



**UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO**

**PLAN DE ESTUDIOS COMBINADOS EN MEDICINA  
FACULTAD DE MEDICINA**

**Minería de datos biológicos de dianas epigenéticas para el  
desarrollo de compuestos polifarmacológicos y  
combinaciones de fármacos**

**TESIS**

que para optar por el grado de:

**DOCTOR EN MEDICINA**

PRESENTA:

**JOSÉ DE JESÚS NAVEJA ROMERO**

TUTOR PRINCIPAL:

**Dr. José Luis Medina Franco**  
Facultad de Química, UNAM

MIEMBROS DEL COMITÉ TUTOR:

**Dr. José Correa Basurto**  
Escuela Superior de Medicina, IPN

**Dr. José de la Luz Díaz Chávez**  
Instituto Nacional de Cancerología

**Dr. Maximino Aldana González**  
Instituto de Ciencias Físicas, UNAM

**Dr. Iwein Roger Maria Leenen**  
Facultad de Psicología, UNAM



Universidad Nacional  
Autónoma de México

Dirección General de Bibliotecas de la UNAM

**Biblioteca Central**



**UNAM – Dirección General de Bibliotecas**  
**Tesis Digitales**  
**Restricciones de uso**

**DERECHOS RESERVADOS ©**  
**PROHIBIDA SU REPRODUCCIÓN TOTAL O PARCIAL**

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

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

*A mis padres, de cuyo esfuerzo soy resultado*

*A mis hermanos, con todo mi cariño y admiración*

*A José Luis, mi guía académico y amigo, por llevarme paso a paso por el camino  
de la ciencia*

*A Carlos Cirlos e Iwin Leenen, los profesores de los que más pude aprender;  
con suerte, algún día seré como ellos*

*Al PECEM, experimento del que fui sujeto, y a la Dra. Flisser, su incansable ejecutora;  
ella es un ejemplo de perseverancia y compromiso con los estudiantes*

*A Susana García; el amor y apoyo mutuo me guiaron como un faro en las dificultades*

*A la UNAM, mi hogar*

## Resumen

En esta tesis se presentan los resultados de diferentes investigaciones dirigidas a un objetivo común: desarrollar metodologías para extraer información de ensayos fenotípicos (en líneas celulares de cáncer) y ensayos contra dianas moleculares, así como combinar la información de ambas fuentes. (Entiéndase “ensayo” como cada una de las repeticiones de un experimento.) Durante el desarrollo del proyecto, se alcanzaron resultados a distintos niveles. A nivel celular, se desarrolló una estrategia *in silico* de propagación de redes para identificar combinaciones de dianas moleculares cuya inhibición simultánea puede resultar en sinergia farmacológica; también se propusieron mecanismos a través de los cuales esta sinergia podría ocurrir. A nivel de dianas moleculares, el análisis se centró en dianas relacionadas con procesos epigenéticos; propusimos comparar en el espacio químico las bibliotecas de inhibidores farmacológicos contra las diferentes dianas. Así, es posible buscar oportunidades de desarrollar moléculas con actividad en múltiples dianas epigenéticas. Por último, a nivel farmacológico desarrollamos varios métodos de análisis nuevos. Algunos de ellos se enfocan en obtener información consistente y químicamente relevante a partir de los ensayos biológicos de alto rendimiento: es de mayor prioridad encontrar una serie de análogos químicos con cierta actividad biológica de interés que moléculas aisladas, las cuales podrían tratarse de un falso positivo en la prueba. Otros métodos de análisis a nivel farmacológico se enfocaron en la visualización y estudio del espacio químico y las relaciones estructura-actividad. Los datos utilizados incluyeron ensayos farmacológicos de inhibición de crecimiento en líneas celulares de cáncer, así como ensayos *in vitro* que muestran los perfiles de actividad de compuestos contra dianas moleculares.

Los **objetivos específicos** fueron:

1. Explorar el efecto del tratamiento con moléculas polifarmacológicas y combinaciones de moléculas en el crecimiento de líneas celulares de cáncer.
2. Identificar y comparar los patrones químicos de los inhibidores de dianas epigenéticas.
3. Desarrollar métodos quimiinformáticos novedosos para el análisis de bibliotecas de moléculas pequeñas.

Para alcanzar los objetivos específicos se siguió la siguiente **estrategia**:

1. Integración de bases de datos públicas con información acerca de inhibidores farmacológicos de dianas moleculares y de líneas celulares de cáncer.
2. Rastreo de combinaciones de dianas moleculares con efecto sinérgico sobre distintas líneas celulares de cáncer.
3. Exploración quimiinformática de las bibliotecas de moléculas pequeñas con actividad contra distintas dianas epigenéticas.
4. Desarrollo de métodos para la visualización, análisis cuantitativos de diversidad y de relaciones estructura-actividad de bases de datos de compuestos de interés biológico. Se enfatizó en bases de datos de compuestos con actividad epigenética y en métodos relacionados con polifarmacología.

Se concluyó que los datos analizados son informativos acerca de combinaciones de dianas potencialmente relevantes en el desarrollo de moléculas polifarmacológicas o combinaciones de moléculas con actividad anticancerígena.



## Abstract

Herein we present the results of research directed towards the common goal of developing computation methods to extract information from phenotypic assays and target-based assays, as well as combining both sources' information. Along the development course of the project, results were produced at different levels. At a cellular level, we developed a network propagation *in silico* strategy to identify target combinations whose simultaneous inhibitions may elicit pharmacological synergism; we also proposed biological mechanisms to explain this synergism. At the level of molecular targets, the analysis was centered on epigenetic targets; we proposed comparing inhibitors libraries of the different targets in the chemical space. Thereby it is possible to search for opportunities to develop multi-target molecules. Last, at a pharmacological level, we developed a variety of novel methods for cheminformatic analysis. Some of them are aimed towards the obtention of consistent and chemically-relevant information from high-throughput screening data. In other words, in the context of a particular biological assay where a large library of chemical compounds is experimentally tested, it would be more interesting to identify analog series that share a biological effect of interest, rather than identifying isolated molecules, which are more likely to be false positives. Other cheminformatic methods focus on the visualization and study of the chemical space and structure-activity relationships. The data included came from cell-inhibition assays as well as target-inhibition assays.

The specific goals were:

1. To explore the effect of the treatment with polypharmacological molecules or combinations of molecules in the growth of cancer cell lines.
2. To identify and compare the chemical patterns of epigenetic inhibitors.
3. To develop novel cheminformatic methods for the analysis of small-molecules libraries.

To met these goals, the following strategy was followed:

1. Integration of public databases regarding pharmacological inhibitors of molecular targets and cancer cell lines.
2. Deconvolution of target combinations with synergistic effects on different cancer cell lines.
3. Chemoinformatic exploration of small-molecule libraries with activity against different epigenetic targets.
4. Development of methods for visualization, quantitative analysis of diversity and structure-activity relationships of chemical databases of biological interest. We centered on chemical databases of epigenetic inhibitors and methods related to polypharmacology in general.

We concluded that the analyzed data are informative about target combinations with potential relevance in the development of polypharmacological molecules or compound combinations as anticancer therapies.

# Índice

<b>1. Introducción</b>	<b>6</b>
1.1. Descubrimiento de fármacos basado en dianas vs. fenotipos . . . . .	6
1.2. Rastreo de dianas y aplicación en cáncer . . . . .	7
1.3. Propagación de redes para el rastreo de combinaciones de dianas . . . . .	7
1.4. Oportunidades de polifarmacología en la epigenética del cáncer . . . . .	8
1.5. Cribado virtual simultáneo contra múltiples dianas . . . . .	8
1.6. Definiciones operativas . . . . .	8
1.6.1. Espacio químico . . . . .	9
1.6.2. Polifarmacología . . . . .	10
<b>2. Objetivos del proyecto</b>	<b>10</b>
2.1. Objetivo general . . . . .	10
2.2. Objetivo secundario . . . . .	10
2.3. Objetivos específicos . . . . .	10
2.4. Hipótesis . . . . .	10
<b>3. Resultados obtenidos y significancia</b>	<b>11</b>
<b>4. Rastreo de combinaciones de dianas en líneas celulares de cáncer</b>	<b>12</b>
<b>5. Análisis de compuestos con actividad epigenética</b>	<b>21</b>
<b>6. Desarrollo y aplicación de métodos nuevos</b>	<b>51</b>
<b>7. Discusión, conclusiones y perspectivas</b>	<b>140</b>

# 1. Introducción

## 1.1. Descubrimiento de fármacos basado en dianas vs. fenotipos

Los métodos clásicos de descubrimiento de fármacos basados en dianas biológicas se enfocan en dianas únicas, contra las cuales se diseñan y prueban diferentes moléculas pequeñas. La simplicidad de estas estrategias ha sido fundamental en el desarrollo y descubrimiento de fármacos. Sin embargo, tales métodos han demostrado ser menos efectivos cuando se utilizan en el contexto de enfermedades complejas, por ejemplo, cáncer [1]. Es posible que el efecto fenotípico de la inhibición farmacológica selectiva de una sola diana sea contrarrestado por los múltiples procesos adaptativos que coexisten en los sistemas biológicos [2].

En este escenario, la inhibición conjunta de múltiples dianas es particularmente útil; se puede producir por la combinación de dos moléculas, en cuyo caso buscaríamos combinaciones con efectos sinérgicos, es decir, la combinación tiene un efecto mayor que la suma de los efectos individuales [3, 4]. También, se ha identificado que algunas moléculas presentan “polifarmacología”, la propiedad de actuar contra múltiples dianas biológicas; es una alternativa prometedora a las combinaciones de fármacos, porque se hipotetiza que se producirían menos efectos adversos utilizando menos fármacos [5]. Se ha propuesto que un diseño racional de compuestos polifarmacológicos podría producir moléculas que actúen estratégicamente en múltiples dianas biológicas de una cascada de regulación celular, de manera que tengan un efecto más importante en el fenotipo de enfermedades complejas [2].

Si bien la polifarmacología resulta un concepto interesante, desarrollar compuestos con esta propiedad desde el enfoque basado en dianas conlleva múltiples retos metodológicos [5]. El principal de estos es identificar las combinaciones de dianas moleculares contra las cuales se debería dirigir el tratamiento; aunque se han hecho protocolos para inhibir dos dianas simultáneamente en células por medio de ARN de interferencia [6], seguramente la cantidad de posibles combinaciones de dianas es mayor de las que es posible probar en experimentos. Otras alternativas se basan en la información del interactoma y la expresión génica de un tipo celular particular para generar modelos *in silico* que puedan identificar susceptibilidades de un tumor particular de un paciente [7]. Una vez identificadas las dianas, surge otro reto: diseñar una molécula o una combinación de moléculas para inhibirlas. Es evidente que el diseño racional de inhibidores polifarmacológicos puede fallar en cualquiera de estos dos puntos.

Ante las múltiples dificultades que demostró el diseño basado en dianas, comenzaron a recuperarse las ideas del descubrimiento de fármacos basado en fenotipos: el método de prueba y error aplicado directamente en modelos biológicos de mayor complejidad [8]. Se han estandarizado ensayos biológicos de alto rendimiento que permiten probar miles de moléculas pequeñas directamente contra líneas celulares de cáncer [9, 10]. Recientemente se utilizó este modelo para probar combinaciones de moléculas; se identificaron dos combinaciones de fármacos que ahora están en ensayos clínicos de fase I contra distintos tumores sólidos resistentes a tratamiento: bortezumib y clofarabina (ensayo clínico NCT02211755) y paclitaxil y nilotinib (ensayo clínico NCT02379416) [11]. Está previsto que ambos estudios concluyan en 2020. También se han generado modelos en *Drosophila melanogaster* [12, 13] e incluso modelos murinos con injertos de tumores de pacientes, que permitirían incluso encontrar tratamientos individualizados por paciente [14].

## 1.2. Rastreo de dianas y aplicación en cáncer

El pilar central del diseño de fármacos basado en la diana es la identificación de compuestos activos y selectivos contra dianas moleculares específicas que se consideran relevantes en el contexto de una enfermedad. Una vez identificados estos inhibidores, típicamente a través de experimentos *in vitro*, se procede a probar si producen cierto efecto fenotípico deseado en un modelo biológico más complejo que una sola diana molecular. Esta estrategia es reduccionista y se aplica con poca efectividad en enfermedades complejas [15]. En consecuencia, las pruebas fenotípicas están resurgiendo como primer cribado para las moléculas pequeñas; estas estrategias se basan en encontrar compuestos activos en sistemas biológicos complejos [8].

El rastreo de dianas hace un vínculo entre los datos obtenidos del cribado basado en la diana y los obtenidos del cribado basado en el fenotipo. Se centra en identificar dianas que podrían ser responsables de efectos fenotípicos deseables que se observan experimentalmente [15, 16, 17]. Se han desarrollado estrategias computacionales para hacer rastreo de dianas o explorar los efectos de inhibir múltiples dianas. Por ejemplo, Gayvert *et al.* usaron *random forests*, un método de inteligencia artificial, para predecir la terminación de ensayos clínicos debido a toxicidad. La mayor parte del poder predictivo del modelo que desarrollaron provenía de información de las dianas, como su distribución en los tejidos [18]. También se publicó un estudio pionero para el rastreo de dianas de efectos adversos de fármacos [19]. Estas estrategias se basan en la premisa de que los fenotipos humanos pueden ser causados por el perfil de dianas que tienen los fármacos.

Recientemente, se han desarrollado estrategias para rastreo de dianas que consideran a la polifarmacología. Al-Ali *et al.* construyeron un método que se basa en máquinas de soporte vectorial (otro método de inteligencia artificial) para analizar datos de inhibidores de cinasas en el crecimiento de axones, y finalmente lograron señalar combinaciones de cinasas que es importante inhibir para lograr este efecto [20]. Cabe resaltar que, en general, el cribado fenotípico tiene cada vez mayor importancia en la comunidad científica, así como las estrategias computacionales para el rastreo de dianas [15, 8, 16, 17, 18, 20, 21].

## 1.3. Propagación de redes para el rastreo de combinaciones de dianas

Se han desarrollado metodologías enfocadas en la identificación de dianas únicas relevantes, a partir de datos fenotípicos y de anotaciones de dianas moleculares [17, 19]. Por otra parte, el desarrollo de métodos para buscar combinaciones de dianas es todavía incipiente, aunque con resultados prometedores. Helal *et al.* crearon los *high-throughput screening fingerprints*, basados en la información de perfiles experimentales contra múltiples dianas biológicas que había de estudios de cribado experimental masivo. Se validó la utilidad de estos descriptores en cribado virtual [22].

En la Sección 4 de este trabajo proponemos una nueva metodología basada en la propagación de redes aplicada en el interactoma humano [23] para estudiar datos experimentales de moléculas polifarmacológicas y combinaciones sinérgicas (tratamientos combinados de moléculas que producen un efecto farmacológico mayor que el esperado por la adición de los efectos individuales); esto permitió identificar posibles mecanismos de sinergia en líneas celulares de cáncer y rastrear combinaciones de dianas de potencial interés.

## 1.4. Oportunidades de polifarmacología en la epigenética del cáncer

Los procesos epigenéticos comparten dos propiedades interesantes: son reversibles y ejercen efectos moduladores sobre el genoma [24]. Además, se ha encontrado que muchos de los mecanismos típicos del cáncer se relacionan directamente con procesos epigenéticos [25]. Por lo anterior, las dianas epigenéticas son blancos interesantes en cáncer.

Se han propuesto aplicaciones terapéuticas de inhibidores epigenéticos en enfermedades cardiovasculares, neurológicas y metabólicas; estas enfermedades se suelen asociar tanto con fenotipos complejos como con desregulación epigenética [26, 27, 28, 29, 30, 31, 32]. Pocos fármacos con actividad sobre dianas epigenéticas (o epidianas) han sido aprobados para su uso clínico [33, 34]. Sin embargo, ahora se sabe que múltiples fármacos ejercen efectos epigenéticos [35, 36, 37, 38].

Los epifármacos (fármacos con actividad epigenética) se pueden clasificar como reprogramadores amplios o como terapias dirigidas [34]. Los primeros tienen numerosos efectos en la expresión génica y modifican el perfil epigenético general de la célula (p.ej., metilación del ADN, marcas de histonas). Las terapias epigenéticas dirigidas toman ventaja de la fisiología aberrante de las células enfermas para diseñar tratamientos selectivos contra procesos epigenéticos puntuales [34].

Notablemente, la mayoría de los procesos epigenéticos pueden ser modificados farmacológicamente a diferentes niveles de su regulación, los cuales son: la colocación de la marca epigenética, su transducción y eliminación. Además, la perturbación simultánea de más de un proceso epigenético puede producir resultados no aditivos y, hasta cierto punto, inesperados [39, 40]. Si se considera además que las moléculas pequeñas usualmente muestran polifarmacología, es decir, actúan en más de una diana [41], sería relevante identificar de manera sistemática a: 1) las moléculas con múltiples dianas epigenéticas; y 2) las dianas epigenéticas cuya inhibición combinada es más relevante para lograr un fenotipo: por ejemplo, la inhibición selectiva de una línea celular de cáncer.

## 1.5. Cribado virtual simultáneo contra múltiples dianas

La Figura 1 ilustra el flujo general que se propone para un proyecto de cribado virtual contra múltiples dianas (*virtual screening* en la bibliografía especializada). En primer lugar, se deben reunir datos que informen acerca de las moléculas que inhiben a ciertas dianas. Si además se conoce el efecto fenotípico de los compuestos (por ejemplo, si inhiben o no el crecimiento en una línea celular), entonces se puede rastrear a las dianas (y combinaciones de ellas) que es más importante inhibir para lograr el fenotipo. Finalmente, una vez que se ha seleccionado una combinación de dianas, se pueden utilizar distintas herramientas para predecir compuestos activos contra las dianas de interés, por ejemplo, a través de búsquedas por similitud estructural [42].

## 1.6. Definiciones operativas

En este trabajo se utilizan varios conceptos técnicos; para algunos de ellos no existe un consenso en la literatura especializada acerca de cómo se deberían definir. Por esta razón, se

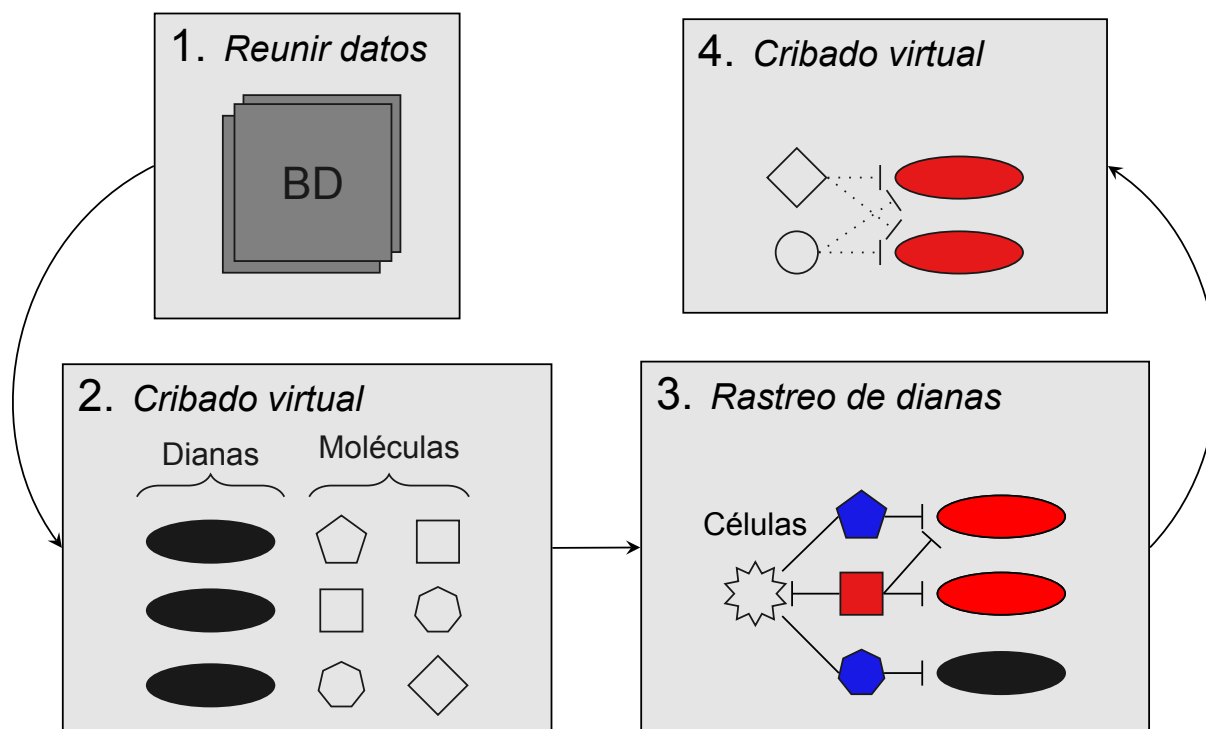


Figura 1: Esquema propuesto para un proyecto de cribado virtual (*virtual screening*) contra múltiples dianas. 1) Reunir datos quimioepigenómicos y de fenotipos; 2) Generar bibliotecas de compuestos enfocadas a distintas dianas; 3) Combinar la información de compuestos-dianas y compuestos-células para rastrear de dianas relevantes en líneas celulares de cáncer; 4) Aplicar la información de las bibliotecas compuesto-dianas para predecir compuestos (o combinaciones de compuestos) que actúen contra las dianas relevantes identificadas en el rastreo de dianas.

presentan las definiciones de dos conceptos fundamentales en el planteamiento de esta tesis: “espacio químico” y “polifarmacología”.

### 1.6.1. Espacio químico

El concepto de “espacio químico” se refiere a todas las moléculas que pueden existir [43]. Se han desarrollado diferentes métodos para explorar el espacio químico, y dependen principalmente del tipo de información que se utiliza para comparar a los compuestos, por ejemplo, propiedades físico-químicas o similitud estructural [44]. Incluso, se han desarrollado herramientas para comparar la información de bibliotecas moleculares completas, considerando la información de las estructuras presentes en cada biblioteca [42]. De la misma manera, hay diferentes métodos para generar representaciones visuales (generalmente aproximadas) del espacio químico en dos o tres dimensiones [45].

### **1.6.2. Polifarmacología**

Existen en la literatura diferentes definiciones para el término “polifarmacología”. En este trabajo, utilizamos la definición de Anighoro *et al.*, donde este término se utiliza para “designar a una sola molécula capaz de interactuar con múltiples dianas de manera específica” [5]. De esta forma, se puede distinguir a la polifarmacología de la “promiscuidad molecular”; esta última se refiere a la propiedad de ciertas moléculas de interactuar con múltiples dianas biológicas por medio de mecanismos inespecíficos [5]. Por metonimia, nos referimos en este trabajo a los compuestos con propiedades polifarmacológicas como “compuestos polifarmacológicos”, como hace Peters [46].

## **2. Objetivos del proyecto**

### **2.1. Objetivo general**

Identificar combinaciones de dianas moleculares cuya inhibición combinada resulte en un efecto sinérgico de inhibición de líneas celulares de cáncer.

### **2.2. Objetivo secundario**

Explorar oportunidades de polifarmacología en dianas epigenéticas.

### **2.3. Objetivos específicos**

1. Explorar el efecto del tratamiento con moléculas polifarmacológicas en el crecimiento de líneas celulares de cáncer.
2. Identificar y comparar los patrones moleculares de los inhibidores de dianas epigenéticas.
3. Desarrollar métodos quimioinformáticos novedosos para el análisis de bibliotecas de moléculas pequeñas.

### **2.4. Hipótesis**

- a. La inhibición conjunta de dos dianas moleculares que produce sinergia contra líneas celulares de cáncer por parte de una combinación de moléculas también se asocia con moléculas polifarmacológicas más potentes contra estas células.
- b. En general, si las bibliotecas de inhibidores de dos dianas epigenéticas son similares, también su función biológica será similar.

### 3. Resultados obtenidos y significancia

En este trabajo se exploraron los efectos de moléculas polifarmacológicas y combinaciones sinérgicas de moléculas en cáncer. De este modo se logró identificar módulos de proteínas en los que actúan los compuestos polifarmacológicos más activos, así como las combinaciones de compuestos probadas experimentalmente. A estos módulos los llamamos “*pathways de sinergia*”. Proponemos que las moléculas polifarmacológicas o combinaciones de moléculas que inhiban a más de una molécula relacionada con estos *pathways* tendrán un mayor efecto en líneas celulares de cáncer.

El conocimiento que se obtuvo indica que es factible encontrar combinaciones de dianas relevantes para inhibir líneas celulares de cáncer, a partir de datos de cribado masivo. Además, los resultados del proyecto se publicaron en 10 artículos en revistas indizadas y un capítulo de libro.

Por otra parte, se utilizaron las bibliotecas de compuestos activos contra las distintas dianas epigenéticas para medir la similitud farmacológica entre ellas. A través de este procedimiento, se identificaron dianas cuyas bibliotecas de inhibidores están más cercanas en el espacio químico; puede ser más fácil desarrollar inhibidores duales contra combinaciones de estas dianas. Las conclusiones a las que apuntan otros estudios de dinámica de redes sugieren que inhibir combinaciones de dianas podría ser más efectivo que apegarse al paradigma de una única diana [7, 47]. Por lo tanto, se puede hipotetizar que las moléculas con actividad en múltiples dianas pueden ser agentes terapéuticos atractivos y sujetas a análisis de rastreo de combinaciones de dianas.

A continuación se adjuntan los trabajos publicados en relación con esta tesis. En la Sección 4 se presenta el resultado principal de la tesis, que tiene como objetivo rastrear múltiples dianas relevantes en líneas celulares de cáncer, así como la propuesta de mecanismos por los que actúan. La Sección 5 incluye dos escritos con respecto al análisis de las bibliotecas de compuestos con actividad epigenética, comparando la información de las distintas dianas. La Sección 6 presenta los resultados del desarrollo de nuevas técnicas de análisis quimioinformáticos para el análisis de datos relacionados con el proyecto. Por último, la sección 7 presenta una discusión general y las conclusiones del proyecto. Al principio de cada sección presentamos un breve apartado con las ideas clave de los artículos incluidos.



## **4. Rastreo de combinaciones de dianas en líneas celulares de cáncer**

### **Ideas clave**

#### **Marco conceptual**

En este artículo investigamos la asociación entre compuestos polifarmacológicos que inhiben a líneas celulares de cáncer y combinaciones de fármacos, considerando como variables latentes los pares de dianas que inhiben. Suponemos que, si dos dianas son responsables de la sinergia observada en una combinación de moléculas, entonces una molécula que inhiba a ambas dianas también será más potente (Fig 1.). Una vez que se identificaron combinaciones de dianas asociadas con sinergia, se utilizó un modelo de propagación de redes para encontrar puntos en común en el interactoma que pudieran explicar la aparición de sinergia ante la inhibición dual.


#### **Datos utilizados**

Para este estudio incluimos bases de datos experimentales de ensayos de alto rendimiento (Fig. 2). Una base de datos es de ensayos con moléculas únicas, mientras que la otra es de ensayos con combinaciones de dos moléculas. Además, como fuente de información acerca de las dianas que inhibidas por las moléculas, utilizamos una de las bibliotecas públicas más grandes y confiables disponibles.

#### **Metodología y resultados**

Para las moléculas polifarmacológicas, primero cada molécula fue anotada con las dianas y células que inhibe. Las moléculas que fueron activas contra todas las células fueron descartadas del análisis, así como las que fueron inactivas contra todas las células. Posteriormente, buscamos combinaciones de dianas donde para cada diana del par existe al menos una molécula inactiva contra la línea celular, para evitar que una sola diana explicara la inhibición del par. Para las combinaciones de moléculas, se consideró que la sinergia debería ser consecuencia de combinaciones de dianas emergentes, es decir, que ocurren por efecto de la combinación. Todos los análisis se hicieron considerando cada línea celular por separado, con la finalidad de poder identificar combinaciones de dianas selectivas contra algún tipo celular. Después se compararon ambos resultados; encontramos que los pares de dianas identificados de moléculas polifarmacológicas son más propensos a producir sinergia cuando se inhiben por una combinación de moléculas (31 % vs. 41 %; Fig. 3). Las combinaciones de dianas consenso fueron estudiadas en el contexto del interactoma por medio de un análisis de propagación de redes (Figs. 4-6), que permitió identificar módulos de proteínas asociadas con cáncer, metabolismo y transporte celular (Fig. 7).

