServer	Mining and filtering chemical compound libraries	2D and 3D molecule visualization, compound filtering: by toxicity, repeated compounds and steric clashes, similarity clustering using molecular fingerprints, data mining, graphical representation and visualization	The application back-end was developed in R programming language, while the front-end is implemented with PHP. 2D/3D display of compounds is accomplished with JChemPaint and Jmol respectively. Compound fingerprints are generated with Open Babel	23
FAF-Drugs4	Mining and filtering chemical compound libraries	Filters compounds based on physicochemical properties, ADMET rules and pan assay interference compounds (PAINS)	The application consists of a set of seven object-oriented Python modules embedded in the RPBS′ Mobyle framework. Each compound processed by FAF-Drugs3 is represented as a molecular object importing methods from the Open Babel toolkit through its Python wrapper Pybel which allows to access to the OpenBabel C++ library	24
BioTriangle	Molecular properties and molecular fingerprints calculation	Computes descriptors that describe chemical features, protein features and DNA/RNA features	The application was implemented in an open source Python framework (Django) for the Graphical User Interface (GUI) and MySQL for data retrieval. The main calculation procedures and transaction processing procedures are written in Python language	25
ChemDes	Molecular properties and molecular fingerprints calculation	Computes more than 3679 molecular descriptors and provides 59 types of molecular fingerprint	The application back-end was developed with Python. Django was chosen as a high-level Python web framework for web interface	26
webMolCS	A web-based interface for visualizing sets of up to 5000 user-defined molecules in 3D chemical spaces and selecting subsets	Computes molecular fingerprints that are used to generate 3D chemical spaces using either principal component analysis (PCA) or similarity mapping (SIM)	This web server was developed using JavaScript and the JChem java chemistry library from ChemAxon	27
Platform for Unified Molecular Analysis (PUMA)	Chemical space and analysis of chemical diversity	Chemical space, molecular properties diversity, scaffold diversity and structural diversity	The application back-end was developed in R programming language: plotly for the interactive plots, rcdk for the chemoinformatic analysis and Shiny for the user interface	28
Consensus diversity plots	Global diversity visualization	Plots to visualize simultaneously several metrics of diversity and classify data sets	The application back-end was developed in R programming language. Shiny package was used for the user interface	29
SwissADME	Molecular and physicochemical properties. Identifies PAINS	Web tool enables the computation of physicochemical, pharmacokinetic, drug-like and related parameters	The website was written in HTML, PHP5, and JavaScript, whereas the backend of computation was mainly coded in Python 2.7	30
MetaTox	Calculation of probability for generated metabolites. Prediction of LD_{50} values	Prediction of xenobiotic's metabolism and calculation toxicity of metabolites based on the structural formula of chemicals	The website uses MySQL server to store the data and PHP and HTML codes to implement the main interface. The Python script is used to generate the prediction and data processing	31
SOMP	Prediction is based on PASS (Prediction of Activity Spectra for Substances) technology and labelled multilevel neighborhoods of atom descriptors	Prediction for drug-like compounds that are metabolized by the main CYP isoforms and UGT	The website uses MySQL server to store the data and PHP and HTML codes to implement the main interface. The Python script is used to produce independent sub-processes to generate input to the prediction program and data processing	32

Table 2 (*Contd.*)

Tool	Primary use	Functions	Implementation	Ref.
CarcinoPred-EL	Computes ensemble machine learning methods to predict carcinogenicity and identify structural features related to carcinogenic effects	This web server computes molecular fingerprints and uses ensemble machine learning methods to discover potential carcinogens	This website uses PaDEL-descriptors[33] to compute the molecular fingerprints and the R package caret for the machine learning methods	34
Pred-Skin	Binary QSAR models	Web-based and mobile application for the identification of potential skin sensitizers	The app is encoded using Flask, uWSGI, Nginx, Python, RDKit, scikit-learn and JavaScript	35
Activity Landscape Plotter	Activity landscape modeling and structure–activity relationships	Structure Activity Similarity (SAS) maps, Structure Activity Landscape Index (SALI) and Dual Activity Difference (DAD) maps	The application back-end was developed in R programming language. Rcdk and Shiny packages are used for the chemoinformatic analysis and user interface, respectively	36
ChemSAR	Structure preprocessing, molecular descriptor calculation, data preprocessing, feature selection, model building and prediction, model interpretation and statistical analysis	This web site computes the standardization of chemical structure representations, 783 1D/2D molecular descriptors and ten types of fingerprints for small molecules, the filtering methods for feature selection, the generation of predictive models	Python/Django and MySQL was used for server-side programming, and HTML, CSS, JavaScript was employed for the web interface	37
Chembench	Chembench is a tool for data visualization, create and validate predictive quantitative structure–activity relationship models and virtual screening	Chembench supports the following chemoinformatics data analysis tasks: Dataset creation, dataset visualization, modeling, model validation and virtual screening	Chembench is a Java-based system. The front end of the website uses Java Server Pages with JavaScript. The struts 2 framework provides the interface between data on the JSPs and Java objects	38



PUMA: Platform for Unified Molecular Analysis, Version 1.0

Choose a csv file: Dataset csv

Upload complete

Select 2D or 3D visualization: 2D

Select a PCa: PC1

Select a PCb: PC2

Select a PCc: PC3

Fig. 1 Screenshot of the Platform for Unified Molecular Analysis (PUMA) server. PUMA is focused on the analysis of chemical space diversity and coverage of compound data sets. The example illustrates the application of the chemical space tab to the visual representation of the chemical space of four data sets using principal component analysis. PUMA is freely available at http://www.difacquim.com/d-tools/.

Molecular diversity of compound data sets can be evaluated employing molecular scaffolds, structural fingerprints and physicochemical properties. Consensus Diversity Plot (CDP) is a novel method to represent in low dimensions the diversity of chemical libraries considering simultaneously multiple molecular representations and to facilitate the classification of data sets into diverse or not diverse.[29] A recent application of CDPlots is the analysis and quantification of the global diversity of 354 natural products from Panama. The diversity of those compounds was compared against the diversity of natural products from Brazil, natural and semi-synthetic molecules used in high-throughput screening, and compounds used in Traditional Chinese Medicine.[39] The CDPlots rapidly led to the conclusion that natural products from Panama have a large scaffold diversity as compared to other databases.

4. Servers to predict ADME and toxicity properties

Computational methods are being used to filter and select compounds based on different molecular characteristics that are considered to be relevant to predict the drug-likeness of molecules. Without the aid of computational methods, the drug development process would be more time-consuming and less efficient, however, it is important to mention that the filtering rules employed by these methods are not absolute answers to the problem and that experimental confirmation is compulsory. A number of compounds fail during clinical phases due to poor pharmacokinetic and safety properties, therefore, the growing number of public and commercial *in silico* tools to predict ADMET (absorption, distribution, metabolism, excretion and toxicity) parameters is not surprising.