# Exploration of Target Synergy in Cancer Treatment by Cell-Based Screening Assay and Network Propagation Analysis

J. Jesús Naveja,<sup>†,‡,§</sup> Dagmar Stumpf,<sup>†</sup> José L. Medina-Franco,<sup>\*,§</sup> and Jürgen Bajorath<sup>\*,†</sup> 

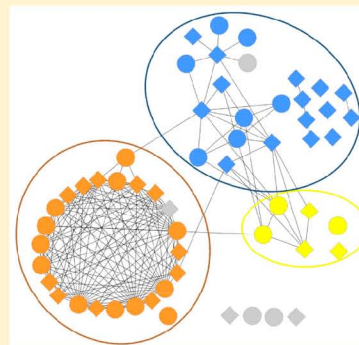
<sup>†</sup>Department of Life Science Informatics, B-IT, LIMES Program Unit Chemical Biology and Medicinal Chemistry, Rheinische Friedrich-Wilhelms-Universität, Endenicher Allee 19c, D-53115 Bonn, Germany

<sup>‡</sup>PECEM, Faculty of Medicine, Universidad Nacional Autónoma de México, Mexico City, 04510, Mexico

<sup>§</sup>Department of Pharmacy, School of Chemistry, Universidad Nacional Autónoma de México, Mexico City, 04510, Mexico

## Supporting Information

**ABSTRACT:** Computational approaches have previously been introduced to predict compounds with activity against multiple targets or compound combinations with synergistic functional effects. By contrast, there are no computational studies available that explore combinations of targets that might act synergistically upon small molecule treatment. Herein, we introduce an approach designed to identify synergistic target pairs on the basis of cell-based screening data and compounds with known target annotations. The targets involved in forming synergistic pairs were analyzed through a novel network propagation algorithm for rationalizing possible common synergy mechanisms. This algorithm enabled further analysis of each synergistic target pair and the identification of “interactors”, i.e., proteins with higher propagation scores than would be expected by adding the individual contributions of each target in the synergistic pair. We detected 137 synergistic target pairs including 51 unique targets. A global network analysis of these 51 targets made it possible to derive a subnetwork of proteins with significant synergy. Furthermore, interactors were identified for 87 synergistic target pairs upon individual analysis of the network propagation of each pair. These interactors were associated with pathways related to cancer and apoptosis, membrane transport, and steroid metabolism and provided possible explanations of synergistic effects.



## ■ INTRODUCTION

Synergy between bioactive compounds refers to the observation that biological effect(s) resulting from combined administration of compounds may exceed the sum of individual compound contributions.<sup>1</sup> There are several pharmacological mechanisms giving rise to synergistic effects.<sup>2</sup> Compound synergy is thought to be relevant for a variety of therapeutic applications.<sup>3–9</sup> Cell-based screening is a major source for studying compound synergy. For example, a screening project termed NCI-ALMANAC (National Cancer Institute—A Large Matrix of Anti-Neoplastic Agent Combinations) was recently carried out by the US National Cancer Institute.<sup>10</sup> Combinations of more than 100 anticancer compounds approved by the FDA (US Federal Drug Administration) were systematically tested on 60 cancer cell lines and synergistic compound pairs were identified.<sup>10</sup> Phenotypic screening has been complemented by computational prediction of synergistic compounds. For example, machine learning methods were applied to predict synergistic compound combinations on the basis of gene expression profiles<sup>11,12</sup> and biological networks were analyzed taking principles of synthetic lethality into consideration.<sup>13–17</sup>

Although several investigations have addressed compound synergy,<sup>4–12</sup> targets that act synergistically upon small molecule treatment have thus far not been identified. Knowledge of compound-dependent target synergy in a

cellular context would substantially aid in designing multitarget therapies and complement results of gene knockout experiments to identify synthetic lethality.

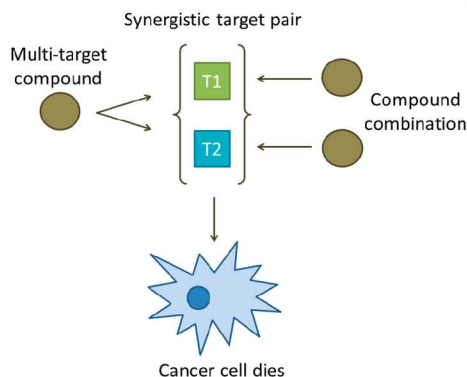
Herein, we introduce a general computational approach for identifying and rationalizing synergistic target pairs on the basis of phenotypic screening and compound activity data deposited in NCI60<sup>18</sup> and ChEMBL,<sup>19</sup> respectively. Our analysis identified pairs of small molecule targets that elicited cellular effects when inhibited in combination but not by individual engagement. The targets involved in forming synergistic target pairs were further analyzed using a novel network propagation algorithm introduced herein that identifies converging nodes in a human protein–protein interactions (PPI) network. Network propagation analysis yielded a subnetwork of proteins linked to targets in synergistic pairs, i.e., proteins potentially involved in mechanisms of synergy. Moreover, the network propagation approach allowed studying each synergistic target pair in greater detail by identifying other proteins termed “interactors” for which the propagating signal was amplified upon simultaneous inhibition of both targets. These interactors were found to be associated with pathways related to cancer and apoptosis, membrane transport, and steroid metabolism and hence provided possible explanations for synergy.

**Received:** January 8, 2019

**Published:** April 23, 2019

## MATERIALS AND METHODS

**Analysis Overview.** In our analysis, target synergy as a consequence of small molecule engagement was rationalized as follows: if coinhibition of two targets caused cell death that was not observed by inhibiting the targets individually, the target pair was classified as synergistic (Figure 1). Although



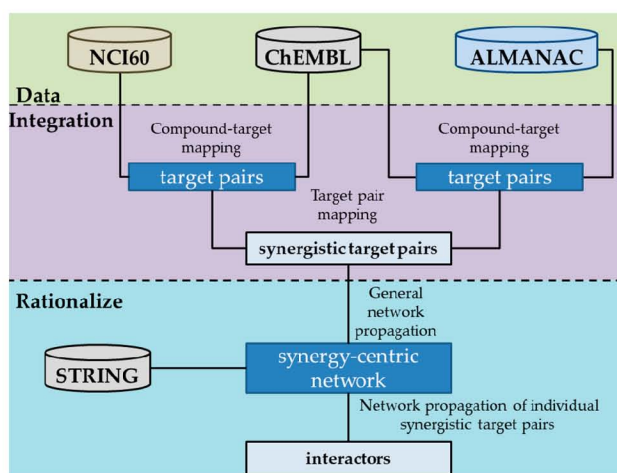
**Figure 1.** Compound-based target synergy. A synergistic target pair can, in principle, be inhibited by a single compound with multitarget activity as well as by a combination of compounds eliciting the same phenotypic effect.

compounds might affect cells in a variety of ways, compound-based target synergy implies that small molecules with activity in cell-based assays specifically interact with targets that are responsible for phenotypic effects. However, if a compound would be consistently active across cell lines, target-independent causes of apparent activity would be likely, such as nonspecific cytotoxic effects, and the compound would not qualify as a probe for target synergy.

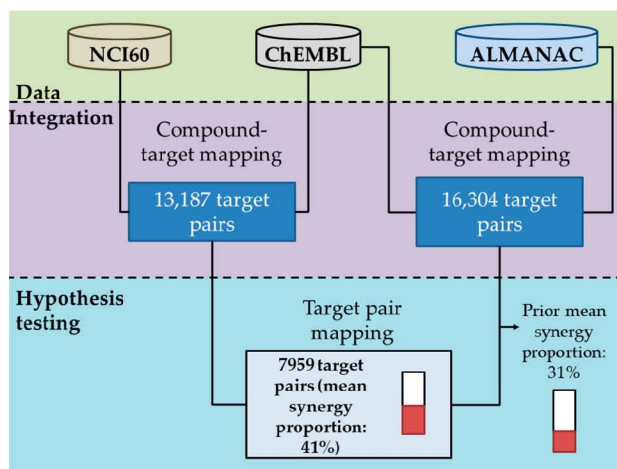
The goal of this study was to identify synergistic target pairs from cancer cell line screening data and derive mechanistic hypotheses on the basis of a human protein–protein interactions (PPI) network. An outline of the methodology is presented in Figure 2. NCI60, ALMANAC, and ChEMBL are the data sources. Consensus synergistic target pairs were inferred by analyzing NCI60/ChEMBL and ALMANAC/ChEMBL. These synergistic target pairs were then further rationalized using network propagation analysis on the basis of the human PPI network derived from interactions collected in STRING.<sup>20</sup> Initial propagation analysis of all targets in synergistic pairs provided a synergy-centric network. This was followed by a network propagation of individual synergistic target pairs to identify interactors, i.e., other proteins for which simultaneous propagation of both targets in a synergistic pair yielded higher signal accumulation than for each target in the pair and any other protein.

Initial evidence for synergistic target pairs was provided by comparing multitarget compounds and compound combinations using screening data (Figure 3). To provide mechanistic hypotheses for synergistic target pairs, we developed a flexible and efficient network propagation algorithm that was applied to protein–protein interactions data (Figure 4). On the basis of propagation analysis, interactions were identified and further analyzed (Figure 5). In the following, details are provided for each analysis step.

**Compound Data and Targets.** The NCI60 data set contained screening data for 40 998 compounds on 73 cancer cell lines. On average, a compound was tested on 56 cell lines



**Figure 2.** Methodological outline. Data integration led to the identification of synergistic target pairs. These synergistic pairs were then rationalized on the basis of a human protein–protein interaction network to identify interactors. NCI60 contains cell screening data for individual compounds, ChEMBL compounds and biological activity data from medicinal chemistry, and ALMANAC compound combinations that were experimentally tested against cell lines. In addition, STRING collects protein–protein interactions.

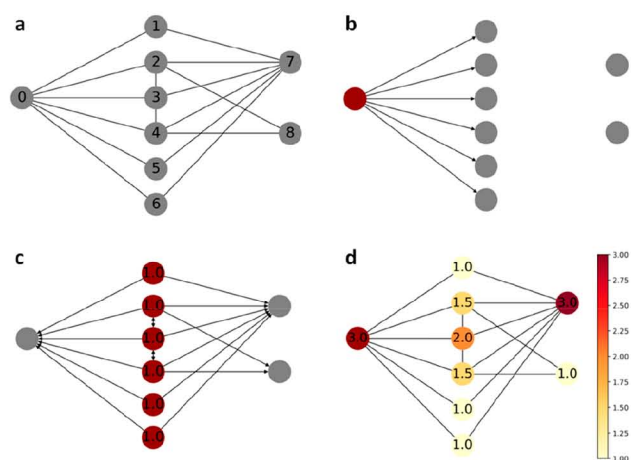


**Figure 3.** Identification of synergistic target pairs. Shown is a schematic representation of the workflow used for identifying synergistic target pairs, which involved data integration from different sources.

and was active on three lines. Compounds classified as “active” or “inactive” for a given screen were selected.

Active screening compounds were mapped to ChEMBL (release 23) and target annotations of detected compounds were collected exclusively on the basis of high-confidence activity data applying criteria established previously.<sup>21</sup> No potency threshold was applied to ChEMBL compounds to ensure that compounds weakly potent against given target(s) were considered in synergy analysis. Instead, it was important to apply high-confidence criteria to compound activity data, requiring, for example, exclusive consideration of clearly defined equilibrium constants ( $K_i$  values or  $IC_{50}$  values).

ALMANAC reports screening data for combinations of 104 compounds approved for cancer treatment on 60 cancer cell lines from the NCI60 screening panel.<sup>10</sup> About a third of all



**Figure 4.** Network propagation algorithm. Network propagation from a single node and perturbation scoring are illustrated. (a) Shows a model network consisting nodes 0 to 8 and (b) a perturbation originating from node 0 (red). (c) As a consequence, all directly connected neighbors are assigned a perturbation score of 1.0 (red). (d) Propagation continues from scored (first-layer) nodes with scores and for each nearest neighbor, a score of 0.5 is added (color-coded). Formally, for a given path with length  $d$ , nodes perturbed in the previous iteration add a score value of  $1/d$  to the score of their nearest neighbors. In this example, convergence is reached at a maximum path length of 2 when ranks of node scores remained constant.

tested compound combinations revealed synergistic effects. Known cancer targets were retrieved from the Therapeutic Target Database<sup>22</sup> and COSMIC.<sup>23</sup>

**Identification of Synergistic Target Pairs.** Two independent experimental data sources were used to identify synergistic target pairs. Compounds tested in NCI60 cell line screening assays were mapped to ChEMBL and annotated with targets on the basis of ChEMBL activity data. Targets linked exclusively to compounds that consistently caused cell death were omitted from target pair analysis because these targets might be essential for cell survival. Furthermore, for a given cell

line, NCI60/ChEMBL compounds were only considered if all other NCI60/ChEMBL compounds annotated with the same target(s) tested against the cell line were also active.

In addition, an independent source for identifying synergistic target pairs were screening data for compound combinations contained in ALMANAC. Only those compounds for which target annotations were available in ChEMBL were considered. From each compound combination, targets inhibited by both compounds were excluded because their inhibition did not result from synergy. The remaining targets were systematically combined, and the proportion of instances any given combination of targets was found to be synergistic was recorded. For example, if a compound from a given combination inhibited targets a, b, and c and the other compound inhibited targets a, d, and e, the following putative synergistic target pairs were obtained: (b,d), (b,e), (c,d), (c,e). In this example, both compounds had multitarget activity.

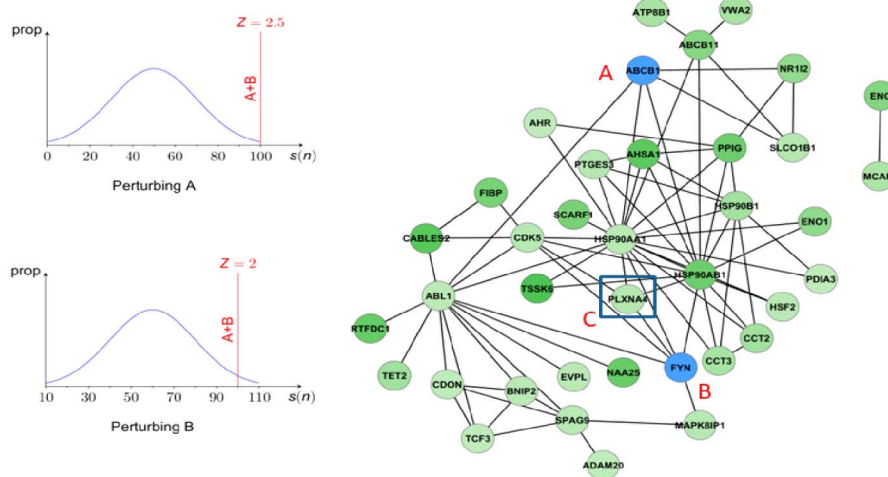
**Synergy Proportion and Enrichment.** For each target pair associated with an ALMANAC cell line screen, the proportion of synergistic and nonsynergistic compound combinations was determined, which yielded the “synergy proportion” for a target pair. A generally expected synergy proportion was calculated as the mean value of individual synergy proportions. The “synergy enrichment” for a target pair was then determined by subtracting the general synergy proportion (31.2%) from the corresponding synergy proportion:

synergy enrichment

$$= (\text{no. of synergistic compound combinations} / \text{no. of all combinations}) - 31.2 \quad (1)$$

For hypothesis testing,  $t$  tests for independent samples were carried out.

**Network Propagation Analysis.** For our analysis, we developed a network propagation algorithm with new iterative propagation and scoring schemes. The algorithm is illustrated in Figure 4.



**Figure 5.** Identification of interactors for a synergistic target pair. When the signal from two nodes A and B forming a synergistic target pair was simultaneously propagated a propagation score was obtained for another node C in the network. This score was compared to the distribution of scores expected to be observed for C if A plus any other node in the network or B plus any other node were propagated. C was termed an interactor of A and B if the obtained score fell above the mean of both distribution of expected propagation values for C when either A or B plus any other node in the network were perturbed. Z-scores measure the distance of a value from the mean in standard deviations.



A network  $G$  representing PPIs consists of nodes (proteins) and edges (pairwise interactions). Given a set  $M \subset G$ , these nodes can be “perturbed” by inducing a signal at node(s) that propagates to the remaining nodes in the network. In the first step, for each node  $m \in M$ , all nearest (first-order) neighbors obtain a score of 1.0. If a node  $n \in G$  is a direct neighbor of more than one node in  $M$ , then node  $n$  receives an additive score of 1.0 from each node in  $M$ . In the next step, nodes with scores propagate the signal, resulting in paths of length 2. Second-order neighbors will receive 1/2 of the original score. During the following iteration(s), additive scores are updated accordingly. Formally, for a given path with length  $d$ , nodes perturbed in the previous iteration add a score value of  $1/d$  to the score of their nearest neighbors. Iterations are carried out until a predefined path length is reached. In our calculations, a path length of 3 was consistently applied because the mean shortest path length between nodes was 3.65 for the global PPI network and 2.82 for the synergy-centered subnetwork. In its current implementation, our algorithm fully reproduced previously reported network heat propagation examples.<sup>24</sup>

For estimating the “propagation noise” distribution at each node, 1000 samples of 50 nodes each were randomly selected. The propagation noise distribution is an empirical distribution of expected scores at a given node resulting from random perturbation.

Perturbation scores were considered statistically significant if they reached or exceeded the top 0.1% of a noise distribution ( $p < 1 \times 10^{-4}$ ).

Nodes with scores significantly exceeding noise distributions were termed “interactors”, as illustrated in Figure 5. For calculating an “interactor score” for a given node in the network, Z-transformed distributions of scores obtained for propagating each target in the pair simultaneously with any other target in the network were calculated. Then, the actual values obtained by propagating the target pair were mapped to the Z-distribution and the Z-scores were multiplied. Since a Z-distribution has a mean of 0 and standard deviation of 1, large values of interactor score were obtained in the presence of large deviations from the mean in both distributions, with positive values indicating deviations in the same direction.

For network generation, analysis, and representation, STRING 10.5, Cytoscape,<sup>25</sup> and NetworkX 2.1<sup>26</sup> were used.

## RESULTS AND DISCUSSION

**Identification of Synergistic Target Pairs.** As detailed in the Materials and Methods section, Figure 2 summarizes the analysis scheme leading to the identification of synergistic target pairs, which can be divided into three main stages: (i) data collection, (ii) integration, and (iii) hypothesis assessment. Importantly, compound-dependent target synergy might be induced by individual small molecules with multitarget activity and/or by compounds that act synergistically in cell-based screens. Thus, central to the analysis concept is the data integration step involving compound-target mapping (Figure 3). Synergistic target pairs were derived on the basis of ChEMBL target annotations for individual compounds screened on the NCI60 panel and synergistic compound combinations from ALMANAC cancer cell line screens. Screening data for compound combinations permitted quantifying the proportion instances for which synergy was observed by coinhibition of two targets. Matching (consensus) target pairs obtained on the basis of both data sources yielded

high-confidence assignments for synergistic target pairs. In the following, the results of the analysis are discussed.

We first searched for the 40 998 NCI60 screening compounds in ChEMBL and detected 933 of them, which were termed “NCI60/ChEMBL” compounds. These 933 NCI60/ChEMBL compounds yielded a total of 43 391 target pairs that involved 162 targets. A subset of 317 compounds was found to be active against at least one cell line and inactive against at least another line. Inactivity against a cell line was required as a control criterion to rule out general toxicity.

The 317 NCI60/ChEMBL compounds that were active and inactive against at least one cell line were designated “cell-relevant” and further analyzed. A subset of 144 compounds was active against a maximum of two cell lines. These 144 compounds were designated “cell-selective”. On the basis of cell-relevant and cell-selective compounds, 13 187 and 470 unique target pairs were obtained, respectively. These target pairs involved 140 and 80 targets, respectively.

Next, the 104 compounds taken from ALMANAC were searched in ChEMBL and 54 compounds were identified. On the basis of these compounds, 16 304 unique target pairs were obtained involving 190 targets.

### Comparing Target Pairs from Different Sources.

Given the hypothetical synergistic target pairs that were identified, hypothesis testing was carried out by mapping NCI60/ChEMBL pairs to ALMANAC/ChEMBL pairs, hence determining consensus pair with high synergy proportion. Table 1 reports the mapping statistics. We found that 7959 of

Table 1. Target Pair Mapping Statistics<sup>a</sup>

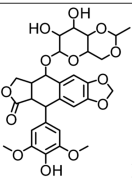
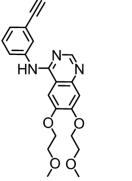
multitarget cpds	NCI60/ChEMBL pairs	consensus pairs	synergy enrichment
all	43391	9987	8.3%
cell-relevant	13187	7959	10.6%
cell-selective	470	164	0.3%

<sup>a</sup>Consensus pairs represented NCI60/ChEMBL pairs that were also found to be ALMANAC/ChEMBL pairs; cpds stands for compounds.

13 187 cell-relevant NCI60/ChEMBL target pairs (60.4%) mapped to ALMANAC/ChEMBL target pairs. Cell-relevant target pairs also included a small subset of 470 target pairs from cell-selective compounds, 164 of which mapped to ALMANAC/ChEMBL target pairs. For consensus target pair, the synergy proportion and synergy enrichment were calculated (see Materials and Methods). The small subset of cell-selective consensus pairs had negligible synergy enrichment close to zero. By contrast, cell-relevant consensus displayed a mean synergy enrichment of nearly 10% ( $p < 1 \times 10^{-5}$ ,  $t$  test). Thus, synergy enrichment was typically observed over multiple cell lines.

Consensus target pairs from different cell lines contained 735 unique target pairs and 102 unique targets, 41 of which were known cancer targets. We assigned high priority to consensus target pairs that were associated with more than five synergistic compound combinations and had a synergy proportion greater than 50%. A subset of 1681 target pairs from different cells lines met these criteria, which contained 137 unique target pairs and 51 unique targets, 21 of which were known cancer targets. These 137 prioritized target pairs included 22 pairs exclusively consisting of known cancer targets, 53 pairs exclusively consisting of other targets, and 62 “mixed” pairs. Thus, the latter pairs provided suggestions for

Table 2. Exemplary Target Pairs<sup>a</sup>

Multi-target compounds	Cell lines	Consensus target pair	Synergy proportion
 Etoposide	Prostate cancer (DU-145), leukemia (HL-60, MOLT-4)	TOP2A + NCOA3	67%
 Erlotinib	Kidney cancer (ACHN, SN12C), ovarian cancer (IGROV1, SK-OV-3) astrocytoma (SNB-19), melanoma (UACC-62)	erbB1 + SLCO1B1	79%

<sup>a</sup>For two multitarget compounds, cell-relevant consensus target pairs and their synergy proportions are reported. Target abbreviations: TOP2A, DNA topoisomerase 2-alpha; NCOA3, nuclear receptor coactivator 3; erbB1, epidermal growth factor receptor erbB1; SLCO1B1, solute carrier organic anion transporter 1B1.

previously unconsidered targets that may act synergistically with known cancer targets.

Table 2 provides exemplary cell-relevant consensus target pairs. The first example shows a target pair derived from target annotations of etoposide, which is a chemotherapeutic agent approved for the treatment of neoplastic disorders including lymphoma and nonlymphocytic leukemia. NCOA3 is a transcriptional coactivator of steroid hormone receptors and its role in breast cancer has been discussed.<sup>27</sup> Prostate cancer is another type of malignancy that is highly dependent on steroid hormone regulation.<sup>28</sup> The second example is a consensus pair identified for erlotinib consisting of epidermal growth factor receptor erbB1 and SLCO1B1, a membrane transporter. Erlotinib is another chemotherapeutic agent with broad spectrum applications in oncology.

We note that the identification of consensus target pairs does not prove the modes of action of implicated compound and drugs but provides hypotheses for further investigation. In the absence of experimental data, as an additional computational analysis step, the synergism of consensus target pairs was further explored through propagation analysis of relevant targets in PPI networks, as discussed in the following.

**Network Propagation Analysis.** For networks of large size such as a global PPI network, currently available propagation algorithms<sup>24</sup> become computationally essentially infeasible. Therefore, we have developed a computationally inexpensive algorithm (see [Materials and Methods](#)), which enabled us to carry out our analysis.

The 137 prioritized consensus pairs contained 51 unique targets that were further analyzed by network to determine if there might be functionally relevant relationships between them and/or involvement in well-defined interaction pathways.

**Global Protein–Protein Interaction Network.** As a starting point for exploring potential relationships, the 51 targets were mapped to a global human protein–protein interaction (PPI) network in which nodes represented proteins and edges pairwise interactions. The global PPI network was taken from STRING and contained 15 154 proteins. Fifty of the 51 targets were found to have high-confidence interactions with others. These 50 targets were densely connected. Based on STRING statistics for this network, on average 21 PPIs

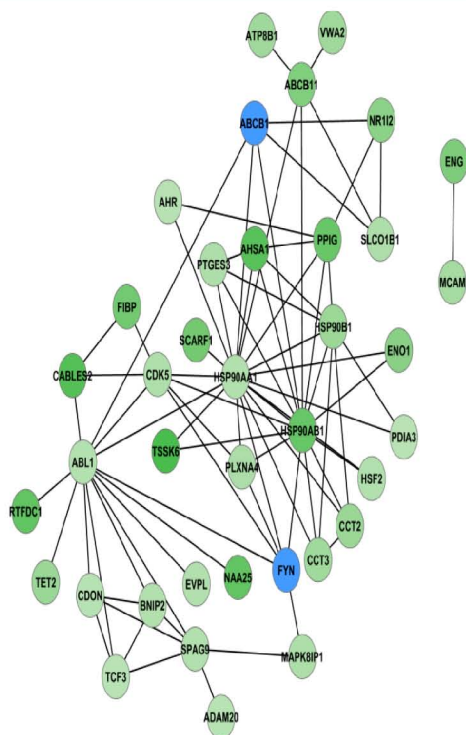
(edges) per target were expected. However, for the 50 targets from synergistic pairs, on average 70 edges were detected ( $p$ -value  $< 1 \times 10^{-16}$ ). Among interaction partners of these targets, there was an enrichment of proteins involved in cell–cell signaling (14 targets), the MAP kinase cascade (11), and the PI3 kinase-Akt signaling pathway (six targets). Given the large size of the global PPI network and small sample size of prioritized targets, conventional network statistics such as shortest path analysis were not applicable in a meaningful way to characterize relationships between these targets and others in the network. Therefore, network propagation analysis was carried out.

**Network Propagation.** Network propagation algorithms evaluate neighborhoods of targets and also quantify the distribution of propagation information throughout an entire network.<sup>24</sup> The underlying idea is to introduce a “perturbation” (signal) at a given node and quantify the progression of the signal (here propagation information) from this node on the basis of perturbation scores. This makes it possible, for example, to identify nodes making largest contributions to interaction pathways throughout a network. When considering a synergistic target pair, the newly developed network propagation method makes it possible to determine noise distributions for each target from a pair and any other target in the network.

**Synergy-Centric Subnetwork.** The algorithm was first applied to the global PPI network to determine if targets from consensus pairs were randomly distributed across the network or enriched in subnetworks. Therefore, 1000 samples of 50 nodes each were randomly drawn from the network and used to calculate the propagation noise distribution for each node in the network. From the distribution, a  $p$ -value for perturbation scores was calculated for each node. Then, the 50 consensus pair targets were simultaneously perturbed, propagation scores calculated, and nodes from the network selected if they obtained statistically significant scores ( $p < 1 \times 10^{-4}$ ). The resulting subnetwork comprised 858 nodes (only ~6% of the original PPI network) but included 32 targets from 86 synergistic pairs. Nearly 17% of nodes in the subnetwork represented known cancer targets, corresponding to a 4-fold enrichment compared to the global PPI network. Furthermore,

for 47 of the 50 targets pairs, also including 15 targets not contained in the subnetwork, perturbation scores were obtained that exceeded any score in the corresponding random noise distribution. Based on STRING statistics, an interaction network comprising 858 was expected to contain 1560 edges, but the synergy-centric subnetwork contained 6832 edges ( $p < 1 \times 10^{-16}$ ). Taken together, these findings revealed that targets in synergistic pairs had substantial influence on PPI propagation and were mostly contained in a confined PPI subnetwork, hence suggesting the presence of functional relationships.