SwissADME is a web tool to compute fast but robust predictive models for physicochemical properties, pharmacokinetics, drug-likeness and identifying PAINS.[30] Other web servers used to predict toxicity are based on the prediction of metabolites formation. This is the case for MetaTox, which can also be used to predict toxicity endpoints,[31] and SOMP, a web-service for the prediction of metabolism by human cytochrome P450.[32] Among various toxicological endpoints, the carcinogenicity of potential drugs is of interest because of its serious effects on human health. In general, the carcinogenic potential of a compound is evaluated using animal models that are time-consuming, expensive, and ethically concerning. The use of computational approaches such as CarcinoPred-El, which predicts carcinogenicity based on chemical structure properties, is an appealing alternative. CarcinoPred-El uses different molecular fingerprints and ensemble machine learning methods to predict the carcinogenicity of diverse organic compounds.[34]

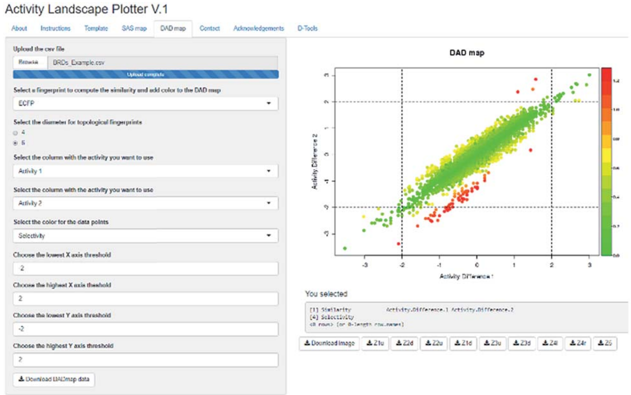
The use of animals for cosmetic experiments is forbidden in Europe, therefore there is a strong need to develop alternative tests to evaluate skin sensitization. Pred-Skin is an app developed to predict the skin sensitization potential of chemicals based on binary QSAR models of skin sensitization potential from human (109 compounds) and murine local lymph node assay (LLNA, 515 compounds) data.[35]

5. Online applications for exploring SAR and SmAR

The increasing availability of chemical biology data (discussed in Section 2) allows researchers to create models capable of predicting the potential chemical and biological behavior of compounds. There is a limited number of public tools available that are able to create models to understand the advantages and disadvantages behind the SAR concept, those models are highly dependent on the quality and quantity of data available, so these models should be selected based on the problem of interest and when available, oriented approaches could be the best choice, but the results obtained by any methodology must be interpreted carefully.

Most of the web-sites developed to perform SAR analysis are focused on QSAR models (Table 2). This is the case of ChemSAR and Chembench. Both are web-based platforms to generate SAR and QSAR classification models employing machine learning methods.[37,38]

Activity Landscape Plotter is an R-based web tool developed to analyze SAR using the concept of activity landscape modeling. The objective of activity landscape modeling is to explore the relationship between structure similarity and activity similarity (or potency difference) of screening data sets.[40,41] There are a number of numerical and visual methods useful for activity landscape modeling. In particular, Activity Landscape Plotter generates structure–activity similarity, dual-activity difference maps and identifies activity cliffs in a data set with biological activity.[36] Dual-activity difference maps are particularly attractive to analyze SAR of data sets with activity data for two biological endpoints. Therefore, these maps are tools to explore SmARs. Fig. 2 shows a screenshot of the Activity Landscape Plotter. It is illustrated the functionality Dual-Activity Difference (DAD) functionality available in the main menu of the server.



Activity Landscape Plotter V.1

Upload the csv file: BRDs_Example.csv

Select a fingerprint to compute the similarity and add color to the DAD map: ECFP

Select the diameter for topological fingerprints
○ 4
● 6

Select the column with the activity you want to use: Activity 1

Select the column with the activity you want to use: Activity 2

Select the color for the data points: Selectivity

Choose the lowest X axis threshold: -2

Choose the highest X axis theshold: 2

Choose the lowest Y axis threshold: -2

Choose the highest Y axis threshold: 2

You selected

```
[1] Similarity          Activity.Difference.1 Activity.Difference.2
[4] Selectivity
<0 rows> (or 0-length row.names)
```

Fig. 2 Screenshot of the Activity Landscape Plotter server. This server is focused on the analysis of structure–activity relationships of compound data sets. The screenshot illustrates the generation of the Dual Activity Difference (DAD) map for a data set of compounds tested with two biological endpoints. Full description of the server and access to the example data set are freely available at http://www.difacquim.com/d-tools/.

6. Disease and target oriented web-servers

There is an increasing need to develop effective chemogenomic tools focused on integrating the large and growing amount of data available for specific health conditions. Multifactorial diseases that involve many genes, proteins and their interactions would be easier to study with the aid of web servers with databases that integrate and validate reported active compounds, molecular mechanisms and genetic association. This information could be easily reused to accelerate the discovery of novel compounds.

There are a number of servers that are focused on specific target families or diseases. These are summarized in Table 3. For complex diseases such as Alzheimer and cancer, useful web servers containing information regarding important targets and their ligands. AlzPlatform[42] and AlzhCPI[43] are web tools implemented for target identification, polypharmacology and virtual screening of active compounds for the treatment of Alzheimer disease. CDRUG,[44] CancerIN,[45] and CanSAR[46] are web servers developed to predict the anticancer activity of compounds. All these disease-oriented web tools contain valuable information such as genes, related proteins, drugs approved and in clinical trials, compounds associated with biological activity, as well as information on biological assays.

Similar web servers have been implemented for specific targets. Kinase and GPCR SARfari are chemogenomic tools implemented on ChEMBL to incorporate and link GPCR and kinase sequences, structures, compounds and screening data.[47] Other web servers such as KIDFamMap were developed to design selective kinase inhibitors.[49] This has been a challenging task given the evolutionary conserved ATP binding site where

Table 3 Servers focused on mining chemical and target spaces of target families or diseases

Tool	Primary use	General approach	Implementation	Ref.
AlzPlatform	Web tool implemented for target identification and polypharmacology analysis for Alzheimer disease research	Assembled with Alzheimer disease-related chemogenomics data records. Uses TargetHunter and/or HTDocking programs for identification of multitargets and polypharmacology analysis and also for screening and prediction of new Alzheimer disease active small molecules	AlzPlatform was constructed based on the molecular database prototype CBID, 8, 9 with a MySQL database and an apache web server. OpenBabel10 is the search engine for chemical structures. The web interface is written in PHP language	42
AlzhCPI	This server will facilitate target identification and virtual screening of active compounds for the treatment of Alzheimer disease	AlzhCPI predicts chemical–protein interactions based on multitarget quantitative structure–activity relationships (mt-QSAR) using naive Bayesian and recursive partitioning algorithms	The web server was designed using HTML and CSS technology	43
Kinase SARfari	This is an integrated chemogenomics workbench focused on kinases. The system incorporates and links kinase sequence, structure, compounds and screening data	Kinase SARfari data is accessible *via*: compound-similarity and substructure searching, target keyword and sequence similarity searching. Provides target and screening data through compound initiated queries	The ChEMBL web services are written in Python programming language within Django software framework	47
KIDFamMap	First tool to explore kinase-inhibitor families (KIFs) and kinase-inhibitor-disease (KID) relationships for kinase inhibitor selectivity and mechanisms	This tool includes 1208 KIFs, 962 KIDs, 55 603 kinase-inhibitor interactions (KIIs), 35 788 kinase inhibitors, 399 human protein kinases, 339 diseases and 638 disease allelic variants. KIDFamMap searches the kinase candidates (K′) with significant sequence similarity (*E*-values $\leq e^{-10}$) using BLASTP[48] and also searches the compound candidates (I′) with significant topology similarity ($\geq$0.6) using atom pairs and moiety composition from the annotated KII database ($\leq$10 μM)	Not reported	49
GLIDA	This web server provides interaction data between GPCRs and their ligands, along with chemical information on the ligands, as well as biological information regarding GPCRs	GLIDA includes a variety of similarity search functions for the GPCRs and for their ligands. Thus, GLIDA can provide correlation maps linking the searched homologous GPCRs (or ligands) with their ligands (or GPCRs)	GLIDA was constructed on the LAMP (Linux, Apache, MySQL and PHP) platform	50
GPCR SARfari	GPCR SARfari is an integrated chemogenomics research and discovery workbench for class A G protein coupled receptors	GPCR data is accessible *via* compound-similarity and substructure searching, target keyword and sequence similarity searching. Provides target and screening data through compound initiated queries	The ChEMBL web services are written in Python programming language within Django software framework	47
CancerIN	The web server uses machine learning and potency score based methods to classify compounds as anticancer and non-anticancer	This server provides various facilities that includes; virtual screening of anticancer molecules, analog based drug design, and similarity with known anticancer molecules	CancerIN was built using python scripts	45
CDRUG	CDRUG is a web server for predicting anticancer efficacy of chemical compounds	CDRUG uses a novel molecular description method (relative frequency-weighted fingerprint) to implement the compound 'fingerprints'. Then, a hybrid score was calculated to measure the similarity between the query and the active compounds. Finally, a confidence level (*P*-value) is calculated to predict whether the query compounds have, or do not have, the activity of anticancer	CDRUG employs both Python and Java to implement prediction of anticancer activity. Pybel is used to calculate the daylight fingerprint and use jCompoundMapp to calculate the kernel fingerprint	44
CanSAR	Tool to identify biological annotation of a target, its structural characterization, expression levels and protein	A large set of descriptors is calculated for each of the compounds to enable clustering of compounds into chemically related groups. Bemis and Murcko	CanSAR is running on an Apache web server implemented in PHP, JavaScript, Perl and Java. Chemical compound search and	46

Table 3 (*Contd.*)

Tool	Primary use	General approach	Implementation	Ref.
	interaction data, as well as suitable cell lines for experiments, potential tool compounds and similarity to known drug targets	frameworks are calculated for all compounds. The interface allows users to rapidly obtaining biological and chemical annotation together with druggability considerations, explore genomic variation and gene-expression data, identify relevant cell lines for experiments, and tool compounds for analysis	handling is supported by the Accelrys direct cartridge. The data processing pipelines are written in Perl, Python and Java and utilize OpenBabel, CDK and Pipeline Pilot	
HEMD	HEMD provides a central resource for the display, search, and analysis of the structure, function, and related annotation for human epigenetic enzymes and chemical modulators focused on epigenetic therapeutics	User may paste a SMILES or sketch a potential epigenetic compound. Submitting the query launches a structure similarity search tool in HEMD. In addition to these structure similarity searches, the "Modulator search" utility also supports compound searches on the basis of physicochemical properties and chemical formulas	Not reported	51

the majority of inhibitors are expected to bind. GLIDA is a public GPCR- related chemical genomics database, it provides chemical information on the ligands as well as biological information regarding GPCRs or G-protein coupled receptors, which represent one of the most important families of drug targets in pharmaceutical development.[50]

Epigenetics became of great importance for researchers when it was discovered that gene function could be altered by more than just changes in sequence. Today a number of diseases have been linked to amplification, mutation, and other alterations of epigenetic enzymes. Therefore, analyzing the most appropriate epigenetic enzymes involved in different diseases is a prerequisite for epigenetic therapeutics. HEMD is a web server that provides the utilities to display, search and analyze the structure, function and related annotation of human epigenetic enzymes and chemical modulators focused on epigenetic therapeutics.[51]

7. Data automatization with customizable workflows

In addition of web servers that are being increasingly used by experts and non-experts in chemoinformatics, there are open source applications that enable the generation of workflows and highly facilitate the automatization of data analysis. Among the advantages of these workflows is their customizability and adaptability to meet specific needs. KNIME is perhaps the most widely used such environment that is open access, and it is further described in this section.

KNIME's modular workflow design, along with its ability to automatically parallelize many operations, free distribution, and simplicity to communicate analysis pipelines, has made it widely successful in diverse areas of analytics. It is also quite flexible and allows integration of different software and tools.[52] For a detailed explanation of the "workflow" concept, as well as other software following this approach, see the review by Tiwari and Sekhar.[53] In the following subsections, the issues that can be addressed through chemistry applications or plugins implemented in KNIME are presented.

Data curation

It has not escaped the attention of chemoinformaticians that there is a vital necessity to produce reliable libraries prior to computational modeling.[54–56] Therefore, there are emerging several tools useful for processing and assessing chemical data (*e.g.*, parsing molecules, removing mixtures, and salts, optimizing pH and pK_a, standardizing chemotypes, managing tautomers, standardizing synonyms, and visualizing chemical graphs).[54] KNIME includes plugins able to perform these operations. Some of these are open source (*e.g.*, RDKit, Indigo, CDK), while others are commercial, though available at no additional cost to anyone holding a license for the standard software (*e.g.*, Schrödinger, MOE, ICM, ChemAxon).