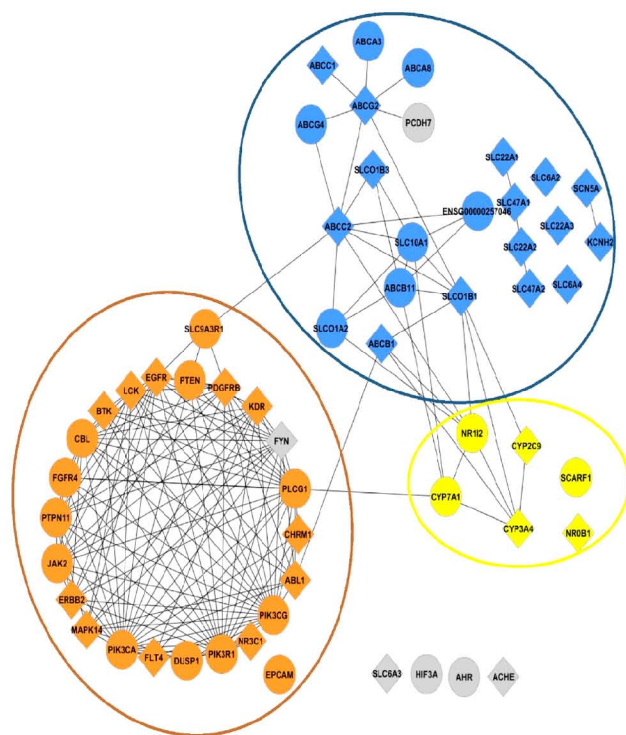
**Pathways and Functional Implications.** To further explore such relationships, the subnetwork was subjected to propagation analysis. In this case, the 32 remaining consensus pair targets were individually perturbed and for each of the 86 synergistic pairs it was determined whether statistically significant perturbation scores for both targets from the pair other shared protein partners were obtained. Proteins with statistically significant scores for both targets were classified as interactors (see [Materials and Methods](#)). On the basis of these calculations, 85 synergistic pairs were identified having one or more shared interactor nodes, which defined pathways between node clusters or pathways involving targets from consensus pairs. [Figure 6](#) shows a representative example. In this cluster, heat shock protein 90 (HSP90) isoforms



**Figure 6.** PPI cluster with consensus pair targets. Shown is an exemplary cluster containing a consensus pair (blue, FYN oncogene and ATP-binding cassette B1 protein) and interactor nodes. The cluster was extracted from the synergy-centric network. Interactor nodes are color-coded based on their combined interaction scores for both targets in synergistic pairs using a spectrum from green (highest score) to white. Such clusters represent local communities in PPI networks and are formed by proteins that are typically functionally related (corresponding to the “guilt-by-association” principle that is often applied in network analysis).

represented hubs and are interactor nodes for targets from consensus pairs. However, other nodes displayed larger enrichment in the propagated signal than HSP90. This was the case because chaperones such as HSP90 interact promiscuously with many proteins in the PPI network and are not specifically relevant for targets from synergistic pairs.

We identified 25 interactors that were associated with at least half of the consensus pairs, forming interactions with 32 unique targets in synergistic target pairs. [Figure 7](#) shows these frequent interactors, associated targets in synergistic pairs, and interactions they form.



**Figure 7.** Frequent interactors. Shown are interactors (circles) found in network clusters of at least 50% of the consensus pairs. Targets from consensus pairs are represented as diamonds. Nodes are colored by functional annotations including proteins involved in cancer and apoptosis (orange), membrane transport (blue), or steroid metabolism (yellow). Nodes of proteins without high-confidence functional annotations are shown in gray.

As can be seen, interactors and targets from consensus pairs form well-defined modules that were highly enriched with proteins having similar functions. One of two large modules was formed by proteins implicated in cancer and apoptosis and the other by proteins involved in membrane transport. These observations provided further support for the presence of functional relationships between targets from synergistic pairs and preferred interactors. The identification of interactors depended on the algorithm presented herein

## CONCLUDING REMARKS

In this study, we have systematically identified and rationalized synergistic target pairs in cancer cell lines on the basis of compound activity data from different sources and network propagation analyses. While compound synergy has been intensely studied, target synergy has thus far not been explored computationally. Our approach prioritized target pairs that



caused death of cancer cell lines when inhibited in combination, while inhibition of individual targets had no effect. Identified target pairs must be considered within their cell line context. To ensure a high level of confidence of the analysis, target pair hypotheses were first derived from compounds with multitarget activity and then evaluated by mapping to target pairs that were independently derived from synergistic compound combinations. The data-driven analysis identified more than 13 000 consensus target pairs. Among these synergistic target pairs, more than 700 were recurrent, which implicated a total of 102 unique targets involved in synergistic effects. A subset of 137 target pairs involving 51 unique targets (including 21 known cancer targets) had strong compound and synergy proportion support, thus providing focal points for follow-up investigations. Mapping of these targeted to a global PPI interaction network and iterative network propagation analysis using a novel algorithm provided substantial support for the presence of functional relationships between these targets and interactor targets. To enable follow-up analyses, Tables S1 and S2 of the Supporting Information report the 137 prioritized synergistic target pairs and specify the targets forming them. These targets pairs should be further experimentally validated, for instance with knockout and knock-down studies.

## ■ ASSOCIATED CONTENT

### Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jcim.9b00036.

Supplementary Tables S1 and S2 report targets in synergistic pairs with UniProt and STRING identifiers and prioritized synergistic target pairs, respectively (PDF)

## ■ AUTHOR INFORMATION

### Corresponding Authors

\*Email: [bajorath@bit.uni-bonn.de](mailto:bajorath@bit.uni-bonn.de) (J.B.)

\*Email: [medinajl@unam.mx](mailto:medinajl@unam.mx) (J.L.M.-F.).

### ORCID

Jürgen Bajorath: 0000-0002-0557-5714

### Notes

The authors declare no competing financial interest.

## ■ ACKNOWLEDGMENTS

The authors thank Martin Vogt, Filip Miljković, Swarit Jasial, and Erik Gilberg for helpful discussions. J.J.N. is grateful to Consejo Nacional de Tecnología, Mexico (CONACyT), for a scholarship (grant no. 622969) and to the German Academic Exchange Service (DAAD) for a short-term research grant (program no. 53378443).

## ■ DEDICATION

This work is dedicated to Gerald Maggiora on the occasion of his 80th birthday.

## ■ REFERENCES

- (1) Berenbaum, M. C. What is synergy? *Pharmacol. Rev.* **1989**, *41*, 93–141.
- (2) Jia, J.; Zhu, F.; Ma, X.; Cao, Z. W.; Li, Y. X.; Chen, Y. Z. Mechanisms of drug combinations: interaction and network perspectives. *Nat. Rev. Drug Discovery* **2009**, *8*, 111–128.

- (3) Kong, D.-X.; Li, X.-J.; Zhang, H.-Y. Where is the hope for drug discovery? Let history tell the future. *Drug Discovery Today* **2009**, *14*, 115–119.

- (4) Veldstra, H. Synergism and potentiation with special reference to the combination of structural analogues. *Pharmacol. Rev.* **1956**, *8*, 339–387.

- (5) Goldin, A.; Mantel, N. The employment of combinations of drugs in the chemotherapy of neoplasia: a review. *Cancer Res.* **1957**, *17*, 635–654.

- (6) Odds, F. C. Synergy, antagonism, and what the checkerboard puts between them. *J. Antimicrob. Chemother.* **2003**, *52*, 1–1.

- (7) Nesbitt, S. D. Antihypertensive combination therapy: optimizing blood pressure control and cardiovascular risk reduction. *J. Clin. Hypertens.* **2007**, *9*, 26–32.

- (8) Setter, S. M.; Iltz, J. L.; Thams, J.; Campbell, R. K. Metformin hydrochloride in the treatment of type 2 diabetes mellitus: a clinical review with a focus on dual therapy. *Clin. Ther.* **2003**, *25*, 2991–3026.

- (9) Lehár, J.; Krueger, A. S.; Avery, W.; Heilbut, A. M.; Johansen, L. M.; Price, E. R.; Rickles, R. J.; Short, G. F.; Staunton, J. E.; Jin, X.; Lee, M. S.; Zimmermann, G. R.; Borisy, A. A. Synergistic drug combinations tend to improve therapeutically relevant selectivity. *Nat. Biotechnol.* **2009**, *27*, 659–666.

- (10) Holbeck, S. L.; Camalier, R.; Crowell, J. A.; Govindharajulu, J. P.; Hollingshead, M.; Anderson, L. W.; Polley, E.; Rubinstein, L.; Srivastava, A.; Wilsker, D.; Collins, J. M.; Doroshow, J. H. The National Cancer Institute ALMANAC: a comprehensive screening resource for the detection of anticancer drug pairs with enhanced therapeutic activity. *Cancer Res.* **2017**, *77*, 3564–3576.

- (11) Preuer, K.; Lewis, R. P. I.; Hochreiter, S.; Bender, A.; Bulusu, K. C.; Klambauer, G. DeepSynergy: predicting anti-cancer drug synergy with deep learning. *Bioinformatics* **2018**, *34*, 1538–1546.

- (12) Bansal, M.; Yang, J.; Karan, C.; Menden, M. P.; Costello, J. C.; Tang, H.; Xiao, G.; Li, Y.; Allen, J.; Zhong, R.; Chen, B.; Kim, M.; Wang, T.; Heiser, L. M.; Realubit, R.; Mattioli, M.; Alvarez, M. J.; Shen, Y.; Gallahan, D.; Singer, D.; Saez-Rodriguez, J.; Xie, Y.; Stolovitzky, G.; Califano, A. A community computational challenge to predict the activity of pairs of compounds. *Nat. Biotechnol.* **2014**, *32*, 1213–1222.

- (13) Beijersbergen, R. L.; Wessels, L. F. A.; Bernards, R. Synthetic lethality in cancer therapeutics. *Annu. Rev. Cancer Biol.* **2017**, *1*, 141–161.

- (14) Berlow, N.; Davis, L. E.; Cantor, E. L.; Séguin, B.; Keller, C.; Pal, R. A new approach for prediction of tumor sensitivity to targeted drugs based on functional data. *BMC Bioinf.* **2013**, *14*, No. e239.

- (15) Szalay, K. Z.; Csermely, P. Perturbation centrality and turbine: a novel centrality measure obtained using a versatile network dynamics tool. *PLoS One* **2013**, *8*, No. e78059.

- (16) Tang, J.; Karhinen, L.; Xu, T.; Szwajda, A.; Yadav, B.; Wennerberg, K.; Aittokallio, T. Target inhibition networks: predicting selective combinations of druggable targets to block cancer survival pathways. *PLoS Comput. Biol.* **2013**, *9*, No. e1003226.

- (17) He, L.; Tang, J.; Andersson, E. I.; Timonen, S.; Koschmieder, S.; Wennerberg, K.; Mustjoki, S.; Aittokallio, T. Patient-customized drug combination prediction and testing for T-cell prolymphocytic leukemia patients. *Cancer Res.* **2018**, *78*, 2407–2418.

- (18) Shoemaker, R. H. The NCI60 human tumour cell line anticancer drug screen. *Nat. Rev. Cancer* **2006**, *6*, 813–823.

- (19) 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*, D945–D954.

- (20) Szklarczyk, D.; Franceschini, A.; Wyder, S.; Forslund, K.; Heller, D.; Huerta-Cepas, J.; Simonovic, M.; Roth, A.; Santos, A.; Tsafou, K. P.; Kuhn, M.; Bork, P.; Jensen, L. J.; von Mering, C. STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.* **2015**, *43*, D447–452.



(21) Hu, Y.; Bajorath, J. Influence of search parameters and criteria on compounds selection, promiscuity, and pan assay interference characteristics. *J. Chem. Inf. Model.* **2014**, *54*, 3056–3066.

(22) Li, Y. H.; Yu, C. Y.; Li, X. X.; Zhang, P.; Tang, J.; Yang, Q.; Fu, T.; Zhang, X.; Cui, X.; Tu, G.; Zhang, Y.; Li, S.; Yang, F.; Sun, Q.; Qin, C.; Zeng, X.; Chen, Z.; Chen, Y. Z.; Zhu, F. Therapeutic Target Database update 2018: enriched resource for facilitating bench-to-clinic research of targeted therapeutics. *Nucleic Acids Res.* **2018**, *46*, D1121–D1127.

(23) Forbes, S. A.; Beare, D.; Boutselakis, H.; Bamford, S.; Bindal, N.; Tate, J.; Cole, C. G.; Ward, S.; Dawson, E.; Ponting, L.; Stefancsik, R.; Harsha, B.; Kok, C. Y.; Jia, M.; Jubb, H.; Sondka, Z.; Thompson, S.; De, T.; Campbell, P. J. COSMIC: somatic cancer genetics at high-resolution. *Nucleic Acids Res.* **2017**, *45*, D777–D783.

(24) Cowen, L.; Ideker, T.; Raphael, B. J.; Sharan, R. Network propagation: a universal amplifier of genetic associations. *Nat. Rev. Genet.* **2017**, *18*, 551–562.

(25) Shannon, P.; Markiel, A.; Ozier, O.; Baliga, N. S.; Wang, J. T.; Ramage, D.; Amin, N.; Schwikowski, B.; Ideker, T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **2003**, *13*, 2498–2504.

(26) Hagberg, A. A.; Schult, D. A.; Swart, P. J. Exploring network structure, dynamics, and function using NetworkX. In: *Proceedings of the 7th Python in Science Conference*; Varoquaux, G.; Vaught, T.; Millman, J.; Eds.; Pasadena, CA, USA, 2008; Vol. 1, pp 11–15.

(27) Amelio, I.; Lisitsa, A.; Knight, R. A.; Melino, G.; Antonov, A. V. Polypharmacology of approved anticancer drugs. *Curr. Drug Targets* **2017**, *18*, 534–543.

(28) Feldman, B. J.; Feldman, D. The development of androgen-independent prostate cancer. *Nat. Rev. Cancer* **2001**, *1*, 34–45.

## **5. Análisis de compuestos con actividad epigenética**

### **Ideas clave**

#### **Marco conceptual**

En esta sección se presentan dos investigaciones exploratorias acerca del potencial de desarrollar moléculas polifarmacológicas contra dianas epigenéticas. El primer artículo es una revisión que se enfoca en estudiar la diversidad y similitud de las bibliotecas de inhibidores reportados contra dianas epigenéticas; la finalidad era estudiar el espacio químico y la diversidad de estas bibliotecas. El segundo artículo es una extensión del primero, en cuanto a que también se investigan las relaciones estructura-actividad en cada una de las bibliotecas. Proponemos que si las bibliotecas de dos dianas son parecidas, y además se conservan las relaciones estructura-actividad, entonces hay potencial de desarrollar o identificar moléculas polifarmacológicas que actúen contra ambas dianas.

#### **Datos utilizados**

Se analizan las bibliotecas químicas de inhibidores de 52 dianas epigenéticas. La información estaba disponible en diferentes bases de datos públicas, por lo que fue integrada y armonizada. Se incluyen más de 9 familias de dianas epigenéticas (Tab. 1, Art. 1).

#### **Metodología y resultados**

Encontramos que si las bibliotecas de inhibidores de dos dianas epigenéticas son parecidas, entonces las dianas también son más parecidas en su su función y estructura (Fig. 4, Art. 1). También se realizó un análisis de núcleos base de Bemis y Murcko para identificar motivos estructurales asociadas con mayor actividad y selectividad contra distintas dianas (Fig. 3, Art. 1). Se identificaron regiones en el espacio químico donde se conservan mejor las relaciones estructura-actividad (Fig. 1, Art. 2). Concluimos que es factible desarrollar inhibidores polifarmacológicos epigenéticos.



# Insights from pharmacological similarity of epigenetic targets in epipolypharmacology

J. Jesús Naveja<sup>1,2</sup> and José L. Medina-Franco<sup>1</sup>



<sup>1</sup> Facultad de Química, Departamento de Farmacia, Universidad Nacional Autónoma de México, Avenida Universidad 3000, Mexico City 04510, Mexico  
<sup>2</sup> PECEM, Facultad de Medicina, Universidad Nacional Autónoma de México, Avenida Universidad 3000, Mexico City 04510, Mexico

As the number of compounds tested against epigenetic targets grows, exploration of the possible associations in chemical space among these targets could lead to the identification of new drugs or new designs of epipolypharmacological molecules. Thus, here we review compound–epitarget associations of public databases. Specifically, we explore the structure–multitarget activity relationships and diversity of over 7000 compounds tested against 52 epigenetic-related targets. We found that, whereas inhibitors of histone deacetylases and other epigenetic targets are clustered in the chemical space, the chemical space of inhibitors of different DNA methyltransferases (DNMTs) did not overlap, indicating DNMT selectivity. These and other compound–epitarget relationships discussed here could be useful for both drug repurposing and the rational design of epipolypharmacological compounds.

## Introduction

The definition of ‘epigenetics’ is still a topic of debate and involves ill-defined processes, such as cell memory [1]. Many mechanisms occur simultaneously within the cell, contributing to cellular adaptation through the regulation of transcription patterns, and mediating the stability and inheritance of these patterns [2]. The most-studied epigenetic events are signaled through DNA and histone modifications [3], as well as transcription factors [4] and noncoding RNA transcripts [5]. Notably, the interplay among these mechanisms shapes genic expression and, therefore, the cell phenotype. Moreover, complex regulation is involved at every level. For instance, the expression of histone post-translational modifications (PTMs) is regulated through ‘writers’ (i.e., enzymes that place the PTM), ‘readers’ (i.e., proteins that recognize the PTM and transduce the signal to other protein mediators), and ‘erasers’ (i.e., enzymes that catalyze the removal of the PTM) [6–8]. Given that different PTMs can coexist in either the same or neighboring histones, and these combinations result in different chromatin responses, a ‘histone code’ thus emerges [9]. DNA methylation follows a similar mechanism, where DNMTs catalyze the addition

of methyl groups to CpG sites [10], whereas methyl-binding domains (MBDs) present in many proteins act as ‘readers’ [11], and ten-eleven translocation (TET) enzymes oxidize methylcytosines, leading to demethylation [12]. The regulation of other epigenetic mechanisms is also as intricate [2].

The plasticity and dynamism of epigenetic features place them as interesting pharmacological targets for many chronic diseases in which the cell phenotype remains perturbed on a long-term basis [13]. Nevertheless, fewer than ten drugs in the market have been accepted for their direct pharmacodynamic effects on epigenetic targets, such as DNMT and histone deacetylase (HDAC) inhibitors. Most of these drugs have antineoplastic clinical indications against hematological malignant or premalignant processes, although some are currently in advanced clinical trials against solid tumors [14,15]. Moreover, many other pharmacological agents exert epigenetic actions that might elicit adverse or therapeutic side effects that are not exclusively oncology related, thereby opening a path for drug repurposing [16–19].

A general classification of epidrugs is to consider them as either broad reprogrammers or targeted therapies [15]. Among the broad reprogrammers are DNMT, bromodomain and extra terminal (BET), and HDAC inhibitors. These agents have wide and dramatic effects on gene expression and effectively alter the epigenetic cell

Corresponding authors: Naveja, J.J. (naveja@comunidad.unam.mx), Medina-Franco, J.L. (medinajl@unam.mx)

signature. Targeted epigenetic therapies take advantage of the aberrant physiology of some cancer cells, which would make them more susceptible than normal cells to this kind of therapy [15]. For an in-depth review of epidrugs in clinical trials, see Ref. [20].

Other potential applications for epidrugs arise in cardiovascular, neurological, and metabolic diseases, which tend to have complex phenotypes and epigenetic dysregulations [21]. For instance, BET inhibitors have already been tested in preclinical studies against heart failure, inflammatory processes, and HIV reactivation, with promising results [22–24]. Furthermore, HDAC inhibitors have had promising results in murine models of Alzheimer's disease [25]. Related to metabolic diseases, some advances have resulted from studying epigenetic targets for diabetes and obesity treatments, particularly HDACs, histone acetyltransferases (HATs), DNMTs, and protein arginine methyltransferase (PRMTs) [26].

Interestingly, each epigenetic process can be pharmacologically approached at different regulatory levels, sometimes with variable results [27]. In this regard, researchers are described the molecular libraries of compounds associated with a variety of epigenetic targets [28–30], contributing to the exploration of the epigenetic relevant chemical space (ERCS).

Combinations of the inhibition of epigenetic pathways can lead to unpredictable and nonadditive results [31]. However, since complex diseases, such as cancer, are multifactorial and involve the dysregulation of many pathways, therapies aiming at more than a single target might be beneficial [32]. Furthermore, given that single molecules often show polypharmacology (i.e., are able to act on more than one target), it would be of interest to identify molecules with multiple epigenetic targets [33]. Structure–multiple activity relationships (SmART) have emerged for the study of polypharmacology in epigenetics [34,35].

Here, we report a survey of a comprehensive epigenomics database assembled from data available in the public domain. In contrast to related approaches that find associations between epigenetic targets through their sequence similarities [36], the relationship among epigenetic targets is explored here through data provided by the chemical structures of their reported inhibitors. Both analyses look at the data from a different perspective and are complementary. The starting point of this survey was an epigenomics database that contains 7820 nonduplicate compounds, of which 3456 (44.2%) have information regarding more than one target. The database contains 16 102 compound–target associations, of which 15 887 (98.7%) have quantitative potency data associated with them. In terms of the degrees of polypharmacology (i.e., number of targets/compound) and polyspecificity (i.e., number of compounds/target in the database [37]), the mean targets/compound ratio is 2.1, and the mean of compounds/target is 268.4. The database contains associations with 60 epigenetic targets. However, only targets with at least ten active compounds were included for further discussion and, therefore, only 52 remain (Table 1). A summary of the statistics of the molecular descriptors calculated in this study is available in the Supplemental information online. The Supplemental information online also presents a complete description of the mining methods used throughout this review. Here, exemplary epidrugs are discussed in the context of major epigenetic targets of proven clinical relevance.

This review is organized in eight main sections. Section 1 discusses the profile of physicochemical properties of therapeutic relevance

for all the chemical compounds in the epigenomics database. Section 2 presents a survey of the chemical diversity using molecular fingerprints (that, in contrast to the molecular scaffolds discussed in Section 3, consider the entire molecular structure). Section 3 discusses briefly the content and diversity of the chemical scaffolds. The next section then analyses the global diversity of the data sets, integrating the diversity based on properties, fingerprints, and molecular scaffolds. In this section, consensus diversity plots (CDPs) are used, which are recently introduced chemoinformatic tools [38]. Section 5 focuses on the structure–activity relationships (SARs) of the epigenetic targets based on molecular scaffolds. In this section, we discuss whether there are molecular scaffolds enriched with active molecules. In the next section, we analyze the epigenetic data sets using the concept of database fingerprint (DFP), a novel condensed representation of compound databases [39]. Section 7 overviews the chemical space of the entire epigenomics database based on a visual representation of the space generated using DFPs. Lastly, Section 8 addresses another major aspect of compound data sets: their structural complexity [40]. Finally, summary conclusions and an outlook are presented. Target names are provided in Table 1.

### Physicochemical properties

Physicochemical properties are chemical features that provide insights that are usually relevant to drug discovery and lead optimization. A classical example are the properties used in the Lipinski Rule of 5 for oral bioavailability [41]. Here, the distribution of six of the most commonly used physicochemical properties is surveyed, namely: calculated logarithm of the partition coefficient (SlogP); topological polar surface area (TPSA); molecular weight (MW); hydrogen-bond donors (HBD); hydrogen-bond acceptors (HBA); and rotatable bonds (RB). The results are summarized in the Supplemental information online. Physicochemical properties of the epigenomics compound database are, in general, homogenous. EP300, DOT1L, DNMT3B, MGEA5, PRMT6, and WDR5 are remarkably different to the mean in many of these properties, suggesting these occur in a novel and underexplored region in the chemical space.

### Fingerprint and 3D shape-based diversity

Pairwise structural similarity among compounds in a database gives an idea of how diverse that database is. Namely, a higher median in this variable indicates lower diversity of the data set [29]. For all data sets, all pairwise structural similarity values are computed using 2D and 3D molecular representations. Molecular fingerprints are a widely used chemical representation, which is usually bidimensional, such as molecular access system (MACCS) keys and extended connectivity fingerprint, diameter 4 (ECFP4). Many studies have found different, although complementary, results when using 3D structural representations [22–25]. Throughout the article, we refer to the 3D similarity of the conformers as OMEGA-Rapid overlay of chemical structures (ROCS), whose calculation is described in more detail in the Supplemental information online. Figure 1 summarizes the 2D and 3D diversity of the data sets through the median similarity of MACCS keys and OMEGA-ROCS, respectively. The color indicates the family of the target and the number of compounds associated with it. According to Figure 1, KDMs (data points in brown) are diverse based on both the 3D (OMEGA-ROCS) and 2D (MACCS keys) representations. By contrast, HDACs have average diversity (i.e., are towards the

TABLE 1

Targets included in this survey.<sup>a</sup>

Target	Function	Families (HGNC)	Cluster (manually annotated)	Molecules	Scaffolds	% Active
BAZ2B	Acetylated histone reader	PHD finger proteins, methyl-CpG binding domain containing	BRD	53	27	25
BRD2	Histone PTM reader	NA	BRD	277	91	87
BRD3	Histone PTM reader	NA	BRD	263	89	95
BRD4	Histone PTM reader	NA	BRD	643	259	80
BRD9	Histone PTM reader	NA	BRD	13	9	77
BRPF1	Histone PTM reader	PHD finger proteins, PWWP domain containing	BRD	27	15	89
DNMT1	DNA methyltransferase	Zinc fingers CXXC-type, seven-beta-strand methyltransferase motif containing	DNMT	248	194	60
DNMT3A	DNA methyltransferase	PWWP domain containing	DNMT	47	30	55
DNMT3B	DNA methyltransferase	PWWP domain containing	DNMT	40	22	50
CREBBP	Histone acetyltransferase	Zinc fingers ZZ-type, lysine acetyltransferases	HAT	180	65	64
EP300	Histone acetyltransferase	Zinc fingers ZZ-type, lysine acetyltransferases	HAT	73	52	78
KAT2A	Histone acetyltransferase	Lysine acetyltransferases, ATAC complex, SAGA complex, GCN5 related N-acetyltransferases	HAT	27	20	41
KAT2B	Histone acetyltransferase	Lysine acetyltransferases, ATAC complex, SAGA complex, GCN5 related N-acetyltransferases	HAT	121	40	61
NCOA1	Histone acetyltransferase	Basic helix-loop-helix proteins, lysine acetyltransferases	HAT	634	568	22
NCOA3	Histone acetyltransferase	Basic helix-loop-helix proteins, lysine acetyltransferases, trinucleotide repeat containing	HAT	564	517	32
HDAC1	Histone deacetylase	Histone deacetylases class I, EMSY complex, NuRD complex, SIN3 histone deacetylase complex	HDAC	3304	1418	90
HDAC2	Histone deacetylase	Histone deacetylases class I, EMSY complex, NuRD complex, SIN3 histone deacetylase complex	HDAC	942	427	84
HDAC3	Histone deacetylase	Histone deacetylases class I	HDAC	854	395	80
HDAC4	Histone deacetylase	Histone deacetylases class IIA	HDAC	704	348	69
HDAC5	Histone deacetylase	Histone deacetylases class IIA	HDAC	235	150	58
HDAC6	Histone deacetylase	Histone deacetylases class IIB, protein phosphatase 1 regulatory subunits	HDAC	1706	697	86
HDAC7	Histone deacetylase	Histone deacetylases class IIA	HDAC	257	151	51
HDAC8	Histone deacetylase	Histone deacetylases class I, X-linked mental retardation	HDAC	1176	493	79
HDAC9	Histone deacetylase	Histone deacetylases class IIA	HDAC	209	123	55
HDAC10	Histone deacetylase	Histone deacetylases class IIB	HDAC	243	116	78
HDAC11	Histone deacetylase	Histone deacetylases class IV	HDAC	200	104	73
DOT1L	Histone lysine methyltransferase	Lysine methyltransferases, seven-beta-strand methyltransferase containing	HKM	81	36	58
EHMT1	Histone lysine methyltransferase	Lysine methyltransferases, ankyrin repeat domain containing, SET domain containing	HKM	23	19	44
EHMT2	Histone lysine methyltransferase	Lysine methyltransferases, ankyrin repeat domain containing, SET domain containing	HKM	110	75	56
KMT5A	Histone lysine methyltransferase	Lysine methyltransferases, SET domain containing	HKM	71	33	14
SMYD2	Histone lysine methyltransferase	Lysine methyltransferases, zinc fingers MYND-type, SET domain containing	HKM	13	11	77
KDM1A	Histone lysine demethylase	Lysine demethylases	KDM	453	153	54
KDM2A	Histone lysine demethylase	Lysine demethylases, PHD finger proteins, Zinc-fingers CXXC type, F-box and leucine rich repeat proteins	KDM	61	36	69
KDM3A	Histone lysine demethylase	Lysine demethylases	KDM	70	33	51
KDM4A	Histone lysine demethylase	Lysine demethylases, Tudor domain containing	KDM	153	85	50
KDM4C	Histone lysine demethylase	Lysine demethylases, Tudor domain containing	KDM	247	162	37
KDM4E	Histone lysine demethylase	Lysine demethylases	KDM	86	43	34
KDM5A	Histone lysine demethylase	Lysine demethylases, PHD finger proteins, AT-rich interaction domain containing, EMSY complex	KDM	105	74	62
KDM5C	Histone lysine demethylase	Lysine demethylases, PHD finger proteins, AT-rich interaction domain containing, X-linked mental retardation	KDM	50	30	74
CBX7	Histone methylation reader	Chromobox family	KMeR	160	97	44
L3MBTL1	Histone methylation reader	Zinc fingers C2CH-type, sterile alpha motif domain containing, MBT domain containing	KMeR	126	97	44
L3MBTL3	Histone methylation reader	Sterile alpha motif domain containing, MBT domain containing	KMeR	115	92	77

TABLE 1 (Continued)

Target	Function	Families (HGNC)	Cluster (manually annotated)	Molecules	Scaffolds	% Active
L3MBTL4	Histone methylation reader	Sterile alpha motif domain containing, MBT domain containing	KMeR	98	77	56
TP53BP1	Histone methylation reader	Tudor domain containing	KMeR	75	65	69
WDR5	Histone methylation reader	WD repeat domain containing	KMeR	64	21	73
CARM1	Histone arginine methyltransferase	Protein arginine methyltransferases	PRMT	73	38	63
PRMT1	Histone arginine methyltransferase	Protein arginine methyltransferases	PRMT	141	89	43
PRMT6	Histone arginine methyltransferase	Protein arginine methyltransferases	PRMT	33	18	61
PRMT8	Histone arginine methyltransferase	Protein arginine methyltransferases	PRMT	24	12	92
MAP3K7	Kinase	Mitogen-activated protein kinase kinase kinases	Other	147	100	96
MGEA5	Histone O-N-acetylglucosamine transferase	NA	Other	74	19	93
SMARCA2	Chromatin remodeler	NA	Other	220	187	6

<sup>a</sup>Their function, families by HGNC, manually annotated cluster, number of molecules associated and number of distinct scaffolds are summarized.

middle of the plot); DNMT1 and DNMT3A are more diverse in 2D than in 3D; lysine methylation readers (KMeR) and PRMTs are more diverse in 3D than in 2D; and bromodomains (BRDs) and DNMT3B tend to a (lower) diversity in both 2D and 3D.

### Molecular scaffolds: content and diversity

Table 1 includes the number of molecular scaffolds for each data set. The data revealed that many targets have high scaffold diversity (i.e., similar numbers of scaffolds and total compounds).

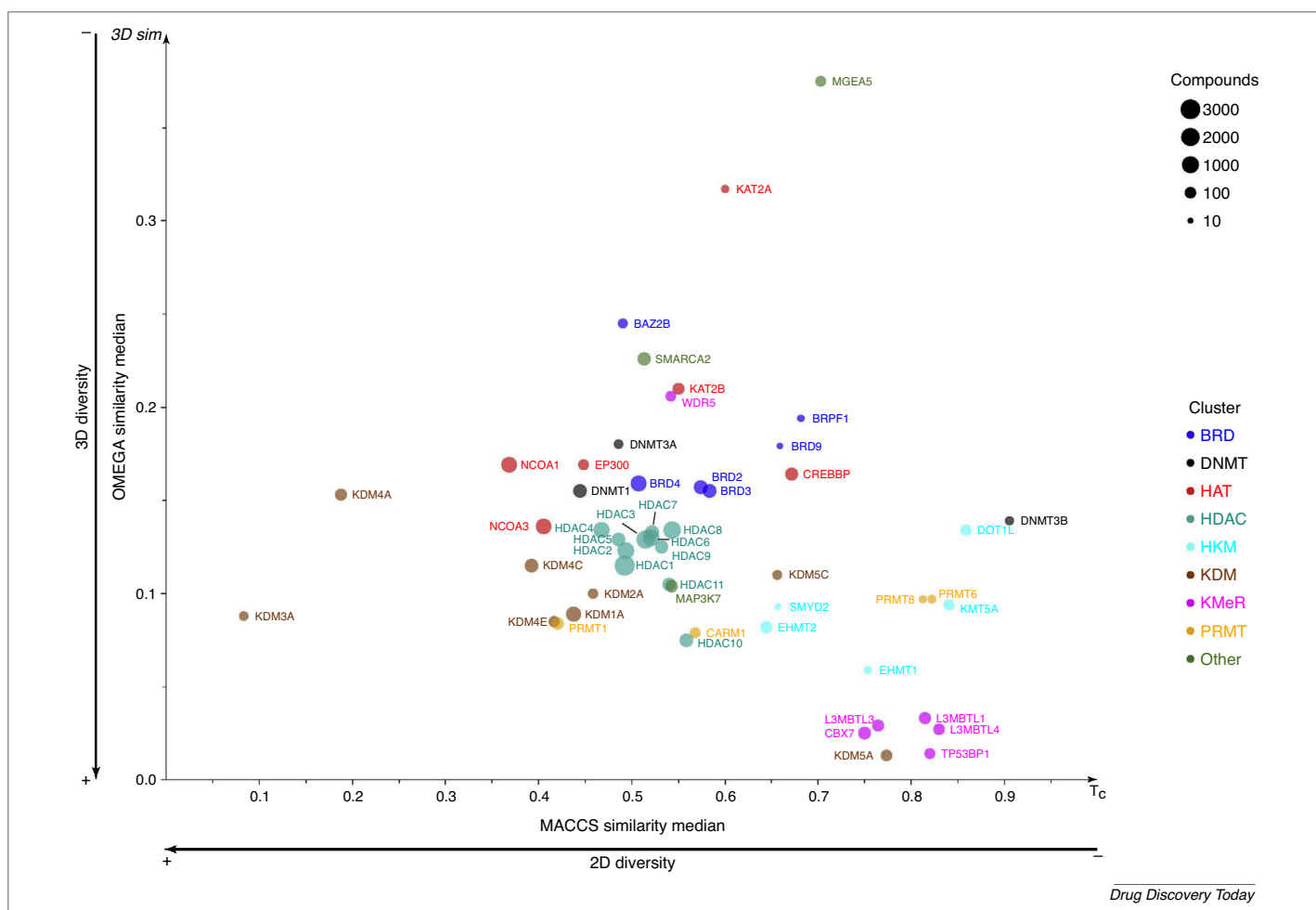


FIGURE 1

Consensus 3D diversity plot. The plot is 3D similarity of conformers (OMEGA) versus the median of molecular access system (MACCS) keys similarity. Note that, on both axes, a higher number denotes lower diversity. Dot size represents the chemical library size, and the color the family of the target. For definitions of abbreviations, please see the main text.



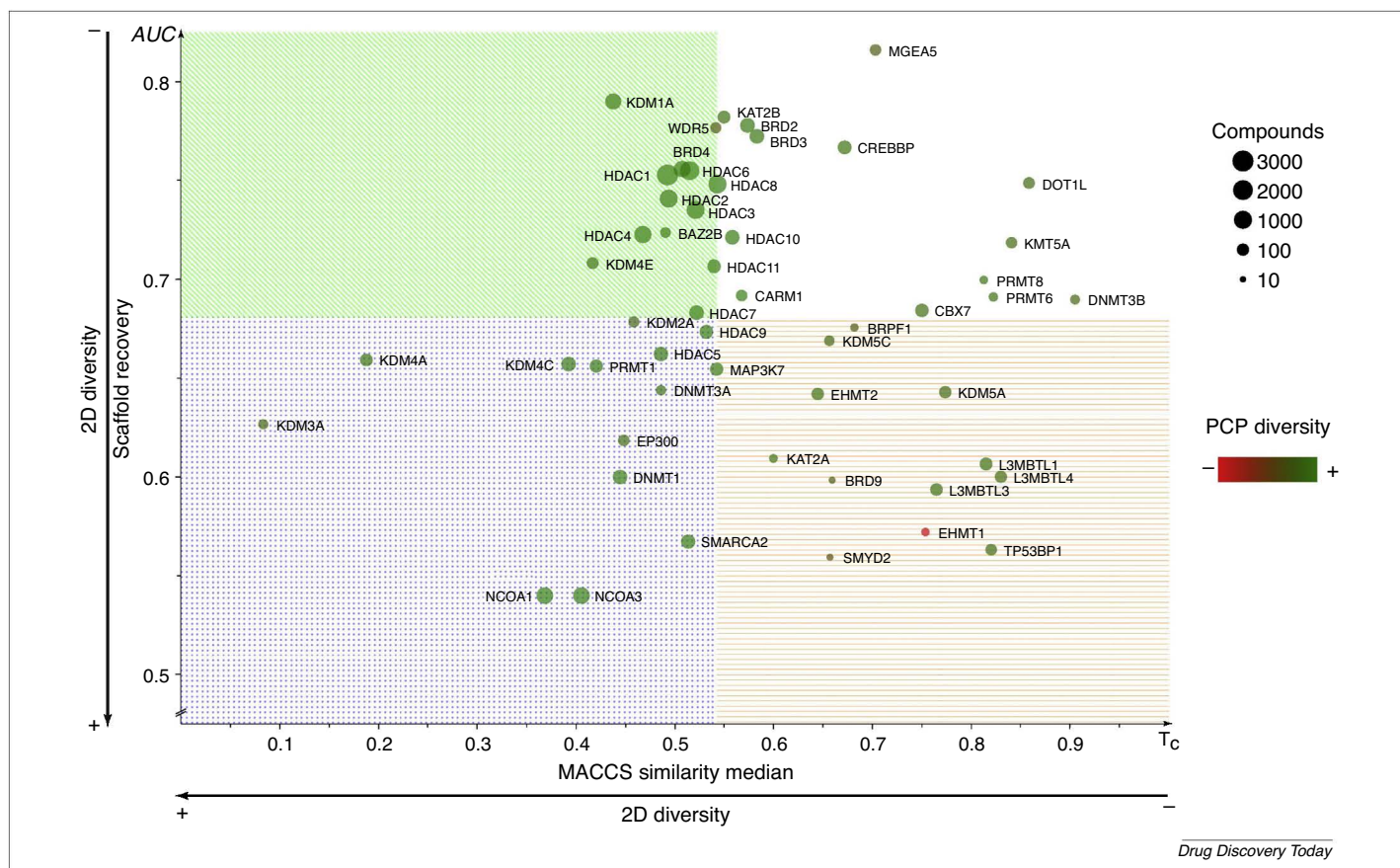
HDACs have larger libraries and, therefore, a lower scaffold/compound ratio is to be expected.

### Global molecular diversity

The total or global diversity of the data sets using multiple criteria (i.e., fingerprint, Bemis–Murcko scaffolds and physicochemical properties) can be analyzed simultaneously using CDPs [38]. Figure 2 shows this plot for the 52 epigenetic regulators in Table 1, comparing the diversity of the data sets in terms of their physicochemical properties (measured by Euclidean distance of Z-scaled MW, TPSA, HBD, HBA, SlogP, and RB; described further in the Supplemental information online), scaffolds (measured by the area under the scaffold recovery curve), and fingerprints (median of MACCS keys/Tanimoto similarity). In this plot, the data sets with the largest overall diversity are located on the left-bottom quadrant (high scaffold and fingerprint diversity). Notably, several KDMs and the two NCOAs showed the largest global diversity. By contrast, data sets in the upper-right quadrant (such as PRMTs, DOT1L, DNMT3B, and some BRDs) have low scaffold and fingerprint diversity. High-throughput screening strategies could be suggested for these last targets to increase the diversity of inhibitors tested.

### SAR analysis (based on molecular scaffolds)

Of the molecular representations considered in this survey, molecular scaffolds are, perhaps, the most easily interpretable for SAR studies. This section provides an overview of the SAR of each target based on calculated enrichment factors (EF) of the scaffolds of each target. Details of the calculation of EF is in the Supplemental information online. Statistically significant EF were identified through Chi-squared tests computing the *P* value from Monte Carlo simulations with 10 000 replicates [42]. Also, it was corrected for multiple hypothesis testing with the false discovery rate (Benjamini–Hochberg) method. In total, 57 statistically significant associations ( $P \leq 0.05$ ) were found, 27 of which with values of EF > 1 (positive enrichment) and the rest with values < 1 (negative enrichment). In addition, 15 targets and 38 scaffolds were involved in these associations. Figure 3 and Figure S1 in the Supplemental information online depict these scaffolds along with the targets against which they are enriched. Interestingly, compounds with scaffold **SCAFF4** shows selectivity towards HDAC4, compared with HDAC1, 2, and 6. Most of these associations involve HDACs, KDMs, or BRDs. The knowledge of enriched scaffolds might be useful for further drug design based on these targets.



**FIGURE 2**

Consensus diversity plot (CDP) summarizing the 2D diversity of the data sets. The x-axis indicates the median of the molecular access system (MACCS) keys (166-bits)/Tanimoto similarity, and the y-axis the area under the scaffold recovery curve. In both scales, higher values denote lower diversity. The size of the points is proportional to the number of compounds in the database. The color shift indicates the diversity of physicochemical properties, as measured by the median of the Euclidean distance of AMW, hydrogen bond donors (HBD), hydrogen bond acceptors (HBA), topological polar surface area (TPSA), calculated log partition coefficient (SlogP), and rotatable bonds (RB) (the greener the color, the more diverse the properties). For additional definitions of abbreviations, please see the main text.

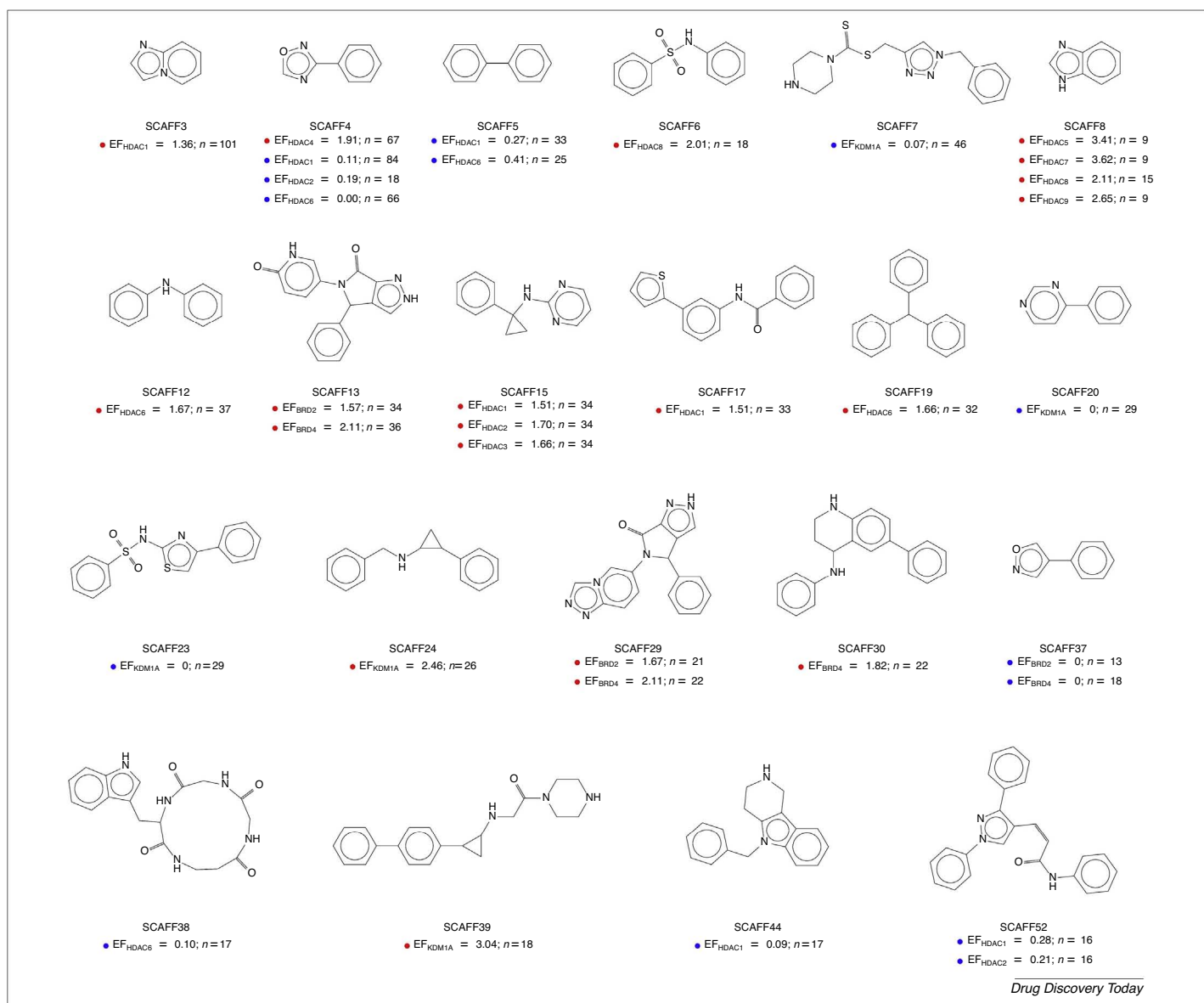


FIGURE 3

Significantly enriched scaffold-target associations (corrected  $P \leq 0.05$ ) with more than 14 distinct reported compounds (see Figure S1 in the Supplemental information online for the remaining compounds). A dot below the names of the structures indicates either positive (red) or negative (blue) enrichment for the corresponding target. Enrichment factors (EF) are shown and the number ( $n$ ) of total compounds with the scaffold that have been tested against the particular target.

Recently, Schneider et al. thoroughly analyzed the promiscuity statistics of many scaffolds in the ChEMBL database [43]. The main purpose of their work was to identify features that could predict whether a scaffold would be more specific and, therefore, better suited to target-focused design. To this end, they relied upon both physicochemical and complexity properties. By incorporating this information in our analysis, we found a correlation, albeit weak ( $r = 0.33$ ) between the scaffold promiscuity (information) reported for these scaffolds and the degree of polypharmacology calculated for our database.

### Epitargets DFPs

DFPs represent a novel approach that attempts to summarize the most common chemical motifs in a database [39]. This approach is particularly suited in this analysis because of the large number of

compounds data sets present in the epigenomic data set. DFPs were generated for each target by considering its associated chemical compound library. Table 2 highlights the most relevant information obtained through DFPs. For the internal validation of the usage of DFPs for describing targets, the recovery rate of the active compounds and the corresponding area under the ROC curve (AUC) were computed. This would be useful as an assessment of the virtual screening capabilities of per-target DFPs. The recovery of compounds was good for many epigenetic targets, as indicated by the areas under the ROC curves (a value of 0.5 equals random selection, and a value of 1 a perfect selection). The performance of DFP recovery rates was not influenced significantly by using either MACCS keys (166-bits) (a dictionary-based fingerprint) or Extended Connectivity diameter 4 (a radial fingerprint). Unsurprisingly, performance of DFP recovery rates anticorrelated with the diversi-



TABLE 2

## Summary of the DFP for each epitarget.

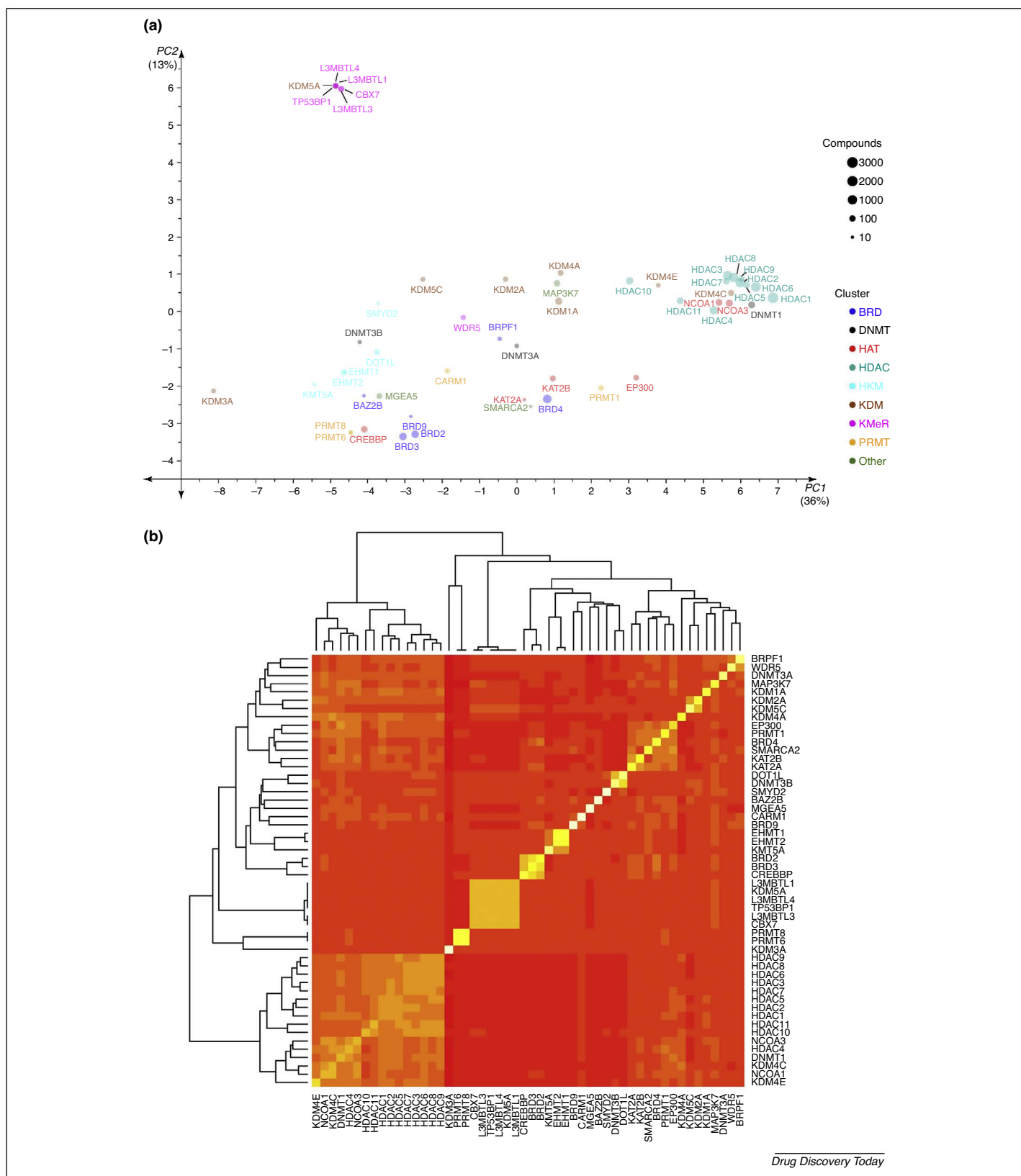
Target	Cluster	Number of '1' bits in DFP	Shannon entropy	Area under the ROC curve
BAZ2B	BRD	30	79	0.914
BRD2	BRD	23	115	0.902
BRD3	BRD	25	115	0.907
BRD4	BRD	13	138	0.847
BRD9	BRD	36	58	0.969
BRPF1	BRD	28	54	0.86
DNMT1	DNMT	14	153	0.429
DNMT3A	DNMT	26	100	0.928
DNMT3B	DNMT	53	62	0.978
CREBBP	HAT	32	103	0.911
EP300	HAT	12	144	0.690
KAT2A	HAT	16	83	0.855
KAT2B	HAT	10	93	0.946
NCOA1	HAT	11	145	0.503
NCOA3	HAT	10	145	0.506
HDAC1	HDAC	14	145	0.664
HDAC2	HDAC	14	134	0.690
HDAC3	HDAC	15	140	0.699
HDAC4	HDAC	11	128	0.659
HDAC5	HDAC	14	130	0.677
HDAC6	HDAC	16	143	0.715
HDAC7	HDAC	16	128	0.755
HDAC8	HDAC	15	145	0.701
HDAC9	HDAC	14	126	0.754
HDAC10	HDAC	18	127	0.771
HDAC11	HDAC	19	127	0.753
DOT1L	HKM	60	65	0.996
EHMT1	HKM	40	72	0.872
EHMT2	HKM	40	93	0.861
KMT5A	HKM	46	56	0.843
SMYD2	HKM	46	63	0.883
KDM1A	KDM	21	117	0.763
KDM2A	KDM	24	72	0.909
KDM3A	KDM	0	49	0.500
KDM4A	KDM	6	96	0.543
KDM4C	KDM	10	125	0.549
KDM4E	KDM	13	83	0.843
KDM5A	KDM	27	61	0.834
KDM5C	KDM	37	48	0.919
CBX7	KMeR	26	76	0.840
L3MBTL1	KMeR	27	49	0.985
L3MBTL3	KMeR	26	58	0.987
L3MBTL4	KMeR	27	47	0.999
TP53BP1	KMeR	27	48	0.999
WDR5	KMeR	29	93	0.662
CARM1	PRMT	41	100	0.753
PRMT1	PRMT	8	131	0.401
PRMT6	PRMT	36	64	1.000
PRMT8	PRMT	36	73	0.957
MAP3K7	Other	19	151	0.678
MGEA5	Other	18	71	0.993
SMARCA2	Other	14	79	0.975

ty measurements and, most importantly, with those based on 2D molecular fingerprints (correlation coefficients for AUC ECFP4 DFPs versus AUC MACCS DFPs = 0.87; versus number of compounds = -0.235; versus median of OMEGA-ROCS = 0.023; versus median of MACCS Tc = 0.747).

### Visual representation of chemical space

Chemical space is a broad concept related to mapping, analyzing, and visualizing chemical compounds, and has applications in

virtual screening and SAR analyses [44–48]. Figure 4a depicts a visual representation of the chemical space of the epitargets (i.e., an approach to the visual representation of the ERCS). The visual representation was generated by principal component analysis of the similarity matrix computed with DFPs. Perhaps unsurprisingly, HDACs form a distinct cluster. Indeed, with few exceptions (DNMTs, KDM3 and 4, CREBBP, and WDR5), targets grouped with others that have similar functions. The visualization of this map could guide discussions on the feasibility of repurposing com-



**FIGURE 4** Visual representation of the epigenetic chemical space. (a) Chemical space for the epigenetic targets. Principal components (PC) 1 and PC2 are shown, which capture 49% of the variance. Dot size represents the chemical library size, and the color the family of the target. The nearer the dots, the more similar their database fingerprints (DFPs) are. (b) Heatmap with hierarchical clustering of the epigenetic targets studied using data from the principal component analysis (PCA) of the DFPs similarity matrix, generated in R. Of note, histone deacetylases (HDACs) are clustered closer, whereas, for example, DNA methyltransferases (DNMTs) are not. Another cluster is formed by histone lysine demethylases (KDMs). A third cluster is formed by KDM5A, CBX7, TP53BP1, L3MBTL1, and L3MBTL3, most of which act as histone methylation readers. For additional definitions of abbreviations, please see the main text.

pounds among epitrgeters, as well as in the design of epipolypharmacological small molecules. For example, it is known that many molecules are not selective of a particular HDAC or BRD, but indeed are active against many of them. The chemical space in Figure 4a recapitulates this knowledge, and suggests that other proteins are more similar (at least pharmacologically) than was previously thought, for example DNMT1 to some histone methylation readers and HDACs.

Figure 4b shows a heatmap and hierarchical clustering obtained with the similarity matrix of the target DFPs. Interestingly, it brings HDACs together, as well as proteins with a methyl-lysine reading function (CBX7, TP53BP1, and L3MBTLs) and histone lysine demethylases (KDMs). In addition, the relative closeness of CARM1, some PRMTs, and DOT1L, and that of NCOA1 and NCOA3 (both in the heatmap and Figure 4), is in agreement with the results of Cabaye *et al.*, who used pocketome data for a related analysis of epigenetic targets [36].

By contrast, the results of this survey suggest that CREBBP and EP300, two structurally related histone acetyl-transferases, are pharmacologically different. Similarly, in Figure 4, it can be seen that DNMTs do not appear to be closely clustered. This could be indicative of selectivity among the libraries, although it is not conclusive. Also, it should be kept in mind that these data are restricted to publicly available sources.

### Structural complexity

Chemical complexity, albeit a concept challenging to quantify in a robust manner, is generally thought to be related to features such as target specificity and, therefore, fewer adverse effects [37,40]. Here, we survey the complexity of the data sets using two metrics widely used in drug discovery projects.

For the whole database, the mean number of chiral atoms was 0.71 but with significant variation among the data sets. For example, CARM1, CBX7, DNMT1, DNMT3B, DOT1L, EHMT1, EHMT2, EP300, HDAC8, KAT2B, KDM1A, MGEA5, PRMT1, PRMT6, and WDR5 had means >1, suggesting that inhibition of these targets requires more-complex molecules. Also, many of them exhibited a higher fraction of sp<sup>3</sup> carbon atoms (FCSP3) than the general mean (31%; see Supplemental information online for details). It has been described that approved drugs have a FCSP3 of 47%, which decreases for compounds in lower development stages [49]. Accordingly, because most of the compounds in the libraries used are in preclinical studies, the general mean is lower than this. However, CBX7, DNMT3B, DOT1L, EHMT2, KDM5A, KMT5A, L3MBTLs, MGEA5, SMYD2, and TP53BP1 have mean FCSP3 values higher than approved drugs, suggesting the higher selectivity of their inhibitors. Notably, there is no correlation among these complexity metrics and the degree of polypharmacology in this database. However, this could be a result of the high sparsity (~97%) of the matrix [37].