A prior step to data curation involves, of course, reading a chemical database. There are many kinds of files in which chemical information may be stored, including CSV, SDF, SQL and XML. KNIME provides extensions able of reading most, if not all, of them. Regarding data curation pipelines, a recent publication by Gally *et al.* proposed a workflow for preliminary molecule preparation in KNIME.[57] Also, a useful and comprehensive tutorial for KNIME application into chemical data curation has been recently published elsewhere.[58]

Chemical properties and calculations

A variety of chemical features can be assessed through the KNIME chemoinformatics extensions mentioned above, such as physicochemical (*e.g.*, atomic molecular weight, SlogP, topological polar surface area, number of hydrogen bond acceptors and donors, rotatable bonds) and complexity (*e.g.*, fraction of sp^3 atoms, number of chiral atoms) descriptors, enumeration of

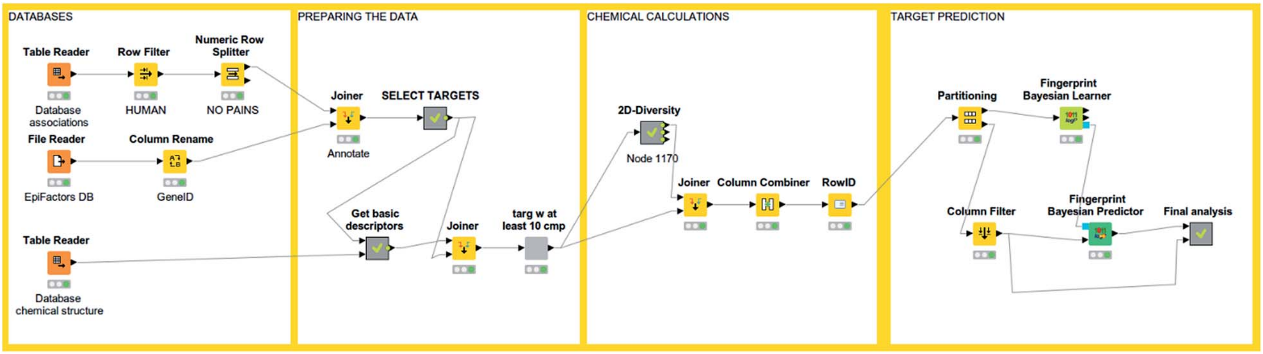


Fig. 3 An example of KNIME workflow for reading a chemical dataset and performing target prediction.

heteroatoms, a wide variety of chemical fingerprints, similarity calculations, virtual screening, R group decomposition and so forth. Also, tautomer lists, 3D functionalities such as 3D optimization, conformer generation and 3D similarity assessment are available in both free and commercial extensions. Docking is available mostly from commercial packages (GLIDE, ICM, MOE, Schödinger, *etc.*), although using AutoDock within KNIME is also an option.[59] Of note, 3D-e-Chem-VM, a recently developed application, integrates KNIME with public domain resources for analyzing protein–ligand interaction data. Its tools aid in virtual screening, metabolism prediction and rational ligand design in kinases and G-coupled protein receptors.[60]

Machine learning and SAR analysis

An interesting feature from KNIME is the incorporation of scalable machine learning. Some of these algorithms perform virtual screening by similarity searching or naïve Bayesian models with some options given, but mostly predetermined (see Fig. 3). Nonetheless, an option to enhance flexibility in KNIME workflows is to integrate scripts of programming languages with libraries specialized in machine learning (such as R and Python). Murcko scaffolds can be computed as well, followed by enrichment factor calculations.[61] There are even specific nodes for studying activity cliffs.[59] Notably, deep learning nodes have been recently incorporated.[62]

Examples of applications and a published KNIME workflow

In this section we describe two applications of KNIME to chemoinformatics. A more comprehensive review by Mazanetz *et al.* has been published, including also applications for data analysis applied to next generation sequencing and high throughput screening.[59]

PAINs filter workflow. Identification of PAINs (pan assay interference compounds) is becoming increasingly relevant, as they are thought (not without controversy)[63] to have higher rates of false-positives and unspecific promiscuity in screening studies.[64] Therefore, for many screening purposes it is widely preferred to sort them out, or at least identify them. Saubern *et al.* made available a KNIME workflow for identifying PAINS, after adequate molecule preprocessing.[65] They incorporated a previously published list of structural features intended to identify PAINS,[66] converted it to SMARTS format and used them to iteratively search through a chemical library of 10 000 compounds. The algorithm outputs a file with structures that do not match any of the features, as well as and another file with structures that match, along with the labels of the matching PAINS features. They compared the results of using Indigo or RDKit KNIME nodes for substructure search *versus* the hits from the original reference,[66] finding a higher overlap when Indigo nodes were used.

Rule of 0.5 of an approved drug's metabolite-likeness. Given prior insights that metabolites and approved drugs share chemical features,[67] O'Hagan *et al.* evaluated this hypothesis using KNIME nodes.[68] They pre-processed DrugBank approved drugs database and a human metabolites chemical database, calculated MACCS-166 bits fingerprints, and then evaluated the similarity among both datasets. They discovered that most (∼90%) of the approved drugs have a Tanimoto similarity of 0.5 of higher to their 'nearest' metabolite. Therefore, they suggested a '0.5 metabolite-likeness rule' that characterizes post marketed drugs.

8. Conclusions and future directions

The amount of information in drug discovery continues to increase rapidly. This is true for both the size of the screening libraries and the biological activity data. Therefore, the increasing amount of information *i.e.*, big data (particularly in the public domain), has boosted the development of tools for the comprehensive assessment of the coverage and diversity of the chemical space of compound libraries. Likewise, there is a need to develop automated applications for the rapid exploration of SAR and SmARTs, and to simplify the communication of the results across research teams. There are numerous chemoinformatic resources available to implement protocols that analyze different aspects of chemical space and SAR/SmART. These resources are being implemented in open web servers or workflows. These tools benefit not only chemoinformaticians but also to members of the multidisciplinary teams working on drug discovery projects that are non-experts or lack time to generate their own code or workflows from scratch. It is anticipated that these tools will continue to evolve and improve. Importantly, it is desirable that the easy-to-use web server applications do not become black boxes. It is of great importance that the user is fully aware of the calculations that are done, in order to fully maximize the interpretation of the results and that he/she is aware of the approximation and eventual limitations of the application or workflow. It is also expected a continuous development of web servers dedicated to explore the SAR and chemical space of a disease or target family. The improvement and refinement of these servers will certainly benefit from the constant increase of chemical biology information available in the public domain.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

We thank funding from the 'Programa de Apoyo a la Investigación y el Posgrado' (PAIP) 50009163, Facultad de Química, UNAM. MG-M thanks CONACyT-FUNED for the scholarship number 2017-000001-02EXTF-00177. JJN and NS-C are thankful to CONACyT for the granted scholarships number 622969 and 335997, respectively.