### Concluding remarks

Here, we presented a general exploration of the bioactivity landscape of 52 epitrgeters, based on publicly available data. Relevant SAR information from the molecular scaffolds is discussed from a polypharmacology approach. Molecular fingerprints were computed for each target by summarizing the small molecules that have been proven active against them. This allowed further comparisons and visualization of the chemical space. More studies comparing pharmacological profiles of epitrgeters will become feasible as the size of the public databases grows.

The global structural diversity analysis of the 52 data sets tested against epigenetic targets revealed that targets that might not seem structurally similar could in fact be pharmacologically similar. This has profound implications in predicting and designing polypharmacological compounds. By contrast, targets that might be structurally similar might not resemble each other pharmacologically, providing challenges to proteochemometric approaches.

As part of this survey, epitrgeter DFPs were calculated, which could be used for similarity-based virtual screening and multiple-target virtual screening, approaches that appear feasible given their favorable recovery rates. The epigenomics compound database generated in this work can be used as a starting point for further SMART analyses to 'get SMART in epigenetics'. The epitrgeter database used in this survey is available in SDF format in the Supplemental information online.

Finally, the insights from this study could be applied to guide drug discovery approaches for the design of more selective compounds (e.g., specific inhibitors of DNMT1, 3A, or 3B), or even the rational design of compounds with relevant polypharmacological properties in epigenetics. Moreover, this is the first attempt to cluster epigenetic targets from a pharmacological point of view, which is likely to evolve as more compounds are tested. Major next steps in this approach are the in-depth study of the physicochemical properties, SAR, and degrees of polypharmacology in individual epigenetic targets, as well as the development of specific multiple epigenetic target prediction tools.

### Acknowledgments

The authors thank OPEN Eye Scientific Software for the academic license. They are also grateful to Gisbert Schneider for providing the list of scaffolds published in Ref. [43]. J.J.N. is thankful to CONACyT for the granted scholarship number 622969. J.L.M.F. is thankful to PAPIIT project IA204016. Insightful discussions with Oscar Palomino-Hernández during the integration of the database are highly appreciated.

### Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at <https://doi.org/10.1016/j.drudis.2017.10.006>.

### References

- 1 Henikoff, S. and Gready, J.M. (2016) Epigenetics, cellular memory and gene regulation. *Curr. Biol.* 26, R644–R648
- 2 Allis, C.D. and Jenuwein, T. (2016) The molecular hallmarks of epigenetic control. *Nat. Rev. Genet.* 17, 487–500
- 3 Waldmann, T. and Schneider, R. (2013) Targeting histone modifications—epigenetics in cancer. *Curr. Opin. Cell Biol.* 25, 184–189
- 4 Suvà, M.L. *et al.* (2013) Epigenetic reprogramming in cancer. *Science* 339, 1567–1570

- 5 Li, L.-C. (2014) Chromatin remodeling by the small RNA machinery in mammalian cells. *Epigenetics* 9, 45–52
- 6 Ruthenburg, A.J. *et al.* (2007) Methylation of lysine 4 on histone H3: intricacy of writing and reading a single epigenetic mark. *Mol. Cell* 25, 15–30
- 7 Musselman, C.A. *et al.* (2012) Perceiving the epigenetic landscape through histone readers. *Nat. Struct. Mol. Biol.* 19, 1218–1227
- 8 Bowman, G.D. and Poirier, M.G. (2015) Post-translational modifications of histones that influence nucleosome dynamics. *Chem. Rev.* 115, 2274–2295
- 9 Jenuwein, T. and Allis, C.D. (2001) Translating the histone code. *Science* 293, 1074–1080
- 10 Law, J.A. and Jacobsen, S.E. (2010) Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nat. Rev. Genet.* 11, 204–220
- 11 Lopez-Serra, L. and Esteller, M. (2008) Proteins that bind methylated DNA and human cancer: reading the wrong words. *Br. J. Cancer* 98, 1881–1885
- 12 Rasmussen, K.D. and Helin, K. (2016) Role of TET enzymes in DNA methylation, development, and cancer. *Genes Dev.* 30, 733–750
- 13 Wayne, T.F. (2015) Epigenetics in the development, modification, and prevention of cardiovascular disease. *Mol. Biol. Rep.* 42, 765–776
- 14 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. Discovery and Development of Small Molecule Epigenetic Drugs and Probes* (Medina-Franco, J.L., ed.), pp. 1–20, Academic Press
- 15 Jones, P.A. *et al.* (2016) Targeting the cancer epigenome for therapy. *Nat. Rev. Genet.* 17, 630–641
- 16 Naveja, J.J. *et al.* (2016) Drug repurposing for epigenetic targets guided by computational methods. In *Epi-Informatics. Discovery and Development of Small Molecule Epigenetic Drugs and Probes* (Medina-Franco, J.L., ed.), pp. 327–357, Academic Press
- 17 Csoka, A.B. and Szyf, M. (2009) Epigenetic side-effects of common pharmaceuticals: a potential new field in medicine and pharmacology. *Med. Hypotheses* 73, 770–780
- 18 Hunter, P. (2015) The second coming of epigenetic drugs: a more strategic and broader research framework could boost the development of new drugs to modify epigenetic factors and gene expression. *EMBO Rep.* 16, 276–279
- 19 Raynal, N.J.-M. *et al.* (2017) Repositioning FDA-approved drugs in combination with epigenetic drugs to reprogram colon cancer epigenome. *Mol. Cancer Ther.* 16, 397–407
- 20 Nebbioso, A. *et al.* (2012) Trials with ‘epigenetic’ drugs: an update. *Mol. Oncol.* 6, 657–682
- 21 Heerboth, S. *et al.* (2014) Use of epigenetic drugs in disease: an overview. *Genet. Epigenet.* 6, 9–19
- 22 Meng, S. *et al.* (2014) BET inhibitor JQ1 blocks inflammation and bone destruction. *J. Dent. Res.* 93, 657–662
- 23 Anand, P. *et al.* (2013) BET bromodomains mediate transcriptional pause release in heart failure. *Cell* 154, 569–582
- 24 Banerjee, C. *et al.* (2012) BET bromodomain inhibition as a novel strategy for reactivation of HIV-1. *J. Leukoc. Biol.* 92, 1147–1154
- 25 Kilgore, M. *et al.* (2010) Inhibitors of class 1 histone deacetylases reverse contextual memory deficits in a mouse model of Alzheimer’s disease. *Neuropsychopharmacology* 35, 870–880
- 26 Arguelles, A.O. *et al.* (2016) Are epigenetic drugs for diabetes and obesity at our door step? *Drug Discov Today* 21, 499–509
- 27 Dekker, F.J. *et al.* (2014) Small molecule inhibitors of histone acetyltransferases and deacetylases are potential drugs for inflammatory diseases. *Drug Discov. Today* 19, 654–660
- 28 Prieto-Martínez, F.D. *et al.* (2016) A chemical space odyssey of inhibitors of histone deacetylases and bromodomains. *RSC Adv.* 6, 56225–56239
- 29 Gortari, E.F. and Medina-Franco, J.L. (2015) Epigenetic relevant chemical space: a chemoinformatic characterization of inhibitors of DNA methyltransferases. *RSC Adv.* 5, 87465–87476
- 30 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
- 31 Sato, T. *et al.* (2017) Transcriptional selectivity of epigenetic therapy in cancer. *Cancer Res.* 77, 470–481
- 32 Benedetti, R. *et al.* (2015) Epigenetic-based therapy: from single- to multi-target approaches. *Int. J. Biochem. Cell. Biol.* 69, 121–131
- 33 de Lera, A.R. and Ganesan, A. (2016) Epigenetic polypharmacology: from combination therapy to multitargeted drugs. *Clin. Epigenetics* 8, 105
- 34 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
- 35 García-Sánchez, M.O. *et al.* (2017) Quantitative structure-epigenetic activity Relationships. In *Advances in QSAR Modeling*, (24) (Roy, K., ed.), pp. 303–338, Springer International Publishing
- 36 Cabaye, A. *et al.* (2015) Structural diversity of the epigenetics pocketome. *Proteins* 83, 1316–1326
- 37 Maggiora, G. and Gokhale, V. (2017) A simple mathematical approach to the analysis of polypharmacology and polyspecificity data. *F1000Res* 6, 788
- 38 Saldívar-Medina, M. *et al.* (2016) Consensus diversity plots: a global diversity analysis of chemical libraries. *J. Cheminform.* 8, 63
- 39 Fernández-de Gortari, E. *et al.* (2017) Database fingerprint (DFP): an approach to represent molecular databases. *J. Cheminform.* 9, 9
- 40 Méndez-Lucio, O. and Medina-Franco, J.L. (2017) The many roles of molecular complexity in drug discovery. *Drug Discov. Today* 22, 120–126
- 41 Lipinski, C.A. (2004) Lead- and drug-like compounds: the rule-of-five revolution. *Drug Discov. Today Technol.* 1, 337–341
- 42 Hope, A.C. (1968) A simplified Monte Carlo significance test procedure. *J. R. Statist. Soc. B* 30, 582–598
- 43 Schneider, P. and Schneider, G. (2017) Privileged structures revisited. *Angew. Chem. Int. Ed. Engl.* 56, 7971–7974
- 44 Naveja, J.J. and Medina-Franco, J.L. (2017) ChemMaps: towards an approach for visualizing the chemical space based on adaptive satellite compounds. *F1000Res* 6, 1134
- 45 Maggiora, G.M. and Bajorath, J. (2014) Chemical space networks: a powerful new paradigm for the description of chemical space. *J. Comput. Aided Mol. Des.* 28, 795–802
- 46 Osolodkin, D.I. *et al.* (2015) Progress in visual representations of chemical space. *Expert Opin. Drug Discov.* 10, 959–973
- 47 Reymond, J.-L. (2015) The chemical space project. *ACC Chem. Res.* 48, 722–730
- 48 López-Vallejo, F. *et al.* (2012) Expanding the medicinally relevant chemical space with compound libraries. *Drug Discov. Today* 17, 718–726
- 49 Lovering, F. *et al.* (2009) Escape from flatland: increasing saturation as an approach to improving clinical success. *J. Med. Chem.* 52, 6752–6756



# Computational Methods for Epigenetic Drug Discovery: A Focus on Activity Landscape Modeling

J. Jesús Naveja<sup>\*,†</sup>, C. Iluhí Oviedo-Osornio<sup>\*</sup>, José L. Medina-Franco<sup>\*,1</sup>

<sup>\*</sup>Facultad de Química, Universidad Nacional Autónoma de México, Mexico City, Mexico

<sup>†</sup>PECEM, Facultad de Medicina, Universidad Nacional Autónoma de México, Mexico City, Mexico

<sup>1</sup>Corresponding author: e-mail addresses: medinajl@unam.mx; jose.medina.franco@gmail.com

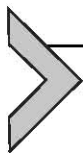
## Contents

1. Introduction	66
2. Epigenetic Targets	67
3. Activity Landscape Modeling	68
3.1 SAS Maps	69
3.2 Structure–Activity Landscape Index	71
3.3 SAS Maps of Epigenetic Targets	72
3.4 Activity Cliffs Generators	73
3.5 Epigenetic Targets With Continuous SAR	79
3.6 Epigenetic Targets With Scaffold Hops	79
4. Conclusions and Perspectives	80
Acknowledgments	80
References	81

## Abstract

Epigenetic drug discovery is an emerging strategy against several chronic and complex diseases. The increased interest in epigenetics has boosted the development and maintenance of large information on structure–epigenetic activity relationships for several epigenetic targets. In turn, such large databases—many in the public domain—are a rich source of information to explore their structure–activity relationships (SARs). Herein, we conducted a large-scale analysis of the SAR of epigenetic targets using the concept of activity landscape modeling. A comprehensive quantitative analysis and a novel visual representation of the *epigenetic activity landscape* enabled the rapid identification of regions of targets with continuous and discontinuous SAR. This information led to the identification of epigenetic targets for which it is anticipated an easier or a more difficult drug-discovery program using conventional hit-to-lead approaches. The insights of this work also enabled the identification of specific structural changes associated with a

large shift in biological activity. To the best of our knowledge, this work represents the largest comprehensive SAR analysis of several epigenetic targets and contributes to the better understanding of the *epigenetic activity landscape*.



## 1. INTRODUCTION

Epigenetics, despite of all its intrinsic complexity, is quite promising in drug discovery, given the expected potential of modifying or even reversing complex gene expression patterns associated to chronic diseases (Dueñas-González, Naveja, & Medina-Franco, 2016). Also, the possibility of inhibiting epigenetic processes at writing, reading, and erasing times adds to the diversity of potential therapies (Chung, 2015). Briefly, epigenetic writers add marks to either DNA or histones, which are in turn transduced by readers, therefore leading to a cell response. In antagonism to writers, epigenetic erasers remove marks and thereby inhibit the subsequent signal (Allis & Jenuwein, 2016). Examples of writers are DNA methyltransferases (DNMTs) and histone acetyltransferases (HATs); examples of readers are bromodomains (BRDs), while histone deacetylases (HDACs) account for the most studied epigenetic erasers.

The increasing awareness of the role of epigenetic targets for the treatment of several diseases has boosted the development of compounds that can act as potential inhibitors. Indeed, the number of compounds tested and the number of targets under study have been increasing (Lundstrom, 2017). This far, the chemical space of small molecules targeting major epigenetic targets has been explored. Examples are inhibitors of HDACs and BRDs (Prieto-Martínez, Gortari, Méndez-Lucio, & Medina-Franco, 2016). These studies have contributed to the charting of the so-called Epigenetic Relevant Chemical Space (ERCS) (Gortari & Medina-Franco, 2015). More recently, the chemical space of 52 epigenetic targets has been characterized (Naveja & Medina-Franco, 2018). In that study, the recently developed concept of *database fingerprints* was used to map targets according to their pharmacological similarity (Fernández-de Gortari, García-Jacas, Martínez-Mayorga, & Medina-Franco, 2017). Further analysis of this type might confirm the feasibility of rational design of epi-polypharmacological compounds.

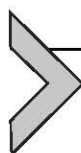
A major next step to develop effective compounds for epigenetic targets or epi-drugs is the analysis of the structure–activity relationships (SARs) associated with active molecules. One computational approach to systematically characterize the SAR of compound data sets is activity landscape



modeling. Activity landscape modeling can be conceptualized as the association between the chemical space and biological activity. This far, the activity landscape of few, although major, epigenetic targets has been analyzed. For instance, Naveja et al. analyzed the activity landscape of DNMTs (Naveja & Medina-Franco, 2015a, 2015b). Saldívar-González et al. explored the landscape of HDACs (Saldívar-González, Naveja, Palomino-Hernández, & Medina-Franco, 2017), and García-Sánchez et al. the landscape of BRD inhibitors (García-Sánchez, Cruz-Montegudo, & Medina-Franco, 2017). However, there is a need to analyze the activity landscape of many more relevant epigenetic targets associated with the ERCS.

The main goal of this study is to characterize the epigenetic activity landscape associated with ERCS currently known. The specific goals are: (a) to identify the epigenetic targets with more continuous and discontinuous SAR; (b) to quantify the proportion of activity cliffs for each target while identifying the most frequent activity cliffs, i.e., activity cliff generators (ACGs); and (c) to provide a structure-based rationale of the ACGs and epigenetic targets with more discontinuous SAR. To achieve these goals Structure–Activity Similarity (SAS) maps and Structure–Activity Landscape Index (SALI) values were employed. These are well-known computational approaches for activity landscape modeling (Guha & Van Drie, 2008; Méndez-Lucio, Pérez-Villanueva, Castillo, & Medina-Franco, 2012).

This chapter is organized into four major sections. After this introduction, Section 2 discusses the major epigenetic targets covered in this chapter. Section 3 presents the results of the activity landscape modeling of the epigenetic targets. This section is further divided into several subsections each discussing the major components of the epigenetic activity landscape. Section 4 presents conclusions and perspectives.



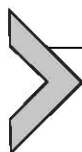
## 2. EPIGENETIC TARGETS

The activity landscape modeling presented in this work builds upon a recently published preliminary study on the chemical space of epigenetic targets (Naveja & Medina-Franco, 2018). In that work, the chemical libraries of major families of epigenetic targets, i.e., HDACs, DNMTs, HATs, BRDs, KMeR, PRMTs, HKMs, KDMs, and others (see summary in Table 1), revealed interesting similarity associations. For instance, inhibitors of HDACs tended to form well-defined clusters, while other families of targets, such as DNMTs, had little resemblance among each other (Naveja & Medina-Franco, 2018). Regarding SAR analysis, enrichment factors for

**Table 1** Main Epigenetic Targets Considered in This Work

Family	Function	Targets
BRD	Histone acetylation reading	BAZ2B, BRD2, BRD3, BRD4, BRD9, BRPF1
DNMT	DNA methylation writing	DNMT1, DNMT3A, DNMT3B
HAT	Histone acetylation writing	CREBBP, EP300, KAT2A, KAT2B, NCOA1, NCOA3
HDAC	Histone acetylation erasing	HDAC: 1–11
HKM	Histone lysine methylation writing	DOT1L, EHMT1, EHMT2, KMT5A, SMYD2
KDM	Histone lysine methylation erasing	KDM: 1A, 2A, 3A, 4A, 4C, 4E, 5A, 5C
KMeR	Histone lysine methylation reading	L3MBTL: 1, 3, 4 CBX7, TP53BP1, WDR5
PRMT	Histone arginine methylation writing	PRMT: 1, 6, 8 CARM1
Others	Miscellaneous	MAP3K7 (kinase), MGEA5 (histone O-N-acetylglucosamine transferase), SMARCA2 (chromatin remodeler)

Bemis–Murcko scaffolds were calculated, finding distinct molecular motifs susceptible of further development. However, an exhaustive SAR analysis of all these epigenetic targets has not been performed until now. An overview of the epigenetic targets included in this work, aggregated by family, is presented in [Table 1](#). Of note, epigenetic writers, readers, and erasers are represented in this list. For detailed reviews on epigenetic targets and their functions, please see [Allis and Jenuwein \(2016\)](#), [Jenuwein and Allis \(2001\)](#), [Medvedeva et al. \(2015\)](#), and citations therein.



### 3. ACTIVITY LANDSCAPE MODELING

Activity landscape modeling is a useful strategy in medicinal chemistry and drug discovery to explore and describe the SAR of chemical data sets ([García-Sánchez et al., 2017](#)). This computational strategy is valuable to guide lead-optimization efforts and to develop *in silico* models, such as



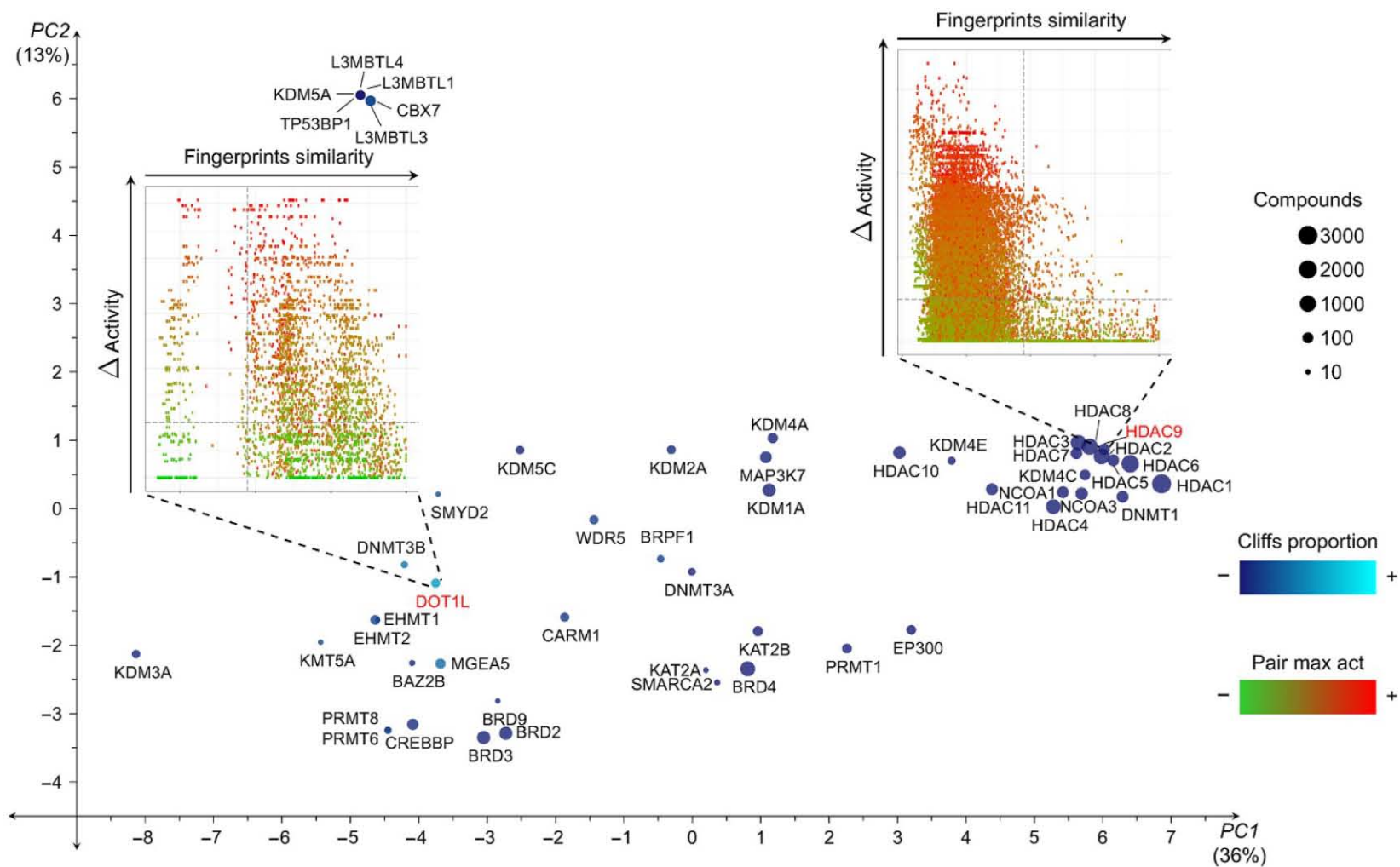
QSAR and similarity-based virtual screening (Medina-Franco, 2012; Peltason & Bajorath, 2007).

Activity landscape can be conceptualized as the association between the chemical space and the activity data (Wassermann, Wawer, & Bajorath, 2010). With this approach we can directly assess the similarity principle, i.e., whether compounds in the data set that are more similar structurally are also alike in activity. A continuous SAR is one where small changes in activity due to small changes in molecular structure can be found, while a discontinuous SAR has many cases where small changes in molecular structure can lead to a significant change in the activity (Pérez-Villanueva et al., 2010). Fig. 1 depicts the chemical space of epigenetic targets as it resulted in Naveja and Medina-Franco (2018). Data points are colored according to the proportion of activity cliffs in the data set using a continuous color scale from low (dark blue) to high (light blue) proportion of cliffs. Two examples of SAS maps (activity landscape depictions—see Sections 3.1 and 3.3) are illustrated: (a) DOT1L, which has a quite discontinuous SAR, and (b) HDAC9, in contrast, with a rather continuous SAR. In SAS maps, each point represents a paired comparison of structure and activity similarity, where points in the upper right quadrant are activity cliffs (Medina-Franco, 2012). Supplementary Fig. S1 in the online version at <https://doi.org/10.1016/bs.apcsb.2018.01.001> shows a 3D representation of the SAR for these targets, along with relevant average physicochemical and complexity properties. Interestingly, regions in the chemical spaces enriched with discontinuous SAR targets seem to exist, and these same regions tend to have higher complexity ( $n$  chiral atoms and FCSP3) (Fig. 1 and Supplementary Fig. S1 in the online version at <https://doi.org/10.1016/bs.apcsb.2018.01.001>).

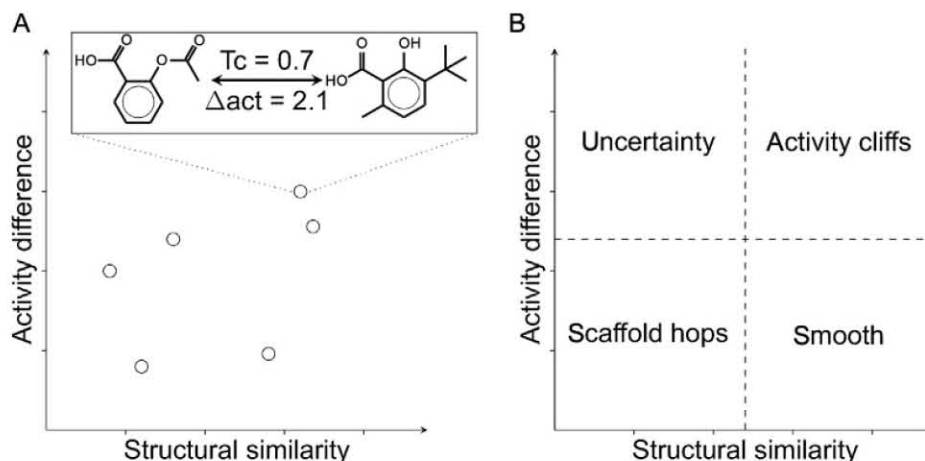
During the last few years, efforts have been made to develop quantitative and visual methods to model activity landscapes. Examples of these approaches are SAS maps, Dual-Activity Difference (DAD) maps, SALI, and Structure-Activity Relationship Index (SARI), to name a few (Méndez-Lucio, 2016). These approaches have been applied to a large number of compound data sets of relevance in medicinal chemistry (e.g., Medina-Franco, 2012; Méndez-Lucio et al., 2012; Naveja & Medina-Franco, 2015a).

### 3.1 SAS Maps

Shanmugasundaram and Maggiora introduced the SAS maps, a 2D activity landscape representation which compares structural similarity (ECFP, MACCS) and activity similarity (for example,  $pIC_{50}$  or  $pK_i$ ) on the basis



**Fig. 1** Chemical space of epigenetic targets colored by the proportion of cliffs in each dataset. Two representative SAS maps (for DOT1L and HDAC9) are shown.



**Fig. 2** (A) Schematic SAS map. Note that each point represents a pair of compounds. (B) Four major quadrants in a SAS map.

of systematic pairwise compound comparisons (Shanmugasundaram & Maggiora, 2001). Each point in a SAS map represents a pair of compounds and is colored according to the most active compound of the pair. The resultant plot (schematically illustrated in Fig. 2) can be roughly divided into four quadrants with thresholds defined a priori: (a) smooth (high structural similarity and low activity difference), (b) activity cliffs (high structural similarity but high activity difference), (c) scaffold hops (low structural similarity but low activity difference), and (d) uncertainty (low structural similarity and high activity difference) (Bajorath et al., 2009; Guha, 2012; Medina-Franco, 2012).

### 3.2 Structure–Activity Landscape Index

SALI was designed to identify activity cliffs and compounds that represent key inflection points on activity landscapes. This metric gives a score to a pair of compounds based on the comparison of the structural similarity and the difference between their potency. SALI is calculated as follows:

$$SALI_{i,j} = \frac{|A_j - A_i|}{1 - \text{sim}(i,j)}$$

where  $A_i$  and  $A_j$  are the potency values of the  $i$ th and  $j$ th molecule, respectively, and  $\text{sim}(i,j)$  is the similarity of the two molecules (Bajorath et al., 2009; Guha, 2012). Higher values of SALI correspond to more pronounced activity cliffs (Bajorath et al., 2009; Guha, 2012).

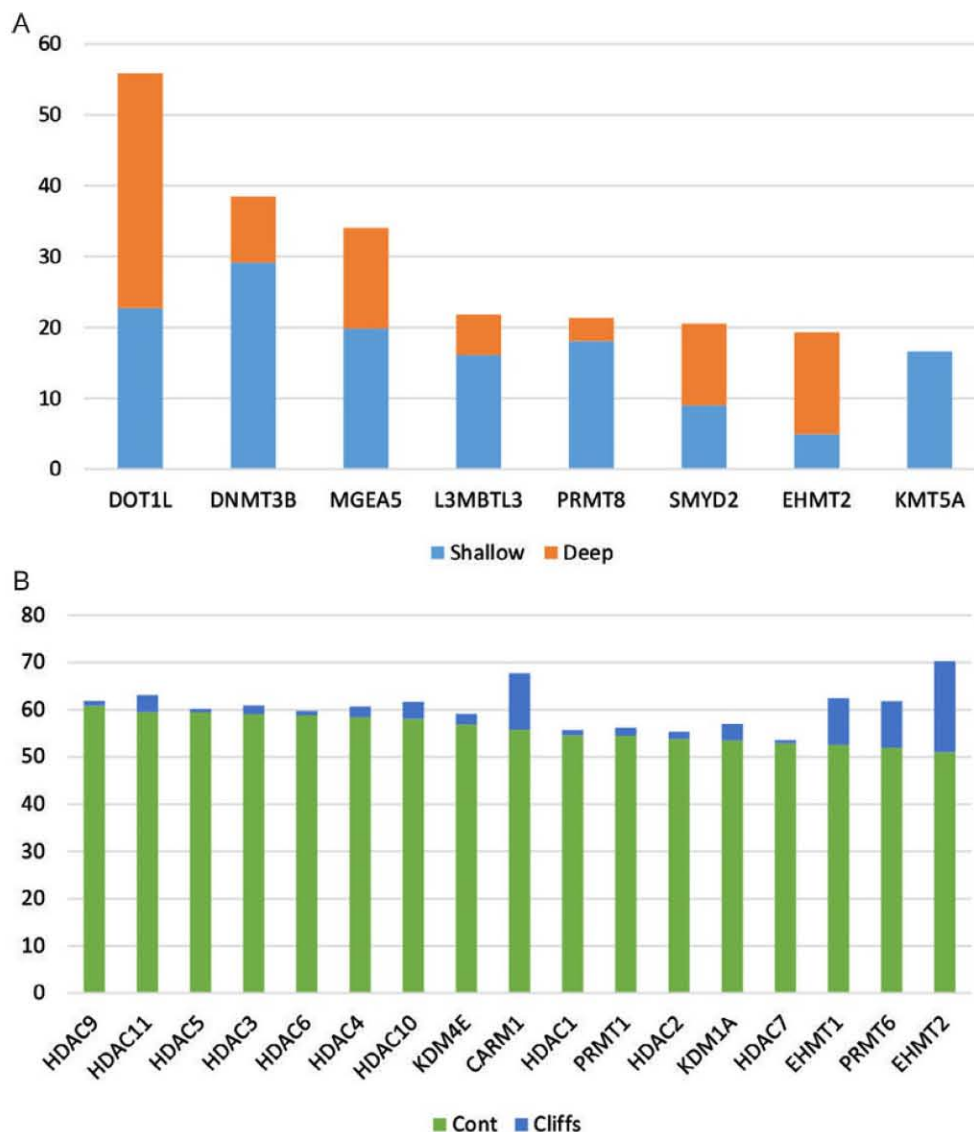
### 3.3 SAS Maps of Epigenetic Targets

As discussed in Section 3.1, exploring the SAR of compound data sets provides key information to determine the approach to further developing the compounds. Thus, if the SAR is continuous (e.g., similar compounds have similar activity), it is possible to implement methods based on the similarity principle. On the other hand, if the data set has a discontinuous SAR, insights from the activity landscape might provide specific information on the pharmacophore features that will be key for further compound development.

Herein, SAS maps for all the 52 epigenetic targets were generated and analyzed (Supplementary Fig. S2 in the online version at <https://doi.org/10.1016/bs.apcsb.2018.01.001>). It was found that the epigenetic targets DOT1L, DNMT3B, and MGEA5 have a significant percentage of cliffs (overall more than 30%) with respect to other epigenetic targets studied in this analysis (Fig. 3). These results suggest that, in general, small changes on the structure of the inhibitors of DOT1L, DNMT3B, and MGEA5 could generate large changes on their biological activity.

DOT1L is the epigenetic target that has the largest percentage of activity cliffs. In addition, similar to the inhibitors of EHMT2, inhibitors of DOT1L have a large percentage of “deep cliffs,” i.e., a small modification in the structure will modify the activity in more than two logarithmic units. In comparison, DNMT3B and PRMT8 are the epigenetic targets with the largest percentage of shallow cliffs, i.e., activity differences between one and two logarithmic units. Such targets with many activity cliffs and therefore a discontinuous SAR are less suited for applying predictive methods that rely upon the similarity principle.

In contrast, other epigenetic targets have an overall continuous SAR. Remarkable examples are most of the HDACs. As shown in Fig. 3B, the majority of the pairs are in the smooth SAR region. This is no surprise, since HDACs are among the most well-understood epigenetic targets. A polarized example is provided by EHMT, which has both ~51% of compounds in the smooth SAR region and ~19% of activity cliffs. Interestingly, SMARCA2 and NCOA1 had a large proportion of scaffold hops, perhaps pointing toward multiple binding sites on the enzymes. Table 2 presents the targets with the highest percentages of points in each of the SAS maps quadrants, for each of the studied epigenetic families. Of note, although HDAC family is the one with the largest number of evaluated compounds, HKM family (and particularly DOT1L and EHMT2) has a larger proportion of



**Fig. 3** Epigenetic targets with the largest proportion of pairs of compounds: (A) in the activity cliff region (shallow and deep cliffs are differentiated) and (B) in the continuous region of the SAS maps.

activity cliffs. Mean SALI values are not quite linearly correlated with the proportion of activity cliffs of each target ( $r = 0.18$ ). However, WDR5 has the highest mean SALI, and it has a discontinuous SAR (Supplementary Figs. S1 and S2 in the online version at <https://doi.org/10.1016/bs.apcsb.2018.01.001>).

### 3.4 Activity Cliffs Generators

An ACG is defined as “a molecule with high probability to form activity cliffs with structurally similar molecules tested in the same biological assay”

**Table 2** Percentage of Data Points in Different Regions of the SAS Maps and Mean SALI

Target Family	<i>n</i>	Activity Cliffs (%)			Scaffold Hops (%)	Continuous (%)	SALI Mean
		Total	Deep	Shallow			
BRD	200,028	16.21	2.37	13.83	43.98	48.79	16.72
	BRD4	BRPF1	BRPF1	BRPF1	BAZ2B	BRD4	BRPF1
DNMT	29,646	38.72	9.36	29.10	67.03	43.57	4.70
	DNMT1	DNMT3B	DNMT3B	DNMT3B	DNMT1	DNMT3A	DNMT3B
HAT	158,766	7.94	0.08	7.80	84.49	39.44	1.75
	NCOA1NCOA3	CREBBP	EP300	CREBBP	NCOA1	NCOA3	CREBBP
HDAC	5,227,761	3.60	1.80	1.79	46.44	60.76	2.57
	HDAC1	HDAC11	HDAC11	HDAC11	HDAC8	HDAC9	HDAC11
HKM	5995	56.11	33.06	22.78	22.13	52.57	17.38
	EHMT2	DOT1L	DOT1L	DOT1L	EHMT1	EHMT1	KMT5A
KDM	102,378	7.54	1.10	6.23	66.17	56.80	6.45
	KDM1A	KDM2A	KDM1A	KDM2A	KDM4A	KDM4E	KDM4A
KMeR	7750	21.83	5.70	16.13	38.08	34.67	25.60
	L3MBTL1	L3MBTL3	L3MBTL3	L3MBTL3	L3MBTL1	WDR5	WDR5
PRMT	9870	21.74	4.19	18.12	39.52	55.71	5.70
	PRMT1	PRMT8	CARM1	PRMT8	PRMT1	CARM1	CARM1
Others	24,090	34.09	14.16	19.86	90.25	37.79	15.12
	SMARCA2	MGEA5	MGEA5	MGEA5	SMARCA2	MAP3K7	MGEA5

(Pérez-Villanueva, Méndez-Lucio, Soria-Arteche, & Medina-Franco, 2015). This kind of molecules provides relevant insights on the pharmacophoric regions. Below, examples of ACGs found in the database are described in detail and illustrated using the chemical neighborhood graphs devised by Namasivayam, Iyer, and Bajorath (2012).

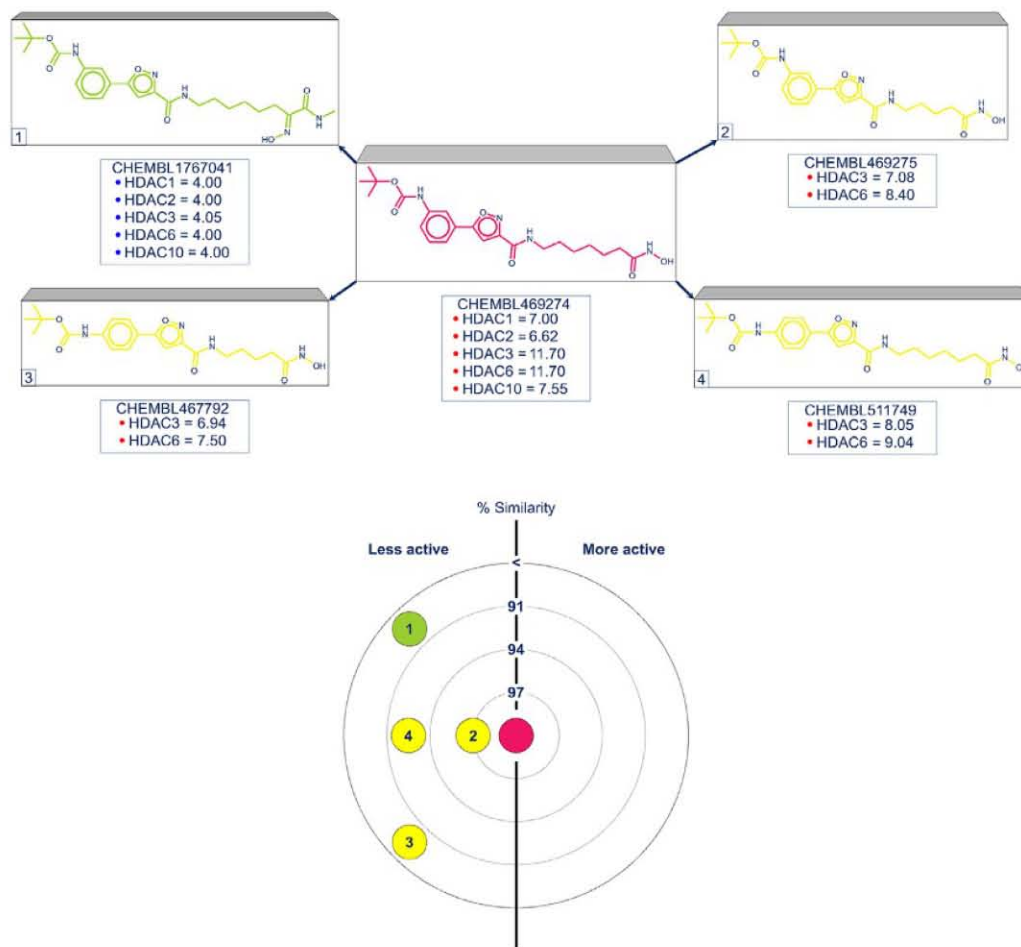
### 3.4.1 Histone Deacetylases

HDACs are enzymes that regulate the gene transcription by catalyzing the deacetylation of the  $\epsilon$ -amino group of lysine side chains of histone and non-histone proteins. Deacetylation promotes a stronger interaction between positively charged histones and negatively charged DNA, leading to a more condensed chromatin structure and gene transcription silencing (Ragno, 2016; Richon, 2006).

There are 18 human HDACs subdivided into four classes (Xu, Parmigiani, & Marks, 2007). Classes I (HDAC 1, 2, 3, and 8), IIa (HDAC 4, 5, 7, and 9), IIb (HDAC 6 and 10), and IV (HDAC 11) are metal dependent and share the catalytic system (Botta et al., 2011; Negmeldin, Padige, Bieliauskas, & Pflum, 2017). Some types of cancer have an altered gene expression of these enzymes. For example, there is an overexpression of HDAC1 in prostate, gastric, colon, and breast carcinomas; in turn, HDAC2 is overexpressed in colorectal, gastric, and cervical cancer (Kim & Bae, 2011; Kral et al., 2014). On the other hand, their inhibition causes histone hyperacetylation associated with cell cycle arrest or apoptosis in tumor cells (Xu et al., 2007). In addition, the acetylation of p53 by HATs increments its binding to DNA, leading to the expression of p53-regulated genes (Botta et al., 2011; Kral et al., 2014; Richon, 2006).

Some HDAC inhibitors have already been approved by the FDA as anti-cancer drugs: e.g., vorinostat, belinostat, panobinostat, and romidepsin. The former three contain a hydroxamic acid group and the latter is a cyclic peptide (Kim & Bae, 2011; Negmeldin et al., 2017). Moreover, HDAC inhibitors might be beneficial in different types of neurodegenerative and cardiovascular diseases, as well as in inflammatory disorders (Kim & Bae, 2011). Hydroxamic acid derivative inhibitors have a similar scaffold which can be divided into three parts: (1) a  $Zn^{2+}$  binding group (ZBG) that coordinates to the catalytic metal atom in the “tube-like” active site; (2) a linker, which helps to find the correct position and fits into the hydrophobic “tube”; and (3) a “cap” region, which interacts with the rim of the active pocket (Botta et al., 2011; Hou et al., 2012).





**Fig. 4** Example of an activity cliff generator (CHEMBL469274) that is active against a variety of HDACs.

CHEMBL469274 is an ACG that is active against a variety of HDACs, but it is prone to forming activity cliffs (Fig. 4). The most dramatic drop in activity is shown by CHEMBL1767041, which lacks an efficient ZBG (Ragno, 2016). On the other hand, although the ZBG is conserved in CHEMBL469275, CHEMBL467792, and CHEMBL511749, the large decrease in the activity values seems to be a result of modifications in the linker length (CHEMBL469275 and CHEMBL467792) or regioisomerism (CHEMBL469275 and CHEMBL469274), which might prevent an adequate fitting in the binding pocket of the enzyme.

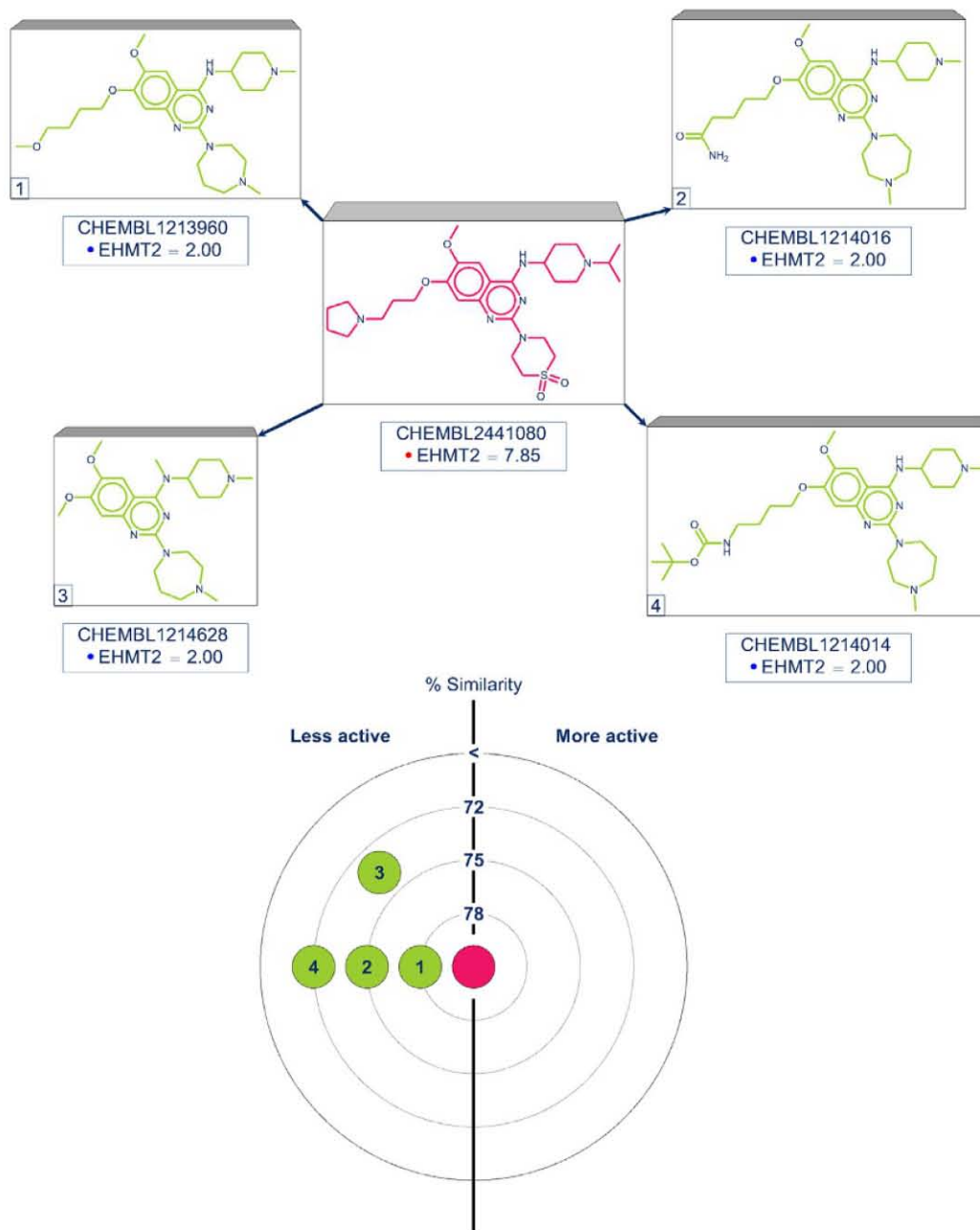
### 3.4.2 Histone Methyltransferases EHMT2 and DOT1L

Histone methyltransferases (HMTs) are currently under research, given the dysregulation of this metabolic pathway in cancer cells (Curry et al., 2015; Lu et al., 2013). Also, DOT1L has been described as a potentially selective

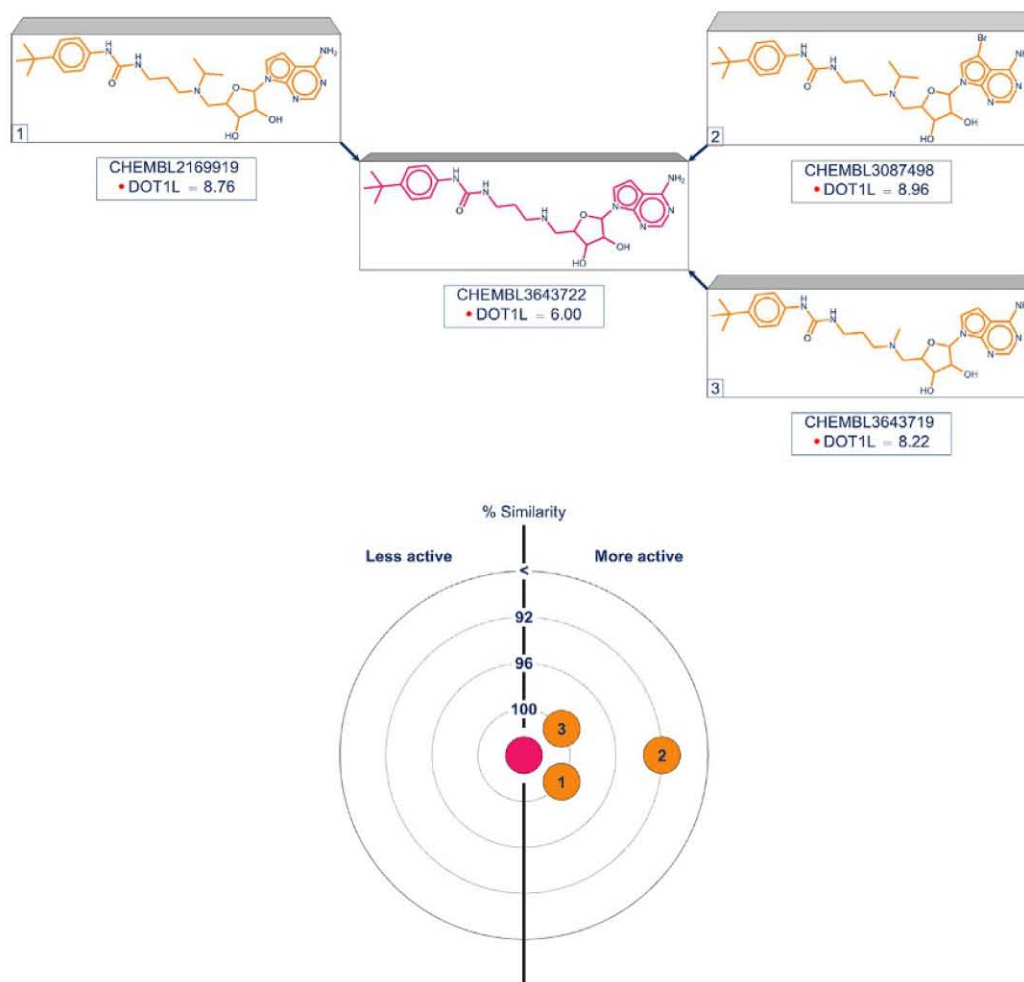


therapy against some types of leukemia (Daigle et al., 2011; Sarkaria, Christopher, Klco, & Ley, 2014).

Some interesting ACGs were identified as inhibitors of HMTs and are presented in Figs. 5 and 6. The most pronounced activity cliffs from an ACG are those in Fig. 5. CHEMBL2441080 is a compound active against EHMT2, a HMT that has been found to be linked to carcinogenic processes (Lu et al., 2013). Fig. 5 shows four structurally similar compounds that



**Fig. 5** Example of an activity cliff generator (CHEMBL2441080) that is active against EHMT2.



**Fig. 6** Example of an activity cliff generator (CHEMBL3643722) that is active against DOT1L, although less active than its pairs.

have lost activity. They all share the elimination of a sulfone functional group, addition of a carbon to the ring where the sulfone is present in the ACG, a substitution of an isopropyl in a lateral chain by a methyl, and they differ in the substitutions done on the longest lateral chain. Despite the high chemical similarity identified through molecular fingerprints, the relatively high number of modifications make difficult to categorically telling which is the culprit of the loss of activity. A study on the SAR of 2,4-diamino-7-aminoalkoxy-quinazolines, such as the activity cliffs surrounding this ACG, showed that, in fact, there are compounds with this scaffold that are very active (Liu et al., 2010). Therefore, the long lateral chain mostly plays a predominant role on the differences observed in Fig. 5. Indeed, as per the crystallographic structure presented therein, the lack of the nitrogen four positions away from the oxygen on this same chain may disrupt electrostatic and cation- $\pi$  interactions with Leu1086 and Tyr1154, respectively (Liu et al., 2010).

A final illustrative example is the case of the ACG CHEMBL3643722 (Fig. 6). This is a compound with moderate activity against DOT1L, which is another HMT. In this case, turning the secondary amine into a tertiary amine enhances activity in more than two logarithmic units. Also, substituting with an isopropyl seems to provide higher activity than with a methyl group.

A SAR study on adenosine analogs as inhibitors of DOT1L found that CHEMBL2169919 (also known as EPZ004777) amine group can be replaced without significant loss of activity by —S—, similar to SAM, the enzyme's cofactor required for methylation (Anglin et al., 2012). However, the ACG identified in this study further suggests that the substitution of the amine is relevant in the activity.

### 3.5 Epigenetic Targets With Continuous SAR

According to Table 2, HDACs present the most continuous SAR in the database. The molecules in this region have high structural similarity and low potency differences (Fig. 2B), i.e., they follow the similarity principle. This property can be used to optimize the scaffold of active compounds in order to generate more potent and selective analogs. Overall, HDAC inhibitors are more predominant in this area (HDAC9 has the maximum, ~60% of pairs in the continuous region). Besides, they also have low presence of activity cliffs (HDAC11 has the maximum, ~4%). These facts, in principle, allow the efficient development of in silico predictive models, such as QSAR. Moreover, some HDAC inhibitors have a complex molecular structure, favoring the generation of a large number of analogs from the same synthetic route. However, the scaffold hop region of HDAC inhibitors is of considerable size (up to 46% in HDAC8), which could suggest that either these targets have multiple binding sites, or there are multiple chemotypes able to inhibit the same binding site.

### 3.6 Epigenetic Targets With Scaffold Hops

In clear contrast to the epigenetic targets with a large proportion of activity cliffs and data points in the continuous region of the SAR, SMARCA2 and HAT are the targets with the two largest proportions of data points in the region of the scaffold hops of the landscape (Table 2). This result suggests there is a large diversity of compounds with similar activity toward SMARCA2 and HAT. If these compounds are active, there would be a significant opportunity to develop the most promising scaffolds with adequate drug-like properties. Also, having different chemical scaffolds with similar activity opens up more venues to synthesize analogs of the most tractable

scaffolds and increases the chances to generate intellectual property (that is a sensitive point in particular to the pharmaceutical industry). From the point of view of the epigenetic targets, those targets with a large proportion of activity cliffs mean that they might be more promiscuous and more adaptable (flexible to accommodate ligands with different scaffolds). In addition, the large proportion of scaffold hops for SMARCA2 and HAT led to the hypothesis that active compounds toward these epigenetic targets may act through different mechanisms of action, e.g., bind in different binding sites. Despite the fact the structure-based interpretation of the scaffolds hops (and other regions of the landscape) is addressed in a separate study, the quantification of the activity landscape (e.g., Table 2 and SAS maps in Supplementary Fig. S2 in the online version at <https://doi.org/10.1016/bs.apcsb.2018.01.001>) points to specific targets that have a distinct landscape.



#### 4. CONCLUSIONS AND PERSPECTIVES

An in-depth epigenetic activity landscape study for 52 epigenetic targets discussed in this work rapidly identified DOT1L and DNMT3B as the epigenetic targets with more discontinuous SAR, e.g., the largest proportion of activity cliffs. The study also found that HDACs are the targets with the most continuous SAR in epigenetics. The significance of this work is manifold: (a) it contributes to identify HDACs, in general, as those epigenetic targets are most suitable to perform a hit-to-lead optimization program (e.g., following the similarity principle), (b) it helped to identify or confirm small structural changes that have a large impact in the biological activity; and (c) it aided to uncover the epigenetic targets suitable to conduct traditional predictive computational approaches (such as QSAR) and those targets prone to scaffold hopping: SMARCA2 and HAT.

A major perspective of this work is to conduct structure–multiepigentic activity relationships. A second major perspective is to rationalize activity cliffs using structure-based approaches such as molecular docking.

Supplementary data to this article can be found online at <https://doi.org/10.1016/bs.apcsb.2018.01.001>.

#### ACKNOWLEDGMENTS

This work was supported by the *Programa de Apoyo a Proyectos de Investigación e Innovación Tecnológica* (PAPIIT) grant IA203718, Universidad Nacional Autónoma de México (UNAM), *Consejo Nacional de Ciencia y Tecnología* (CONACyT) grant 282785, and the *Programa de Apoyo a la Investigación y el Posgrado* (PAIP) grant 5000-9163, Facultad de Química, UNAM. J.J.N. is thankful to *Consejo Nacional de Ciencia y Tecnología* (CONACyT) for the granted scholarship number 622969.