References

1 M. Allarakhia, *Expert Opin. Drug Discovery*, 2014, **9**, 459–465.

2 Toronto International Data Release Workshop Authors, *Nature*, 2009, **461**, 168.

3 K. Hasegawa and K. Funatsu, *Mol. Inf.*, 2014, **33**, 749–756.

4 A. Gaulton, A. Hersey, M. Nowotka, A. P. Bento, J. Chambers, D. Mendez, P. Mutowo, F. Atkinson, L. J. Bellis, E. Cibrián-Uhalte, M. Davies, N. Dedman, A. Karlsson, M. P. Magariños, J. P. Overington, G. Papadatos, I. Smit and A. R. Leach, *Nucleic Acids Res.*, 2017, **45**, D945–D954.

5 Y. Wang, S. H. Bryant, T. Cheng, J. Wang, A. Gindulyte, B. A. Shoemaker, P. A. Thiessen, S. He and J. Zhang, *Nucleic Acids Res.*, 2017, **45**, D955–D963.

6 S. Kim, P. A. Thiessen, E. E. Bolton, J. Chen, G. Fu, A. Gindulyte, L. Han, J. He, S. He, B. A. Shoemaker, J. Wang, B. Yu, J. Zhang and S. H. Bryant, *Nucleic Acids Res.*, 2016, **44**, D1202–D1213.

7 M. K. Gilson, T. Liu, M. Baitaluk, G. Nicola, L. Hwang and J. Chong, *Nucleic Acids Res.*, 2016, **44**, D1045–D1053.

8 S. L. Mathias, J. Hines-Kay, J. J. Yang, G. Zahoransky-Kohalmi, C. G. Bologa, O. Ursu and T. I. Oprea, *Database*, 2013, **2013**, bat044.

9 J. Sun, N. Jeliazkova, V. Chupakhin, J.-F. Golib-Dzib, O. Engkvist, L. Carlsson, J. Wegner, H. Ceulemans, I. Georgiev, V. Jeliazkov, N. Kochev, T. J. Ashby and H. Chen, *J. Cheminf.*, 2017, **9**, 41.

10 A. Chang, I. Schomburg, S. Placzek, L. Jeske, M. Ulbrich, M. Xiao, C. W. Sensen and D. Schomburg, *Nucleic Acids Res.*, 2015, **43**, D439–D446.

11 O. Ursu, J. Holmes, J. Knockel, C. G. Bologa, J. J. Yang, S. L. Mathias, S. J. Nelson and T. I. Oprea, *Nucleic Acids Res.*, 2017, **45**, D932–D939.

12 C. Skuta, M. Popr, T. Muller, J. Jindrich, M. Kahle, D. Sedlak, D. Svozil and P. Bartunek, *Nat. Methods*, 2017, **14**, 759–760.

13 V. Law, C. Knox, Y. Djoumbou, T. Jewison, A. C. Guo, Y. Liu, A. Maciejewski, D. Arndt, M. Wilson, V. Neveu, A. Tang, G. Gabriel, C. Ly, S. Adamjee, Z. T. Dame, B. Han, Y. Zhou and D. S. Wishart, *Nucleic Acids Res.*, 2014, **42**, D1091–D1097.

14 A. S. Brown and C. J. Patel, *Sci. Data*, 2017, **4**, 170029.

15 C. F. Thorn, T. E. Klein and R. B. Altman, in *Pharmacogenomics: Methods and Protocols*, ed. F. Innocenti and R. H. N. van Schaik, Humana Press, Totowa, NJ, USA, 1st edn, 2013, ch. 20, vol. 1015, pp. 311–320.

16 A. Cereto-Massagué, M. J. Ojeda, C. Valls, M. Mulero, G. Pujadas and S. Garcia-Vallve, *Methods*, 2015, **71**, 98–103.

17 C. M. Dobson, *Nature*, 2004, **432**, 824–828.

18 J.-L. Reymond, L. Ruddigkeit, L. Blum and R. van Deursen, *Wiley Interdiscip. Rev.: Comput. Mol. Sci.*, 2012, **2**, 717–733.

19 J. L. Medina-Franco, in *Diversity-Oriented Synthesis*, ed. A. Trabocchi, John Wiley & Sons, Inc., Hoboken, NJ, USA, 1st edn, 2013, ch. 10, vol. 1, pp. 325–352.

20 J.-L. Reymond and M. Awale, *ACS Chem. Neurosci.*, 2012, **3**, 649–657.

21 J.-L. Reymond, *Acc. Chem. Res.*, 2015, **48**, 722–730.

22 T. W. H. Backman, Y. Cao and T. Girke, *Nucleic Acids Res.*, 2011, **39**, W486–W491.

23 E. Athanasiadis, Z. Cournia and G. Spyrou, *Bioinformatics*, 2012, **28**, 3002–3003.

24 D. Lagorce, L. Bouslama, J. Becot, M. A. Miteva and B. O. Villoutreix, *Bioinformatics*, 2017, **33**, 3658–3660.

25 J. Dong, Z.-J. Yao, M. Wen, M.-F. Zhu, N.-N. Wang, H.-Y. Miao, A.-P. Lu, W.-B. Zeng and D.-S. Cao, *J. Cheminf.*, 2016, **8**, 34.

26 J. Dong, D.-S. Cao, H.-Y. Miao, S. Liu, B.-C. Deng, Y.-H. Yun, N.-N. Wang, A.-P. Lu, W.-B. Zeng and A. F. Chen, *J. Cheminf.*, 2015, **7**, 60.

27 M. Awale, D. Probst and J.-L. Reymond, *J. Chem. Inf. Model.*, 2017, **57**, 643–649.

28 M. González-Medina and J. L. Medina-Franco, *J. Chem. Inf. Model.*, 2017, **57**, 1735–1740.

29 M. González-Medina, F. D. Prieto-Martínez, J. R. Owen and J. L. Medina-Franco, *J. Cheminf.*, 2016, **8**, 63.

30 A. Daina, O. Michielin and V. Zoete, *Sci. Rep.*, 2017, **7**, 42717.

31 A. V. Rudik, V. M. Bezhentsev, A. V. Dmitriev, D. S. Druzhilovskiy, A. A. Lagunin, D. A. Filimonov and V. V. Poroikov, *J. Chem. Inf. Model.*, 2017, **57**, 638–642.