## REFERENCES

- Allis, C. D., & Jenuwein, T. (2016). The molecular hallmarks of epigenetic control. *Nature Reviews. Genetics*, 17(8), 487–500. <https://doi.org/10.1038/nrg.2016.59>.
- Anglin, J. L., Deng, L., Yao, Y., Cai, G., Liu, Z., Jiang, H., et al. (2012). Synthesis and structure–activity relationship investigation of adenosine-containing inhibitors of histone methyltransferase DOT1L. *Journal of Medicinal Chemistry*, 55(18), 8066–8074. <https://doi.org/10.1021/jm300917h>.
- Bajorath, J., Peltason, L., Wawer, M., Guha, R., Lajiness, M. S., & Van Drie, J. H. (2009). Navigating structure–activity landscapes. *Drug Discovery Today*, 14(13–14), 698–705. <https://doi.org/10.1016/j.drudis.2009.04.003>.
- Botta, C. B., Cabri, W., Cini, E., De Cesare, L., Fattorusso, C., Giannini, G., et al. (2011). Oxime amides as a novel zinc binding group in histone deacetylase inhibitors: Synthesis, biological activity, and computational evaluation. *Journal of Medicinal Chemistry*, 54(7), 2165–2182. <https://doi.org/10.1021/jm101373a>.
- Chung, C. (2015). Epigenetic drug discovery. In G. Scapin, D. Patel, & E. Arnold (Eds.), *Multifaceted roles of crystallography in modern drug discovery* (pp. 27–40). Dordrecht: Springer Netherlands. [https://doi.org/10.1007/978-94-017-9719-1\\_3](https://doi.org/10.1007/978-94-017-9719-1_3).
- Curry, E., Green, I., Chapman-Rothe, N., Shamsaei, E., Kandil, S., Cherblanc, F. L., et al. (2015). Dual EZH2 and EHMT2 histone methyltransferase inhibition increases biological efficacy in breast cancer cells. *Clinical Epigenetics*, 7(84). <https://doi.org/10.1186/s13148-015-0118-9>.
- Daigle, S. R., Olhava, E. J., Therkelsen, C. A., Majer, C. R., Sneeringer, C. J., Song, J., et al. (2011). Selective killing of mixed lineage leukemia cells by a potent small-molecule DOT1L inhibitor. *Cancer Cell*, 20(1), 53–65. <https://doi.org/10.1016/j.ccr.2011.06.009>.
- Dueñas-González, A., Naveja, J. J., & Medina-Franco, J. L. (2016). Introduction of epigenetic targets in drug discovery and current status of epi-drugs and epi-probes. In *Epi-Informatics* (pp. 1–20). Elsevier. <https://doi.org/10.1016/B978-0-12-802808-7.00001-0>.
- Fernández-de Gortari, E., García-Jacas, C. R., Martínez-Mayorga, K., & Medina-Franco, J. L. (2017). Database fingerprint (DFP): An approach to represent molecular databases. *Journal of Cheminformatics*, 9(1, 9). <https://doi.org/10.1186/s13321-017-0195-1>.
- García-Sánchez, M. O., Cruz-Monteagudo, M., & Medina-Franco, J. L. (2017). Quantitative structure–epigenetic activity relationships. In K. Roy (Ed.), *Advances in QSAR modeling* (pp. 303–338). : Vol. 24. (pp. 303–338). Cham: Springer International Publishing. [https://doi.org/10.1007/978-3-319-56850-8\\_8](https://doi.org/10.1007/978-3-319-56850-8_8).
- Gortari, E. F., & Medina-Franco, J. L. (2015). Epigenetic relevant chemical space: A chemoinformatic characterization of inhibitors of DNA methyltransferases. *RSC Advances*, 5(106), 87465–87476. <https://doi.org/10.1039/C5RA19611F>.
- Guha, R. (2012). Exploring structure–activity data using the landscape paradigm. *Wiley Interdisciplinary Reviews. Computational Molecular Science*, 2(6). <https://doi.org/10.1002/wcms.1087>.
- Guha, R., & Van Drie, J. H. (2008). Structure–activity landscape index: Identifying and quantifying activity cliffs. *Journal of Chemical Information and Modeling*, 48(3), 646–658. <https://doi.org/10.1021/ci7004093>.
- Hou, J., Li, Z., Fang, Q., Feng, C., Zhang, H., Guo, W., et al. (2012). Discovery and extensive in vitro evaluations of NK-HDAC-1: A chiral histone deacetylase inhibitor as a promising lead. *Journal of Medicinal Chemistry*, 55(7), 3066–3075. <https://doi.org/10.1021/jm201496g>.
- Jenuwein, T., & Allis, C. D. (2001). Translating the histone code. *Science*, 293(5532), 1074–1080. <https://doi.org/10.1126/science.1063127>.
- Kim, H.-J., & Bae, S.-C. (2011). Histone deacetylase inhibitors: Molecular mechanisms of action and clinical trials as anti-cancer drugs. *American Journal of Translational Research*, 3(2), 166–179.

- Kral, A. M., Ozerova, N., Close, J., Jung, J., Chenard, M., Fleming, J., et al. (2014). Divergent kinetics differentiate the mechanism of action of two HDAC inhibitors. *Biochemistry*, 53(4), 725–734. <https://doi.org/10.1021/bi400936h>.
- Liu, F., Chen, X., Allali-Hassani, A., Quinn, A. M., Wigle, T. J., Wasney, G. A., et al. (2010). Protein lysine methyltransferase G9a inhibitors: Design, synthesis, and structure activity relationships of 2,4-diamino-7-aminoalkoxy-quinazolines. *Journal of Medicinal Chemistry*, 53(15), 5844–5857. <https://doi.org/10.1021/jm100478y>.
- Lu, Z., Tian, Y., Salwen, H. R., Chlenski, A., Godley, L. A., Raj, J. U., et al. (2013). Histone-lysine methyltransferase EHMT2 is involved in proliferation, apoptosis, cell invasion, and DNA methylation of human neuroblastoma cells. *Anti-Cancer Drugs*, 24(5), 484–493. <https://doi.org/10.1097/CAD.0b013e32835ffdbb>.
- Lundstrom, K. (2017). Epigenetics: New possibilities for drug discovery. *Future Medicinal Chemistry*, 9(5), 437–441. <https://doi.org/10.4155/fmc-2017-0015>.
- Medina-Franco, J. L. (2012). Scanning structure–activity relationships with structure–activity similarity and related maps: From consensus activity cliffs to selectivity switches. *Journal of Chemical Information and Modeling*, 52(10), 2485–2493. <https://doi.org/10.1021/ci300362x>.
- Medvedeva, Y. A., Lennartsson, A., Ehsani, R., Kulakovskiy, I. V., Vorontsov, I. E., Panahandeh, P., et al. (2015). EpiFactors: A comprehensive database of human epigenetic factors and complexes. *Database: The Journal of Biological Databases and Curation*, 2015, bav067. <https://doi.org/10.1093/database/bav067>.
- Méndez-Lucio, O. (2016). Computational structure–activity relationship studies of epigenetic target inhibitors. In *Epi-Informatics* (pp. 359–384). Elsevier. <https://doi.org/10.1016/B978-0-12-802808-7.00013-7>.
- Méndez-Lucio, O., Pérez-Villanueva, J., Castillo, R., & Medina-Franco, J. L. (2012). Identifying activity cliff generators of PPAR ligands using SAS maps. *Molecular Informatics*, 31(11–12), 837–846. <https://doi.org/10.1002/minf.201200078>.
- Namasivayam, V., Iyer, P., & Bajorath, J. (2012). Exploring SAR continuity in the vicinity of activity cliffs. *Chemical Biology & Drug Design*, 79(1), 22–29. <https://doi.org/10.1111/j.1747-0285.2011.01256.x>.
- Naveja, J. J., & Medina-Franco, J. L. (2015a). Activity landscape of DNA methyltransferase inhibitors bridges chemoinformatics with epigenetic drug discovery. *Expert Opinion on Drug Discovery*, 10(10), 1059–1070. <https://doi.org/10.1517/17460441.2015.1073257>.
- Naveja, J. J., & Medina-Franco, J. L. (2015b). Activity landscape sweeping: Insights into the mechanism of inhibition and optimization of DNMT1 inhibitors. *RSC Advances*, 5(78), 63882–63895. <https://doi.org/10.1039/C5RA12339A>.
- Naveja, J. J., & Medina-Franco, J. L. (2018). Insights from pharmacological similarity of epigenetic targets in epi-polypharmacology. *Drug Discovery Today*, 23(1), 141–150. <https://doi.org/10.1016/j.drudis.2017.10.006>.
- Negmeldin, A. T., Padige, G., Bieliauskas, A. V., & Pflum, M. K. H. (2017). Structural requirements of HDAC inhibitors: SAHA analogues modified at the C2 position display HDAC6/8 selectivity. *ACS Medicinal Chemistry Letters*, 8(3), 281–286. <https://doi.org/10.1021/acsmchemlett.6b00124>.
- Peltason, L., & Bajorath, J. (2007). SAR index: Quantifying the nature of structure–activity relationships. *Journal of Medicinal Chemistry*, 50(23), 5571–5578. <https://doi.org/10.1021/jm0705713>.
- Pérez-Villanueva, J., Méndez-Lucio, O., Soria-Arteche, O., & Medina-Franco, J. L. (2015). Activity cliffs and activity cliff generators based on chemotype–related activity landscapes. *Molecular Diversity*, 19(4), 1021–1035. <https://doi.org/10.1007/s11030-015-9609-z>.
- Pérez-Villanueva, J., Santos, R., Hernández-Campos, A., Giulianotti, M. A., Castillo, R., & Medina-Franco, J. L. (2010). Towards a systematic characterization of the antiprotozoal



- activity landscape of benzimidazole derivatives. *Bioorganic & Medicinal Chemistry*, 18(21), 7380–7391. <https://doi.org/10.1016/j.bmc.2010.09.019>.
- Prieto-Martínez, F. D., Gortari, E. F., Méndez-Lucio, O., & Medina-Franco, J. L. (2016). A chemical space odyssey of inhibitors of histone deacetylases and bromodomains. *RSC Advances*, 6(61), 56225–56239. <https://doi.org/10.1039/C6RA07224K>.
- Ragno, R. (2016). Structure-based modeling of histone deacetylases inhibitors. In *Epi-Informatics* (pp. 155–212): Elsevier. <https://doi.org/10.1016/B978-0-12-802808-7.00006-X>.
- Richon, V. M. (2006). Cancer biology: Mechanism of antitumour action of vorinostat (suberoylanilide hydroxamic acid), a novel histone deacetylase inhibitor. *British Journal of Cancer*, 95, S2–S6. <https://doi.org/10.1038/sj.bjc.6603463>.
- Saldívar-González, F. I., Naveja, J. J., Palomino-Hernández, O., & Medina-Franco, J. L. (2017). Getting SMART in drug discovery: Chemoinformatics approaches for mining structure–multiple activity relationships. *RSC Advances*, 7(2), 632–641. <https://doi.org/10.1039/C6RA26230A>.
- Sarkaria, S. M., Christopher, M. J., Klco, J. M., & Ley, T. J. (2014). Primary acute myeloid leukemia cells with IDH1 or IDH2 mutations respond to a DOT1L inhibitor in vitro. *Leukemia*, 28(12), 2403–2406. <https://doi.org/10.1038/leu.2014.235>.
- Shanmugasundaram, V., & Maggiora, G. M. (2001). In *Characterizing property and activity landscapes using an information-theoretic approach CINF-032. Presented at the 222nd ACS national meeting, Chicago, IL*.
- Wassermann, A. M., Wawer, M., & Bajorath, J. (2010). Activity landscape representations for structure–activity relationship analysis. *Journal of Medicinal Chemistry*, 53(23), 8209–8223. <https://doi.org/10.1021/jm100933w>.
- Xu, W. S., Parmigiani, R. B., & Marks, P. A. (2007). Histone deacetylase inhibitors: Molecular mechanisms of action. *Oncogene*, 26(37), 5541–5552. <https://doi.org/10.1038/sj.onc.1210620>.

## **6. Desarrollo y aplicación de métodos nuevos**

### **Ideas clave**

Durante el desarrollo de este proyecto surgieron oportunidades para crear nuevas herramientas de análisis quimioinformático. Considerando que estas herramientas nos resultaron útiles, decidimos compartirlas con la comunidad científica.

### **Series de análogos, núcleos putativos y mapas de constelaciones**

Estos tres conceptos están muy relacionados y se exponen con detalle en cada uno de los primeros tres artículos siguientes. La justificación de estas metodologías radica en que la confiabilidad de los ensayos de alto rendimiento no es elevada, puesto que están diseñados principalmente para probar grandes cantidades de compuestos, por lo que las moléculas que surgen de estos estudios siempre requieren de estudios confirmatorios. Propusimos que es más robusto analizar familias de moléculas (series de análogos) que muestran sistemáticamente cierto efecto en estos ensayos. En los primeros dos artículos se expone la metodología precisa con la que se logra encontrar series de análogos y núcleos putativos; la idea central es que dos moléculas son análogas si se pueden mapear a un núcleo putativo común por reglas de retrosíntesis (p.ej. RECAP), y este núcleo representa una proporción considerable de ambas moléculas. El tercer artículo presenta los mapas de constelaciones: un método de visualización para representar el espacio químico de las series de análogos.

### **Otros métodos y aplicaciones**

El cuarto artículo presenta a “ChemMaps” una metodología para visualizar el espacio químico de bibliotecas químicas muy grandes. En el quinto artículo se presenta el método de barrido de panoramas de actividad; consiste en identificar subgrupos de moléculas en el espacio químico y analizar las relaciones estructura-actividad para cada subgrupo. Los artículos sexto y séptimo presentan el análisis quimioinformáticos, utilizando algunos de los métodos ya mencionados, aplicado en una biblioteca de xenoestrógenos y compuestos químicos presentes en alimentos, respectivamente. El último escrito presentado en esta tesis es un capítulo de libro que resume las técnicas disponibles hasta el momento para estudiar la polifarmacología.



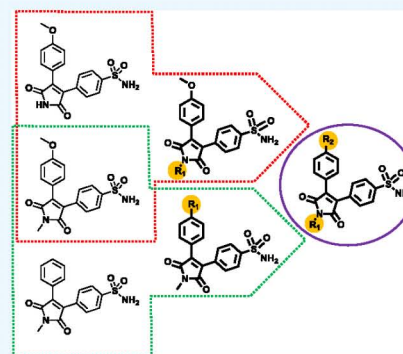
# Systematic Extraction of Analogue Series from Large Compound Collections Using a New Computational Compound–Core Relationship Method

J. Jesús Naveja,<sup>†,‡,§,||</sup> Martin Vogt,<sup>†,||</sup> Dagmar Stumpfe,<sup>†</sup> José L. Medina-Franco,<sup>§</sup> and Jürgen Bajorath<sup>\*,†,||</sup>

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

<sup>‡</sup>PECEM, Faculty of Medicine and <sup>§</sup>Department of Pharmacy, School of Chemistry, Universidad Nacional Autónoma de México, Mexico City 04510, Mexico

**ABSTRACT:** Chemical optimization of organic compounds produces a series of analogues. In addition to considering an analogue series (AS) or multiple series on a case-by-case basis, which is often done in the practice of chemistry, the extraction of analogues from compound repositories is of high interest in organic and medicinal chemistry. In organic chemistry, ASs are a source of alternative synthetic routes and also aid in exploring relationships between compounds from different sources including synthetic vs. naturally occurring molecules. In medicinal chemistry, ASs are the major source of structure–activity relationship information and of hits or leads for drug development. ASs might be identified in different ways. For a given reference compound, a substructure search can be carried out using its scaffold. Alternatively, matched molecular pairs can be calculated to retrieve analogues from a compound repository. However, if no query compounds are used, the identification of ASs in databases is a difficult task. Herein, we introduce a computational approach to systematically identify ASs in collections of organic compounds. The approach involves compound decomposition on the basis of well-established retrosynthetic rules, organization of compound–core relationships, and identification of analogues sharing the same core. The method was applied on a large scale to extract ASs from the ChEMBL database, yielding more than 30 000 distinct series.



## 1. INTRODUCTION

In medicinal chemistry, hit-to-lead and lead optimization campaigns produce a series of analogues. An analogue series (AS) is generally defined as a series of compounds that share the same core structure and carry different R-groups at single or multiple substitution sites.<sup>1</sup> ASs are conventionally represented in R-group tables and are the major source of structure–activity relationship (SAR) information.<sup>1–4</sup> They are usually investigated as individual series in the course of chemical optimization. Computational methods have been introduced to organize large ASs and monitor SAR progression.<sup>2–5</sup>

Going beyond the analysis of individual ASs, another important task is searching for analogues in compound libraries and databases. If one is interested in identifying analogues of given reference compound(s), substructure search approaches can be applied using the core structure of a reference compound as a query.<sup>1,6</sup> For example, this might be attempted in hit expansion when searching for analogues of an interesting active compound. Furthermore, analogues of reference compounds can also be identified without a predefined core structure by searching for matched molecular pairs (MMPs).<sup>7</sup> An MMP is defined as a pair of compounds

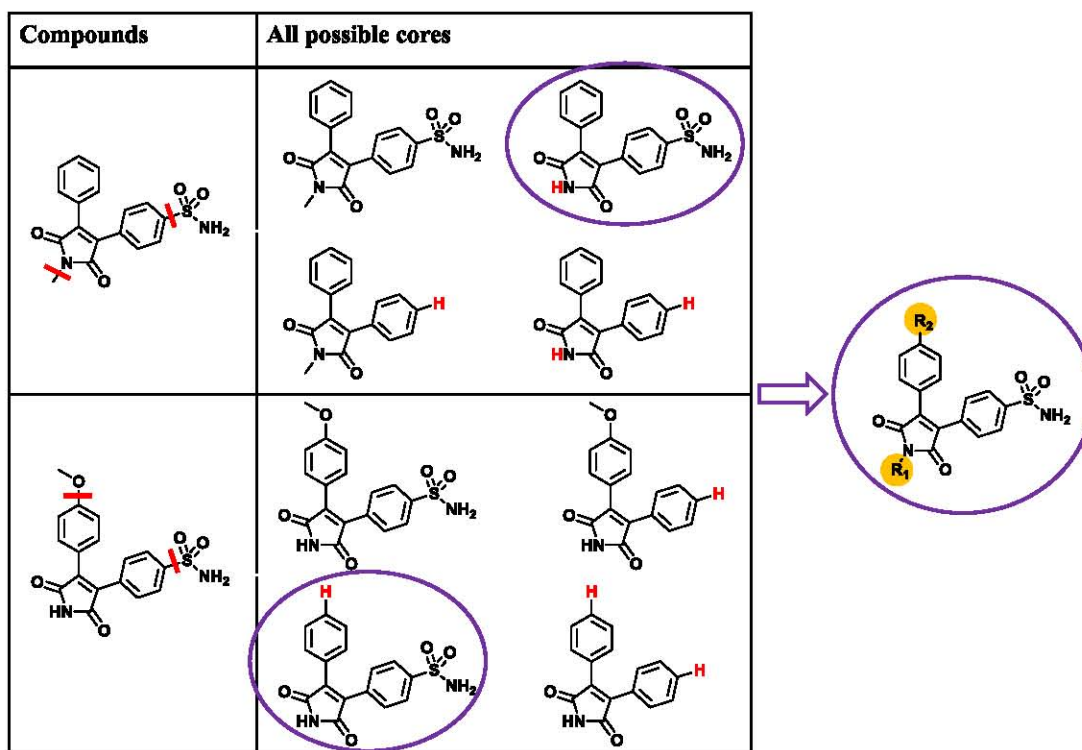
that are only distinguished by a structural modification at a single site.<sup>8</sup> This modification can be rationalized as the exchange of a pair of substructures or a chemical transformation.<sup>9</sup> To detect analogue relationships via MMPs, chemical transformations are restricted in size to focus on typical R-group replacements.<sup>7,10</sup> MMPs can be efficiently generated algorithmically,<sup>9</sup> making MMP-based analogue searching generally applicable<sup>7</sup> and an attractive alternative to substructure search methods.

A much more difficult task than query-based analogue searching is the identification of ASs in large compound data sets, without prior knowledge. However, this task is highly relevant for knowledge extraction from compounds and activity data. In medicinal chemistry, one would like to identify and extract ASs of any composition from heterogeneous compound sources to maximize SAR information retrieval and provide templates for compound optimization. However, to the best of our knowledge, only one computational method for the systematic identification of ASs has so far been

**Received:** December 3, 2018

**Accepted:** January 3, 2019

**Published:** January 14, 2019



**Figure 1.** Concept of the compound–core relationship method. The schematic representation illustrates the identification of analogue series using the CCR approach. For two exemplary compounds (left), all possible cores are shown resulting from the application of retrosynthetic rules and replacement of substitution sites with hydrogen atoms (generalization). In compounds (left), sites of retrosynthetic bond elimination are indicated by red lines. In cores (middle), generalized substitution sites are indicated by red hydrogen atoms. For the two analogues, the largest identical generalized cores and the reconstructed core with two substitution sites (right) are encircled (purple). The reconstructed core contains the invariant sulfonamide group.

introduced.<sup>11</sup> This approach is also based upon the MMP formalism. For a given data set, all possible MMPs are generated and organized in an MMP-based network in which nodes represent compounds and edges pairwise MMP relationships. In this network, separate MMP clusters (MMPCs) are formed by individual ASs that can hence be easily identified.<sup>11</sup> Accordingly, this approach is termed herein an MMP cluster (MMPC)-based method. In the simplest case, an AS from a cluster is formed by a matching molecular series (MMS)<sup>12</sup> having a single substitution site. However, separate clusters in the MMP-based network can also be formed by multiple and overlapping MMSs representing ASs with multiple substitution sites.<sup>11</sup> In this case, each participating MMS contributes a unique single site.

Herein, we introduce another computational methodology for the systematic identification of ASs in repositories of organic compounds, which does not rely on the MMP formalism. Rather, it is based upon the decomposition of single compounds according to well-established retrosynthetic rules and subsequent organization of compound–core relationships (CCRs). In a large-scale application, this new compound–core relationship (CCR) method was applied to systematically extract ASs from the ChEMBL database<sup>13</sup> and compared with the MMPC approach.

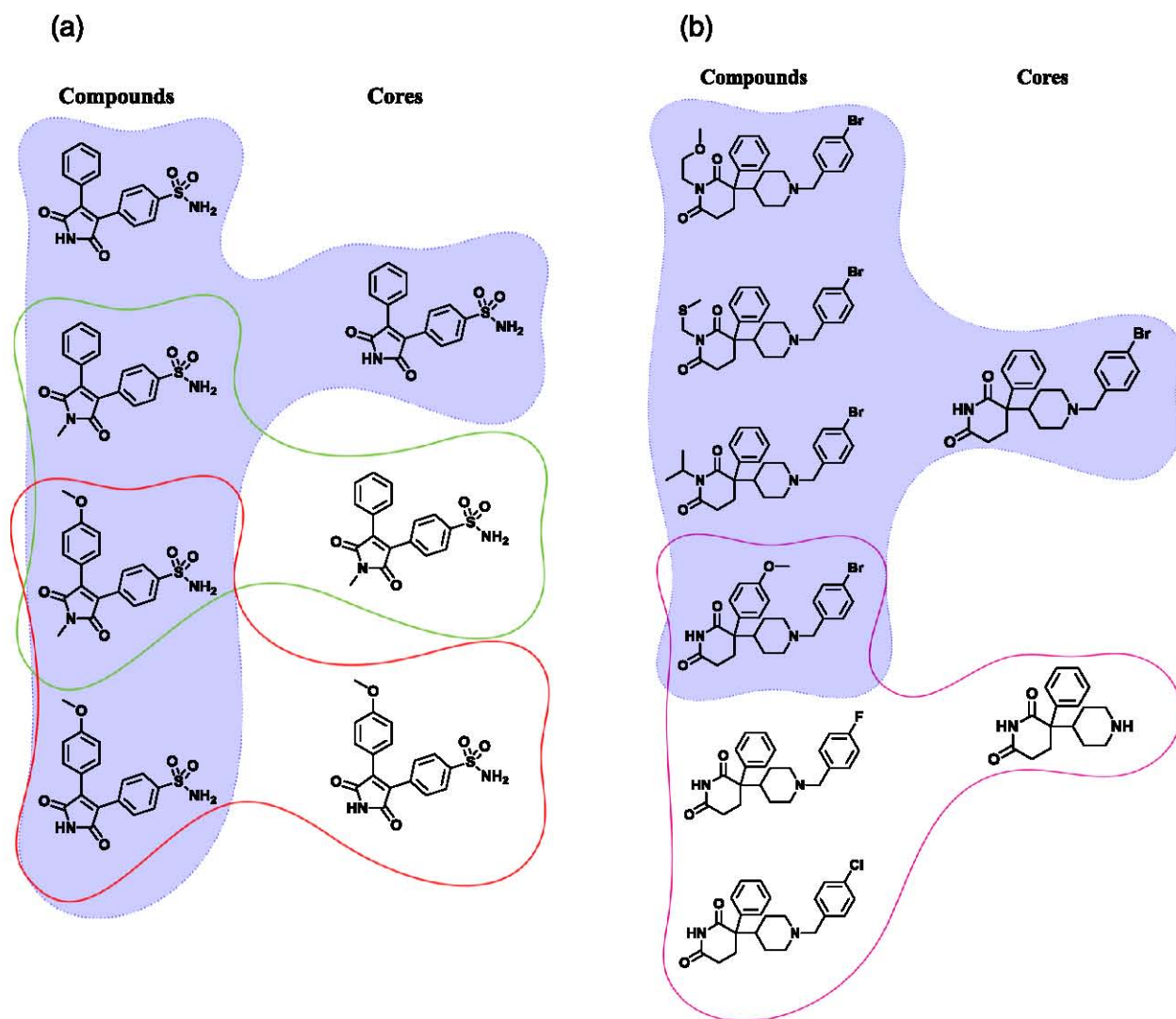
## 2. RESULTS AND DISCUSSION

Extracting ASs from compound repositories without prior knowledge or query compounds is a difficult task. The CCR method introduced herein for systematically identifying ASs in

databases of any composition is conceptually simple and generally applicable. The method comprises three sequential steps, including the generation of cores, exploration of compound–core relationships, and identification of analogue series, which are discussed in the following sections.

**2.1. Methodological Concept.** **2.1.1. Generation of Cores.** The primary goal of the method is the identification of core structures and corresponding analogues such that the compounds can be readily reconstructed from the cores by substitutions at one or more sites and organized into ASs. The basis for reconstruction is provided by systematically applying combinations of possible bond deletions in each compound using retrosynthetic rules.

Specifically, for each database compound, all possible combinations of one to five (or any other predefined number of) bonds are systematically subjected to retrosynthetic cleavage. Hence, a maximum number of five substitution sites per AS are covered. Each combination of applicable retrosynthetic rules leading to the corresponding elimination of single or multiple bonds yields a potential core. The core is considered valid if it consists of a single substructure containing an individual end point (substitution site) for each cleaved bond. Figure 1 illustrates the generation of cores for two analogues having two retrosynthetic cleavage sites. In addition to the three cores obtained from each analogue through retrosynthetic modification, each original compound is recorded as a core with no cleavage sites. Substitution sites in cores are recorded. Furthermore, it is required that the core and eliminated fragments (substituents) meet a predefined size



**Figure 2.** Compound–core relationships and identification of analogue series. (a) AS associated with three retrosynthetic cores. The core at the top represents all analogues (depicted on a purple background), whereas the two remaining cores represent two analogues each (encircled in green and red, respectively). (b) Two overlapping ASs are shown, each of which is associated with an individual core. The core at the top represents four analogues (depicted on a purple background) and the core at the bottom three (encircled in red). One of the analogues is shared by both series.

ratio. In our proof-of-concept study presented herein, we applied the rule that the core must contain at least two-thirds of the heavy atoms comprising the original compound. In other words, the ratio of the number of heavy atoms in the core to the sum of the total number of heavy atoms in all substituents must be at least 2:1. If these requirements are met, a core is accepted for further analysis. For a given database, all possible cores are generated and then “generalized”. During the generalization step, all substitution sites are disregarded by introducing hydrogen atom substitutions at each site such that different cores become identical if they only differ in the position of substitution sites. In Figure 1, two identical cores resulting from generalization are highlighted.

**2.1.2. Exploration of Compound–Core Relationships.** Original database compounds that are identical to hydrogen-substituted cores are assigned to the corresponding cores as the smallest possible analogues. Generalization of cores is followed by reconstruction of recorded substitution sites and the assignment of additional database compounds to cores that differ at given substitution sites. The generalization and

reconstruction steps ensure that compounds with all possible substitutions are assigned to corresponding cores, for example, analogues with ortho-, meta-, and/or para-substitution at one or more rings. We note that this cannot be accomplished on the basis of MMPs. The assignment of compounds to cores with reconstructed substitution sites yields all possible compound–core relationships in an organized form. Figure 1 illustrates the reconstruction of a single core with two substitution sites representing two exemplary analogues.

**2.1.3. Identification of Analogue Series.** An AS is formed if at least two compounds are associated with a core. Because all possible cores meeting the acceptance criteria are involved in CCRs, analogues forming an AS are often associated with multiple cores. ASs might consist of distinct sets of analogues, i.e., analogues belonging to one and only one AS, or overlapping sets of compounds. In addition, an AS might be fully contained as a subset in another series. The latter case is disambiguated by removal of ASs forming a subset of another. In addition, if two ASs contain exactly the same analogues, the one associated with the larger core is retained.

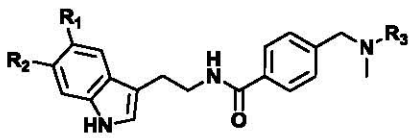




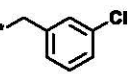


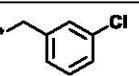


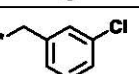


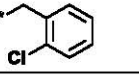


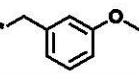


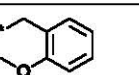
Other possible cases must also be taken into consideration. Figure 2a illustrates the frequently observed situation that an AS is associated with multiple cores. One of the cores might represent the entire AS and others subsets of analogues comprising the series. In this case, the core associated with the entire AS is retained to represent the series.