32 A. Rudik, A. Dmitriev, A. Lagunin, D. Filimonov and V. Poroikov, *Bioinformatics*, 2015, **31**, 2046–2048.

33 C. W. Yap, *J. Comput. Chem.*, 2011, **32**, 1466–1474.

34 L. Zhang, H. Ai, W. Chen, Z. Yin, H. Hu, J. Zhu, J. Zhao, Q. Zhao and H. Liu, *Sci. Rep.*, 2017, **7**, 2118.

35 R. C. Braga, V. M. Alves, E. N. Muratov, J. Strickland, N. Kleinstreuer, A. Trospsha and C. H. Andrade, *J. Chem. Inf. Model.*, 2017, **57**, 1013–1017.

36 M. González-Medina, O. Méndez-Lucio and J. L. Medina-Franco, *J. Chem. Inf. Model.*, 2017, **57**, 397–402.

37 J. Dong, Z.-J. Yao, M.-F. Zhu, N.-N. Wang, B. Lu, A. F. Chen, A.-P. Lu, H. Miao, W.-B. Zeng and D.-S. Cao, *J. Cheminf.*, 2017, **9**, 27.

38 S. J. Capuzzi, I. S.-J. Kim, W. I. Lam, T. E. Thornton, E. N. Muratov, D. Pozefsky and A. Tropsha, *J. Chem. Inf. Model.*, 2017, **57**, 105–108.

39 D. A. Olmedo, M. González-Medina, M. P. Gupta and J. L. Medina-Franco, *Mol. Diversity*, 2017, **21**, 779–789.

40 G. M. Maggiora, *J. Chem. Inf. Model.*, 2006, **46**, 1535.

41 J. L. Medina-Franco, *Chem. Biol. Drug Des.*, 2013, **81**, 553–556.

42 H. Liu, L. Wang, M. Lv, R. Pei, P. Li, Z. Pei, Y. Wang, W. Su and X.-Q. Xie, *J. Chem. Inf. Model.*, 2014, **54**, 1050–1060.

43 J. Fang, L. Wang, Y. Li, W. Lian, X. Pang, H. Wang, D. Yuan, Q. Wang, A.-L. Liu and G.-H. Du, *PLoS One*, 2017, **12**, e0178347.

44 G.-H. Li and J.-F. Huang, *Bioinformatics*, 2012, **28**, 3334–3335.

45 H. Singh, R. Kumar, S. Singh, K. Chaudhary, A. Gautam and G. P. S. Raghava, *BMC Cancer*, 2016, **16**, 77.

46 J. E. Tym, C. Mitsopoulos, E. A. Coker, P. Razaz, A. C. Schierz, A. A. Antolin and B. Al-Lazikani, *Nucleic Acids Res.*, 2016, **44**, D938–D943.

47 A. Gaulton, L. J. Bellis, A. P. Bento, J. Chambers, M. Davies, A. Hersey, Y. Light, S. McGlinchey, D. Michalovich, B. Al-Lazikani and J. P. Overington, *Nucleic Acids Res.*, 2012, **40**, D1100–D1107.

48 S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller and D. J. Lipman, *Nucleic Acids Res.*, 1997, **25**, 3389–3402.

49 Y.-Y. Chiu, C.-T. Lin, J.-W. Huang, K.-C. Hsu, J.-H. Tseng, S.-R. You and J.-M. Yang, *Nucleic Acids Res.*, 2013, **41**, D430–D440.

50 Y. Okuno, A. Tamon, H. Yabuuchi, S. Niijima, Y. Minowa, K. Tonomura, R. Kunimoto and C. Feng, *Nucleic Acids Res.*, 2007, **36**, D907–D912.

51 Z. Huang, H. Jiang, X. Liu, Y. Chen, J. Wong, Q. Wang, W. Huang, T. Shi and J. Zhang, *PLoS One*, 2012, **7**, e39917.

52 M. R. Berthold, N. Cebron, F. Dill, T. R. Gabriel, T. Kötter, T. Meinl, P. Ohl, K. Thiel and B. Wiswedel, *ACM SIGKDD Explor. Newsl.*, 2009, **11**, 26.

53 A. Tiwari and A. K. T. Sekhar, *Comput. Biol. Chem.*, 2007, **31**, 305–319.

54 D. Fourches, E. Muratov and A. Tropsha, *J. Chem. Inf. Model.*, 2010, **50**, 1189–1204.

55 D. Fourches, E. Muratov and A. Tropsha, *Nat. Chem. Biol.*, 2015, **11**, 535.

56 D. Fourches, E. Muratov and A. Tropsha, *J. Chem. Inf. Model.*, 2016, **56**, 1243–1252.

57 J.-M. Gally, S. Bourg, Q.-T. Do, S. Aci-Sèche and P. Bonnet, *Mol. Inf.*, 2017, **36**, 1700023.

58 G. Marcou and A. Varnek, in *Tutorials in Chemoinformatics*, ed. A. Varnek, John Wiley & Sons, Ltd, Chichester, UK, 1st edn, 2017, ch. 1, vol. 1, pp. 1–36.

59 M. P. Mazanetz, R. J. Marmon, C. B. T. Reisser and I. Morao, *Curr. Top. Med. Chem.*, 2012, **12**, 1965–1979.

60 R. McGuire, S. Verhoeven, M. Vass, G. Vriend, I. J. P. de Esch, S. J. Lusher, R. Leurs, L. Ridder, A. J. Kooistra, T. Ritschel and C. de Graaf, *J. Chem. Inf. Model.*, 2017, **57**, 115–121.

61 J. J. Naveja and J. L. Medina-Franco, *Drug Discovery Today*, 2017, DOI: 10.1016/j.drudis.2017.10.006.

62 A. Fillbrunn, C. Dietz, J. Pfeuffer, R. Rahn, G. A. Landrum and M. R. Berthold, *J. Biotechnol.*, 2017, **261**, 149–156.

63 E. Gilberg, D. Stumpfe and J. Bajorath, *RSC Adv.*, 2017, **7**, 35638–35647.

64 J. B. Baell, *J. Nat. Prod.*, 2016, **79**, 616–628.

65 S. Saubern, R. Guha and J. B. Baell, *Mol. Inf.*, 2011, **30**, 847–850.

66 J. B. Baell and G. A. Holloway, *J. Med. Chem.*, 2010, **53**, 2719–2740.

67 P. D. Dobson, Y. Patel and D. B. Kell, *Drug Discovery Today*, 2009, **14**, 31–40.

68 S. O'Hagan, N. Swainston, J. Handl and D. B. Kell, *Metabolomics*, 2015, **11**, 323–339.

ANEXO 2. Artículos complementarios

pharmaceuticals

MDPI

Article

Expanding the Structural Diversity of DNA Methyltransferase Inhibitors

K. Eurídice Juárez-Mercado [1], Fernando D. Prieto-Martínez [1], Norberto Sánchez-Cruz [1], Andrea Peña-Castillo [1], Diego Prada-Gracia [2] and José L. Medina-Franco [1,*]

[1] DIFACQUIM Research Group, Department of Pharmacy, School of Chemistry, National Autonomous University of Mexico, Avenida Universidad 3000, Mexico City 04510, Mexico; kaeuridice@gmail.com (K.E.J.-M.); ferdpm4@hotmail.com (F.D.P.-M.); norberto.sc90@gmail.com (N.S.-C.); andrea.pecas93@gmail.com (A.P.-C.)

[2] Research Unit on Computational Biology and Drug Design, Children's Hospital of Mexico Federico Gomez, Mexico City 06720, Mexico; prada.gracia@gmail.com

* Correspondence: medinajl@unam.mx

Abstract: Inhibitors of DNA methyltransferases (DNMTs) are attractive compounds for epigenetic drug discovery. They are also chemical tools to understand the biochemistry of epigenetic processes. Herein, we report five distinct inhibitors of DNMT1 characterized in enzymatic inhibition assays that did not show activity with DNMT3B. It was concluded that the dietary component theaflavin is an inhibitor of DNMT1. Two additional novel inhibitors of DNMT1 are the approved drugs glyburide and panobinostat. The DNMT1 enzymatic inhibitory activity of panobinostat, a known pan inhibitor of histone deacetylases, agrees with experimental reports of its ability to reduce DNMT1 activity in liver cancer cell lines. Molecular docking of the active compounds with DNMT1, and re-scoring with the recently developed extended connectivity interaction features approach, led to an excellent agreement between the experimental IC_{50} values and docking scores.

Keywords: dietary component; epigenetics; enzyme inhibition; focused library; epi-informatics; multitarget epigenetic agent; natural products; chemoinformatics

check for updates

Citation: Juárez-Mercado, K.E.; Prieto-Martínez, F.D.; Sánchez-Cruz, N.; Peña-Castillo, A.; Prada-Gracia, D.; Medina-Franco, J.L. Expanding the Structural Diversity of DNA Methyltransferase Inhibitors. *Pharmaceuticals* **2021**, *14*, 17. http://dx.doi.org/10.3390/ph14010017

Received: 20 November 2020
Accepted: 23 December 2020
Published: 27 December 2020

1. Introduction

Historically, the term "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 [1,2]. However, "epigenetics" is now used with multiple meanings, for instance, to describe the heritable phenotype (cellular memory) without modification of DNA sequences [3], or the mechanism in which the environment conveys its influence to the cell, tissue, or organism [4]. Regardless of the different definitions, the interest in epigenetic drug discovery has increased, as revealed by the multiple approved epigenetic drugs or compounds in clinical development for epigenetic targets [5,6].

DNA methyltransferases (DNMTs) are one of the primary 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) [7]. 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 [7,8].

DNMTs were the first epigenetic targets for which inhibitors received the approval of the Food and Drug Administration (FDA) of the USA: the nucleoside analogs 5-azacitidine (Vidaza) and decitabine or 5-aza-2′-deoxycytidine (Dacogen) (Figure 1), approved in 2004 and 2006, respectively, for the treatment of the myelodysplastic syndrome [9]. DNMTs

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

RSC Advances

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue

Check for updates

Cite this: *RSC Adv.*, 2020, **10**, 25089

Consensus virtual screening of dark chemical matter and food chemicals uncover potential inhibitors of SARS-CoV-2 main protease†

Marisa G. Santibáñez-Morán,[a] Edgar López-López,[b] Fernando D. Prieto-Martínez,[a] Norberto Sánchez-Cruz[a] and José L. Medina-Franco*[a]

Received 4th June 2020
Accepted 23rd June 2020

DOI: 10.1039/d0ra04922k

rsc.li/rsc-advances

The pandemic caused by SARS-CoV-2 (COVID-19 disease) has claimed more than 500 000 lives worldwide, and more than nine million people are infected. Unfortunately, an effective drug or vaccine for its treatment is yet to be found. The increasing information available on critical molecular targets of SARS-CoV-2 and active compounds against related coronaviruses facilitates the proposal (or repurposing) of drug candidates for the treatment of COVID-19, with the aid of *in silico* methods. As part of a global effort to fight the COVID-19 pandemic, herein we report a consensus virtual screening of extensive collections of food chemicals and compounds known as dark chemical matter. The rationale is to contribute to global efforts with a description of currently underexplored chemical space regions. The consensus approach included combining similarity searching with various queries and fingerprints, molecular docking with two docking protocols, and ADMETox profiling. We propose compounds commercially available for experimental testing. The full list of virtual screening hits is disclosed.

1. Introduction

Coronaviruses (COVs) *per se* can infect humans and other animal species. Some of them cause a variety of previously studied diseases such as Severe Acute Respiratory Syndrome (SARS) and Middle East Respiratory Syndrome (MERS). SARS-CoV-2 is an emergent virus that generates the COVID-19 disease[1] which is currently considered a "pandemic" according to the World Health Organization (WHO), with more than ten million confirmed cases and more than 500 000 deaths worldwide (as per June 30th, 2020).[2]

SARS-CoV-2 has a complex architecture, and as happens with different viruses, there are several proteins involved in viral internalization and replication. The life cycle of SARS-CoV-2 starts with the viral recognition of its spike protein by a cellular receptor (ACE receptor and TMPRSS2). After that, the internalization and uncoating process is mediated by membrane proteins. Once into the host cell, RNA replication, and biosynthesis of viral polypeptides are carried out (RdRp – ribosomes). Finally, the processing of precursors proteins by the main protease (3CLpro or M^{pro}) and the assembly of these, contributes to the generation of new viruses.[3–5] These main targets offer a venue for the development of new treatments *via* rational drug design. Examples include spike protein, RNA polymerase, and chymotrypsin-like cysteine protease (3CLpro or M^{pro}) which are presented in Fig. 1.[3–5] Of these, the main protease (M^{pro}) is a promising target for the design and proposal of new therapies due to the lack of homologous proteins in humans.[6] Also, its selective inhibition would take advantage of the natural life cycle of SARS-CoV-2, avoiding its replication and dissemination. Several research groups are actively pursuing M^{pro} as a molecular target to identify drug candidates for the treatment of COVID-19.

Computational methods represent an approach with the power of efficiently filter large and diverse compound libraries to select potential candidates for drug development.[7,8] Recently published works show a tendency towards drug repurposing and to search structurally different libraries (*e.g.*, with broad scaffold diversity), and natural products.[9–13] Moreover, the search for novel compounds commercially available or with the possibility of being synthesized has had a vital rebound (*e.g.*, screening part or the entire ZINC database).[9,14–16] Table 1 summarizes representative examples of virtual screening (VS) studies directed to different molecular targets, including SARS-CoV-2 M^{pro}. Most of these efforts relied on structure-based drug design (SBDD). Few others include similarity searching and quantitative structure–activity relationship (QSAR) modeling.[17] In this sense, there are many compounds suggested by

[a] *DIFACQUIM Research Group, Department of Pharmacy, School of Chemistry, Universidad Nacional Autónoma de México, Mexico City, Mexico. E-mail: medinajl@unam.mx; jose.medina.franco@gmail.com; Tel: +52 (55) 5622-3899, ext. 44458*

[b] *Department of Pharmacology, Center of Research and Advanced Studies of the National Polytechnic Institute (CINVESTAV), Mexico City, Mexico*

† Electronic supplementary information (ESI) available: Excel file with ten worksheets that report all similarity values, docking scores, and ADMETox profile of the hit compounds outlined in Fig. 2. Structure file of the 1052 queries used for the similarity searching. See DOI: 10.1039/d0ra04922k.

International Journal of
Molecular Sciences

MDPI

Review

Lessons from Exploring Chemical Space and Chemical Diversity of Propolis Components

Trong D. Tran [1,*], Steven M. Ogbourne [1], Peter R. Brooks [1], Norberto Sánchez-Cruz [2], José L. Medina-Franco [2] and Ronald J. Quinn [3]

[1] GeneCology Research Centre, School of Science and Engineering, University of the Sunshine Coast, Maroochydore DC, Queensland 4558, Australia; sogbourn@usc.edu.au (S.M.O.); PBrooks@usc.edu.au (P.R.B.)
[2] Department of Pharmacy, School of Chemistry, Universidad Nacional Autónoma de México, Mexico City 04510, Mexico; norberto.sc90@gmail.com (N.S.-C.); medinajl@unam.mx (J.L.M.-F.)
[3] Griffith Institute for Drug Discovery, Griffith University, Brisbane 4111, Australia; r.quinn@griffith.edu.au
* Correspondence: ttran1@usc.edu.au; Tel.: +61-7-5459-4579

Received: 18 June 2020; Accepted: 14 July 2020; Published: 15 July 2020

Abstract: Propolis is a natural resinous material produced by bees and has been used in folk medicines since ancient times. Due to it possessing a broad spectrum of biological activities, it has gained significant scientific and commercial interest over the last two decades. As a result of searching 122 publications reported up to the end of 2019, we assembled a unique compound database consisting of 578 components isolated from both honey bee propolis and stingless bee propolis, and analyzed the chemical space and chemical diversity of these compounds. The results demonstrated that both honey bee propolis and stingless bee propolis are valuable sources for pharmaceutical and nutraceutical development.

Keywords: honey bee propolis; stingless bee propolis; natural products; phenolics; terpenoids; chemoinformatics; chemical space; chemical diversity

1. Introduction

The emergence of new infectious and chronic diseases makes the need for new drugs paramount [1]. Although the search for new drugs can begin from different sources, natural products have proven to be one of the richest sources of bioactive ingredients and molecules with privileged scaffolds for the discovery and development of new and novel drugs [2–6]. They were historically the sources of all folk medicines [7]. Having evolved over millions of years, structures of natural products have been fine-tuned by nature for optimal bioactivity [5]. Modern studies revealed natural products possess an advantageous structural foundation and cover a wide range of biologically relevant chemical space that cannot be efficiently explored by synthetic compounds [8–10]. These features positively influence the probability of the clinical success of natural product-based drug candidates [11]. A detailed analysis of 1394 new small molecule drugs approved by the US Food and Drug Administration (FDA) between 1981 and 2019 [6] revealed that 32% of those drugs were natural products or direct derivatives of natural products.

Propolis, which is a product of bees, has been used in the folk medicine of many cultures to treat microbial infections since the year 300 B.C. [12]. The name “propolis” originally came from the Greek words meaning “defence of the city” (“pro” meaning “to defend” and “polis” meaning the city) [13]. Historically, the Greeks and the Romans used propolis for treating bruises and suppurating sores; the Egyptians applied propolis for embalming cadavers and preventing infections; the Arabians utilised propolis as an antiseptic, a wound healing agent, and a mouth disinfectant; the Incas described

frontiers
in Pharmacology

REVIEW
published: 10 October 2018
doi: 10.3389/fphar.2018.01144

Inhibitors of DNA Methyltransferases From Natural Sources: A Computational Perspective

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

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

OPEN ACCESS

Edited by:
Shibashish Giri,
University of Leipzig, Germany

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

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

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

Received: *08 August 2018*
Accepted: *21 September 2018*
Published: *10 October 2018*

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

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

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

SECTION 1: INTRODUCTION

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

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

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

Methods in Pharmacology and Toxicology (2018): 3–25
DOI 10.1007/7653_2018_6

Published online: 09 June 2018

Cheminformatics Approaches to Study Drug Polypharmacology

J. Jesús Naveja, Fernanda I. Saldívar-González, Norberto Sánchez-Cruz, and José L. Medina-Franco

This work is dedicated to the loving memory of Nicolás Medina Sandoval.

Abstract

Herein is presented a tutorial overview on selected chemoinformatics methods useful for assembling, curating/preparing a chemical database, and assessing its diversity and chemical space. Methods for evaluating the structure–activity relationships (SAR) and polypharmacology are also included. Usage of open source tools is emphasized. Step-by-step KNIME workflows are used for illustrating the methods. The methods described in this chapter are applied onto a chemical database especially relevant for epi-polypharmacology that is an emerging area in drug discovery. However, the methods described herein could be extended to other therapeutic areas and potentially to other areas of chemistry.

Keywords Chemoinformatics, ChemMaps, Chemical space, Data mining, Epigenetics, Epi-informatics, KNIME, Molecular diversity, Open-access, Polypharmacology, Structure–activity relationships, SmARt

1 Introduction

The rapid growth of chemical information demands efficient and reliable computational algorithms to analyze the accumulated data. Similarly, current trends in drug discovery such as polypharmacology [1, 2] demand the organization and efficient mining of multiple drug–target interactions and study of structure–multiple activity relationships (SMARt) efficiently [3]. Indeed, a plethora of methods and resources for exploiting SMARt and other data relevant to polypharmacology have been published, and many of them are open access [4]. This review includes methodological details for implementing scalable KNIME cheminformatics workflows for:

a. Curating a chemical database;

b. Computing chemical descriptors;

Electronic supplementary material: The online version of this article (https://doi.org/10.1007/7653_2018_6) contains supplementary material, which is available to authorized users.