Figure 2b shows an example of overlapping ASs having different cores. The series share one analogue that is associated with both cores. This example also illustrates the rationale for consistently applying core/substituents size ratio restrictions. We note that the smaller core at the bottom in Figure 2b is a substructure of the larger one at the top. Due to the applied 2:1 size ratio restriction, the three analogues at the top are not presented by the small core at the bottom. This provides a basis for separating the series into two smaller ASs. The confined set of six analogues in this example could have been easily combined into a single AS by assigning two cores to the series. However, application of the size ratio restriction as a criterion for separating overlapping series generally avoids the situation that increasingly large compounds associated with cores that are substructures of each other form elongated “pseudo-AS” that might be artificial in nature and not meaningful chemically. Albeit rarely observed (see Section 2.2), this possible complication should strictly be avoided to ensure chemical relevance of computed ASs. Therefore, in overlapping ASs, each analogue is assigned to the largest AS it belongs to and removed from others. If the number of compounds in alternative ASs is the same, the AS associated with the larger core is selected. Furthermore, if the cores have an identical size, preference is given to the one with fewer substitution sites. Application of these criteria ensures that nearly all overlapping series are disambiguated, as further discussed below. The protocol outlined above guarantees that each AS is ultimately associated with a single core and each compound is associated with no more than one AS. Distinguishing between different CCRs is also of practical relevance. The consistent association of analogues and cores on the basis of size ratio restrictions and the selection of largest possible cores ensures that newly identified ASs are well-defined and can be easily represented in standard R-group tables, as illustrated in Figure 3. Hence, such ASs are readily available for follow-up analysis in medicinal chemistry.

**2.2. Evaluation.** **2.2.1. Large-Scale Search Application.** In a proof-of-concept application, the CCR method was applied to systematically search for ASs in 244 704 active compounds from the ChEMBL database (for details, see the Materials and Methods section). A total of 30 431 ASs containing 145 269 compounds were identified, 8359 of which contained cores with multiple substitution sites. Table 1 reports the size distribution of these ASs, 90% containing between two and nine analogues, 7.5% containing between 10 and 19, and 2.5% containing more than 19 analogues. Furthermore, with increasing size, the proportion of ASs with multiple substitution sites and the average number of substitution sites per AS also increased. For example, the 768 ASs containing at least 20 analogues included 380 series with multiple substitution sites and had on average close to two substitution sites per AS (with a maximum of sites).

Importantly, 18 606 (61%) of the identified ASs containing 64 323 compounds were nonoverlapping and associated with a single core representing the entire series, corresponding to the example shown in Figure 2a. Furthermore, 11 825 ASs (39%) containing 80 946 compounds were obtained from a set of 24



ChEMBL ID	R <sub>1</sub>	R <sub>2</sub>	R <sub>3</sub>
3263732			
2363733			
2363731			
2363730			
2363737			
2363735			

**Figure 3.** Representing identified analogue series in R-group tables. A conventional R-group table for an AS with three substitution sites (R<sub>1</sub>–R<sub>3</sub>) is shown. Six exemplary analogues are listed. The core representing the AS is shown at the top. For each compound, the ChEMBL ID is provided.

202 initially overlapping AS, as illustrated in Figure 2b. Most of the overlapping ASs were separated into well-defined series by uniquely assigning each compound to a single core. Disambiguation (as detailed above) was not possible for a very small subset of 96 overlapping ASs, 82 of which contained less than five compounds.

Thus, taken together, the results of systematic search calculations using the CCR method revealed that the majority of newly identified ASs was distinct from others. In cases where series overlap was detected, separation into non-overlapping ASs was mostly unambiguous. Pseudo-ASs were not detected.

**2.2.2. Method Comparison.** For comparison, search calculations on the basis of 244 704 ChEMBL compounds were repeated using the MMPC approach,<sup>11</sup> the only other computational methodology available to date for systematically identifying ASs. The results are reported in Table 2. MMPC calculations identified 22 111 ASs that covered a total of 103 154 ChEMBL compounds. These series included 3509 ASs (15.9%) with multiple substitution sites. In contrast, the CCR search calculations detected 30 431 ASs that covered a total of 145 326 compounds and included 8359 ASs (27.5%) with multiple substitution sites. Most of the ASs obtained by MMPC were also detected using the CCR method, with some variation in the composition of individual (especially larger) series. Moreover, nearly all analogues (97%) obtained by MMPC were identified using the CCR approach, which yielded 45 508 additional analogues. MMPC calculations yielded 2191 ASs comprising 10 or more analogues. Of these series, 1986 ASs (91%) having more than 50% compound overlap were also identified by CCR including 1406 ASs with at least 80% compound overlap and 730 identical ASs. The

**Table 1. Composition of Analogue Series Identified in ChEMBL Using the CCR Method<sup>a</sup>**

# analogues/series	# series (%)	# series (%), multiple substitution sites	average # substitution sites
2–9	27 391 (90.0%)	7046 (25.7%)	1.38
10–19	2272 (7.5%)	933 (41.1%)	1.71
>19	768 (2.5%)	380 (49.5%)	1.93

<sup>a</sup>Reported are the size distribution of ASs and the fraction of ASs per size range having multiple substitution sites. In addition, the average number of substitution sites per AS of increasing size is given.

**Table 2. Comparison of MMPC- and CCR-Based Retrieval of Analogue Series from ChEMBL<sup>a</sup>**

method	MMPC	CCR
# compounds in ASs	103 154	145 269
# ASs	22 111	30 431
# ASs (%), multiple substitution sites	3509 (15.9%)	8359 (27.5%)

<sup>a</sup>For the MMPC and CCR methods, the total number of ASs extracted from ChEMBL, the number of compounds forming these ASs, and the number (percentage) of ASs with multiple substitution sites are reported.

overlap was calculated as the Jaccard index, i.e., the ratio of the number of shared analogues to the total number of unique analogues in a pair of corresponding series. CCR calculations identified a total of 3040 ASs with 10 or more analogues including 1352 ASs that were not detected using MMPC.

The MMPC/CCR comparison showed that the CCR method identified a significantly larger number of ASs, with a larger proportion of series having multiple substitution sites, and achieved a larger global compound coverage.

**2.4. Conclusions.** The identification of ASs in compound repositories without prior knowledge is of considerable relevance for the practice of organic and medicinal chemistry. ASs and the associated activity information can be used to rationalize and/or guide chemical synthesis and optimization efforts. However, only little has been done so far to automatically identify and extract ASs from databases, leaving much room for further developments. Herein, we have introduced a new computational approach to systematically search for ASs. The CCR method relies on the decomposition of single compounds on the basis of retrosynthetic rules, systematic generation of cores and compound–core relationships, and identification of ASs on the basis of organized and prioritized relationships. By design, the methodology is conceptually simple yet generally applicable. As such, it is thought to represent an attractive addition to the current repertoire of computational methods with utility for organic and medicinal chemistry. In our proof-of-concept investigation, a systematic search for ASs in ChEMBL identified a large number of ASs. The majority of ASs were nonoverlapping and distinct from others and associated with an individual core representing the entire AS. Such series should be of considerable interest for further SAR analysis and the identification of target-selective or promiscuous compounds. In summary, the CCR method introduced herein represents a new and general approach for systematically identifying ASs. It should be of interest to computational as well as organic and medicinal chemists including investigators aiming to explore relationships between compounds from different sources such as natural products and synthetic compounds. Such analyses will provide interesting topics for future application-oriented research.

### 3. MATERIALS AND METHODS

**3.1. Retrosynthetic Rules.** As retrosynthetic rules for compound decomposition, a well-established set of 13 retrosynthetic combinatorial analysis procedure (RECAP) rules was applied.<sup>14</sup> We emphasize that the CCR methodology does not depend on a given set of rules. Depending on individual preferences or project requirements, any chosen set of reaction/retrosynthetic rules can be used. This is particularly relevant for applications in organic chemistry when new synthesis schemes are explored and compared with others.

**3.2. Core Generation Details.** The systematic generation of cores is among the three central components of the CCR method. Further details are provided. Bonds in compounds are cleaved according to RECAP rules and respective substituents are removed. If multiple RECAP rules are applicable to a given compound, all possible combinations are explored to generate cores. For example, if three rules A, B, and C apply, seven cores are obtained, including three with single cleavage sites (A, B, and C), three with dual sites (A/B, A/C, and B/C), and one with three cleavage sites (A/B/C). However, cores are only accepted to establish compound–core relationships if the ratio of the number of heavy atoms forming the core to the number of heavy atoms of all eliminated substituents is at least 2:1. The number of bonds in a compound to which RECAP rules applied was limited (and rarely larger than 20). Consequently, the exhaustive exploration of all possible combinations and resulting cores did not pose a combinatorial problem in most cases. In addition, the 2:1 size ratio restriction further reduced the number of cores for analyzing compound–core relationships. Nonetheless, a computational time restriction of 100 s per compound was implemented for core generation. However, due to this constraint, only 629 of 244 704 ChEMBL compounds failed to produce cores. The protocol for compound decomposition according to retrosynthetic rules was implemented in Java with the aid of the OEChem toolkit.<sup>15</sup>

**3.3. Implementation of the CCR Algorithm.** The CCR algorithm for systematically identifying ASs, as detailed in the [Results and Discussion](#) section, was implemented in Python.

**3.4. Searching for Analogue Series.** Systematic search calculations using the CCR and MMPC reference methods were carried out in a curated version of ChEMBL release 23.<sup>13</sup> Only compounds with direct interactions (target relationship type “D”) with human targets at the highest confidence level (target confidence score 9) and available  $K_i$  or  $IC_{50}$  values were selected, yielding a total of 244 704 active compounds. The application of these selection criteria was not essential for the analysis but ensured that detected ASs exclusively consisted of compounds for which meaningful activity data were available.

### AUTHOR INFORMATION

#### Corresponding Author

\*E-mail: [bajorath@bit.uni-bonn.de](mailto:bajorath@bit.uni-bonn.de). Phone: 49-228-7369-100.

ORCID 

Jürgen Bajorath: 0000-0002-0557-5714

## Author Contributions

||J.J.N. and M.V. contributed equally to this work.

## Author Contributions

The study was carried out and the manuscript was written with contributions of all authors. All authors have approved the final version of the manuscript.

## Notes

The authors declare no competing financial interest.

## ACKNOWLEDGMENTS

J.J.N. is grateful to Consejo Nacional de Tecnología, Mexico (CONACyT) for a scholarship (grant no. 622969) and to the German Academic Exchange Service (DAAD) for a short-term research grant (program no. 53378443). The authors thank OpenEye Scientific Software for an academic software license.

## REFERENCES

- (1) *The Practice of Medicinal Chemistry*; 3rd ed.; Wermuth, C. G., Ed.; Academic Press-Elsevier: Burlington, San Diego, London, U.K., 2008.
- (2) Agrafiotis, D. K.; Shemanarev, K.; Connolly, P. J.; Farnum, M.; Lobanov, V. S. SAR Maps: A New SAR Visualization Technique for Medicinal Chemists. *J. Med. Chem.* **2007**, *50*, 5926–5937.
- (3) Zhang, B.; Hu, Y.; Bajorath, J. AnalogExplorer: A New Method for Graphical Analysis of Analog Series and Associated Structure-Activity Relationship Information. *J. Med. Chem.* **2014**, *57*, 9184–9194.
- (4) Maynard, A. T.; Roberts, C. D. Quantifying, Visualizing, and Monitoring Lead Optimization. *J. Med. Chem.* **2016**, *59*, 4189–4201.
- (5) Shanmugasundaram, V.; Zhang, L.; Kayastha, S.; de la Vega de León, A.; Dimova, D.; Bajorath, J. Monitoring the Progression of Structure-Activity Relationship Information During Lead Optimization. *J. Med. Chem.* **2016**, *59*, 4235–4244.
- (6) Barnard, J. M. Substructure Searching Methods: Old and New. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 532–538.
- (7) Dimova, D.; Stumpfe, D.; Bajorath, J. Systematic Assessment of Analog Relationships between Bioactive Compounds and Promiscuity of Analog Sets. *Med. Chem. Commun.* **2016**, *7*, 230–236.
- (8) Griffen, E.; Leach, A. G.; Robb, G. R.; Warner, D. J. Matched Molecular Pairs as a Medicinal Chemistry Tool. *J. Med. Chem.* **2011**, *54*, 7739–7750.
- (9) Hussain, J.; Rea, C. Computationally Efficient Algorithm to Identify Matched Molecular Pairs (MMPs) in Large Data Sets. *J. Chem. Inf. Model.* **2010**, *50*, 339–348.
- (10) Hu, X.; Hu, Y.; Vogt, M.; Stumpfe, D.; Bajorath, J. MMP-cliffs: Systematic Identification of Activity Cliffs on the Basis of Matched Molecular Pairs. *J. Chem. Inf. Model.* **2012**, *52*, 1138–1145.
- (11) Stumpfe, D.; Dimova, D.; Bajorath, J. Computational Method for Systematic Identification of Analog Series and Key Compounds Representing Series and Their Biological Activity Profiles. *J. Med. Chem.* **2016**, *59*, 7667–7676.
- (12) Wawer, M.; Bajorath, J. Local Structural Changes, Global Data Views: Graphical Substructure-Activity Relationship Trailing. *J. Med. Chem.* **2011**, *54*, 2944–2951.
- (13) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. ChEMBL: A Large-Scale Bioactivity Database for Drug Discovery. *Nucleic Acids Res.* **2012**, *40*, D1100–D1107.
- (14) Lewell, X. Q.; Judd, D. B.; Watson, S. P.; Hann, M. M. RECAP – Retrosynthetic Combinatorial Analysis Procedure: A Powerful New Technique for Identifying Privileged Molecular Fragments with Useful Applications in Combinatorial Chemistry. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 511–522.

(15) *OEChem TK*, version 1.7.7; OpenEye Scientific Software, Inc.: Santa Fe, NM, 2012.

METHODOLOGY

Open Access

# A general approach for retrosynthetic molecular core analysis

J. Jesús Naveja<sup>1,2\*</sup>, B. Angélica Pilon-Jiménez<sup>2</sup>, Jürgen Bajorath<sup>3</sup> and José L. Medina-Franco<sup>2\*</sup>

## Abstract

Scaffold analysis of compound data sets has reemerged as a chemically interpretable alternative to machine learning for chemical space and structure–activity relationships analysis. In this context, analog series-based scaffolds (ASBS) are synthetically relevant core structures that represent individual series of analogs. As an extension to ASBS, we herein introduce the development of a general conceptual framework that considers all putative cores of molecules in a compound data set, thus softening the often applied “single molecule–single scaffold” correspondence. A putative core is here defined as any substructure of a molecule complying with two basic rules: (a) the size of the core is a significant proportion of the whole molecule size and (b) the substructure can be reached from the original molecule through a succession of retrosynthesis rules. Thereafter, a bipartite network consisting of molecules and cores can be constructed for a database of chemical structures. Compounds linked to the same cores are considered analogs. We present case studies illustrating the potential of the general framework. The applications range from inter- and intra-core diversity analysis of compound data sets, structure–property relationships, and identification of analog series and ASBS. The molecule–core network herein presented is a general methodology with multiple applications in scaffold analysis. New statistical methods are envisioned that will be able to draw quantitative conclusions from these data. The code to use the method presented in this work is freely available as an additional file. Follow-up applications include analog searching and core structure–property relationships analyses.

**Keywords:** Analog series-based scaffold, Analog searching, Core structure–property relationships (CSPR), RECAP, Scaffold, Virtual screening

## Introduction

A general trend in drug discovery through big data is emerging [1]. In this context, many exploratory analyses for finding correlations between chemical data and biological activity have been applied, often with satisfactory results [2]. Nonetheless, many of such models require numerical molecule representations in vectors, as opposed to the complex information enclosed in a chemical structure [3]. Chemical fingerprints, a widely applied representation for converting chemical structures into

information vectors, produce a result even when processing complex structures [4]. It is common that such methods detect chemical similarity between molecules even when a synthetic chemist would struggle to find substantial structure commonalities [5].

In contrast to structural fingerprints, molecular scaffolds (and sub-structure methods in general) are alternative representations intuitively interpretable by a chemist, and scaffold analysis is a more chemically conservative approach than a computational prediction of structural resemblance [5]. Several approaches have been proposed to define and generate scaffolds in a consistent manner [6–8]. One of the earliest and still most common scaffold concepts was proposed by Bemis and Murcko [9] and is exemplified in Fig. 1. Section “a” of this figure shows the Bemis and Murcko scaffolds for olanzapine and albendazole. Interestingly, this scaffold concept has evolved. For instance, hierarchies of scaffolds have been proposed,

\*Correspondence: naveja@comunidad.unam.mx; jose.medina.franco@gmail.com

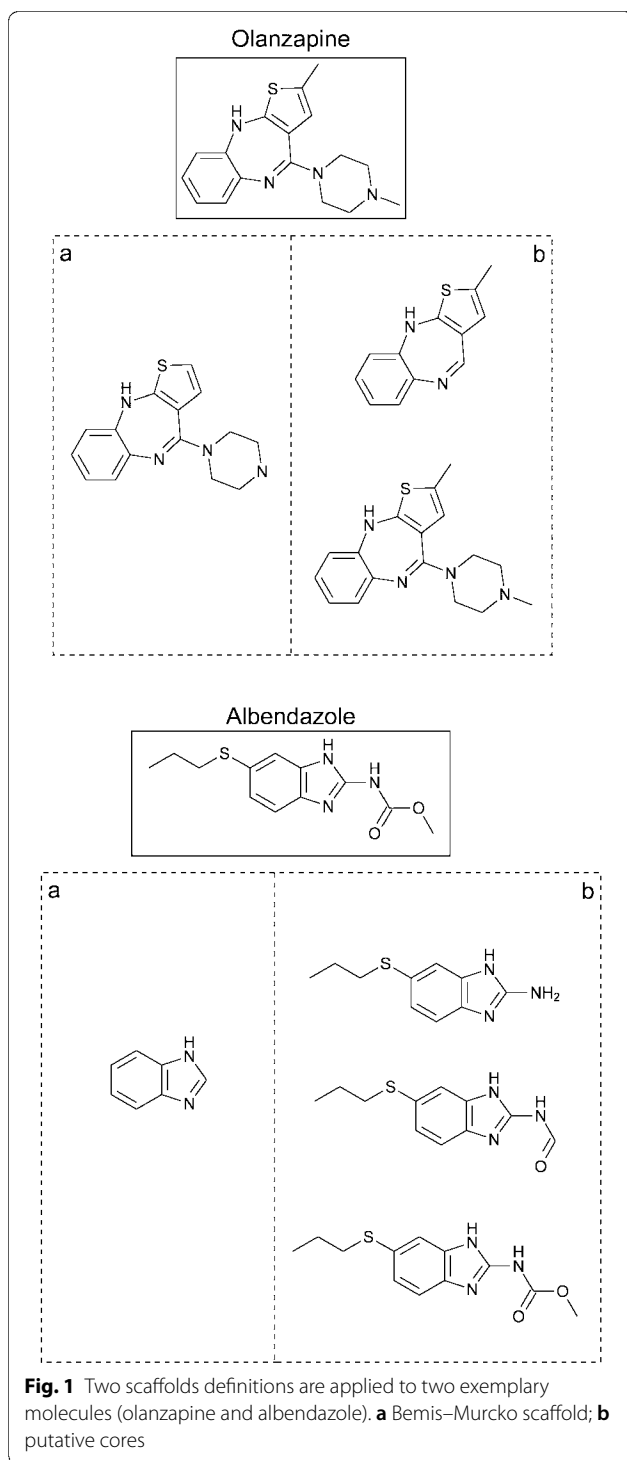
<sup>1</sup> PECEM, School of Medicine, Universidad Nacional Autónoma de México, Avenida Universidad 3000, 04510 Mexico City, Mexico

<sup>2</sup> Department of Pharmacy, School of Chemistry, Universidad Nacional Autónoma de México, Avenida Universidad 3000, 04510 Mexico City, Mexico

Full list of author information is available at the end of the article







which allow to associate scaffolds sharing rings and provide better clustering opportunities than classical scaffold definitions [10–12]. A more comprehensive review on scaffold analysis can be found in [8].

However, these and other classic definitions of scaffolds consider only ring systems, a rather inconvenient feature since it is not uncommon that small rings are conceptualized as side chains or part of substituents by synthetic chemists. Considering the limitations of classical scaffolds, Bajorath et al. developed a novel scaffold concept: the analog series-based scaffold (ASBS) [13] illustrated in section “b” of Fig. 1. In general, ASBS are found through a process that incorporates retrosynthetic information and restrictions in the core/molecule size ratio, thus allowing the identification of chemical analogs that can be summarized in meaningful R-group tables [14, 15]. Hence, ASBS leverage the chemical synthesis and biological relevance of scaffolds [16]. A shortcoming of the current implementation of ASBS is that it depends on the specific dataset [6]. We show below that this is a direct consequence of following the “single molecule–single scaffold” paradigm during the ASBS generation. When using ASBS for analyzing scaffold diversity or comparing scaffolds found in different datasets, it should be taken into consideration that ASBS are by design dataset-dependent.

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

This Methodology paper is organized into two major sections. First, we describe the general approach for constructing molecule–core networks. In the second section, we introduce the application of the method using two case studies, namely: core overlap analysis of two natural products datasets and core structure–activity relationship (CSAR) analysis of an analog series of Akt2 inhibitors. Perspectives for the methodology include, for example, chemical core diversity analysis, advanced SPR, and chemical analog searching. The approach has been used already for the identification of analog series and corresponding scaffolds [15].

## Methods

### Core definition

For any given molecule, a putative core is defined by two criteria [13], herein termed relevance and synthetic feasibility, further clarified as follows:



**Table 1 Comparison of the Bemis–Murcko scaffold and the core framework proposed in this work**

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

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

These two criteria ultimately require the user's input to be further specified. Regarding the first criterion, previous determinations of ASBS have considered a 2:1 ratio of the scaffold vs. all substituents' atoms [13]. The second criterion requires predefining sets of retrosynthesis rules, such as the widely used RECAP rules [17]. A user may implement other sets of available rules [18] or proprietary retrosynthetic schemes.

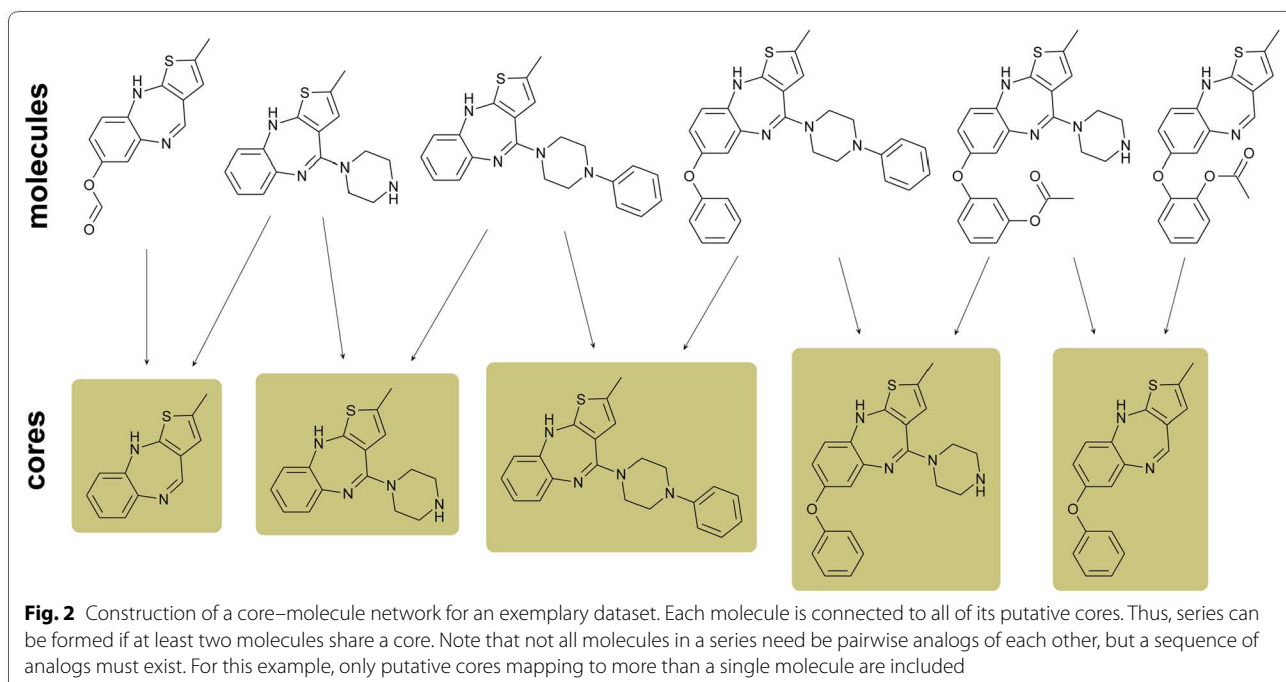
Importantly, given the newly proposed framework, the “single molecule–single core” paradigm underlying various scaffold definitions is no longer compulsory. On the contrary, all substructures of a molecule complying with

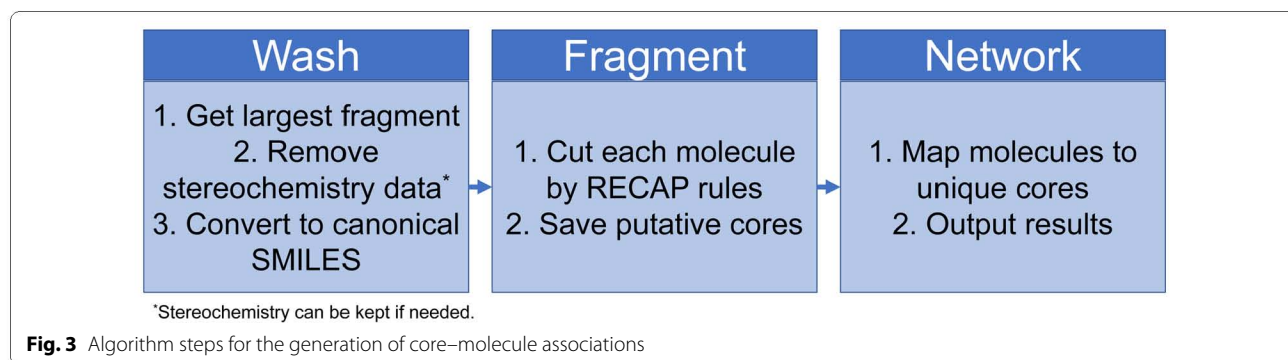
the two criteria above are considered as putative cores, illustrated in Fig. 1b for an exemplary molecule. Our approach is able to include cyclic substructures in both cores and substituents.

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

### Molecule–core network

If the core definition described above is applied to a set of compounds, a bipartite network  $G=(U, V, E)$  can be drawn, where  $U$  is the set of molecules,  $V$  the set of putative cores, and  $E$  the set of edges linking molecules to their putative cores. By definition, if two molecules  $u_1, u_2 \in U$  can be mapped to the same  $v_1 \in V$ , they are considered analogs. An example of a core network is illustrated in Fig. 2, where a set of six exemplary molecules is mapped to all possible cores. Separate clusters represent series. If all compounds in a series can be mapped to a single core, then the series is an analog series, and the comprehensive core is its ASBS. It has been shown that not all sets of related compounds form analog series applying this formalism since in some cases, no single core represents all compounds [15]. Moreover, to a predefined analog series represented by a single core, new molecules might be difficult to add. On the contrary, the use of expandable series with multiple cores makes it easy to include new compounds, which need only to be



**Table 2** Core and Bemis–Murcko scaffold overlap of NuBBE<sub>DB</sub> vs BIOFACQUIM databases

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

decomposed according to the same criteria and incorporated into the network. This is a consequence of accounting for all possible molecule–core relationships.

### Computational implementation

An RDKit–Python [19] implementation of the algorithm is made available in Additional files 1, 2 (see also section Availability of data and materials). The algorithm flow is depicted in Fig. 3. The code is fully parallelized and runs mostly off-memory, which means it can be used to process large chemical libraries. The input is a file with molecular structures represented as SMILES strings as well as an identifier. A “washing” script was added to remove salts, retain the largest molecular component, generate canonical SMILES, and omit stereochemistry information by default. However, stereochemistry can be retained by modifying the data preparation script. Canonical SMILES are annotated with an identifier (WID). Then, each molecule is fragmented independently, and only fragments complying with the core definition (see “Methods”) are saved. Unique cores are annotated with another identifier (MID). Finally, through network analysis, analog series are identified as disjoint subgraphs (clusters). The output is: (1) a file containing molecule–core associations (suffix: “cores.tsv”); (2) a file containing analog series–molecule associations (suffix:

“ASW.tsv”); (3) a file containing analog series–cores associations (suffix: “ASM.tsv”).

### Results

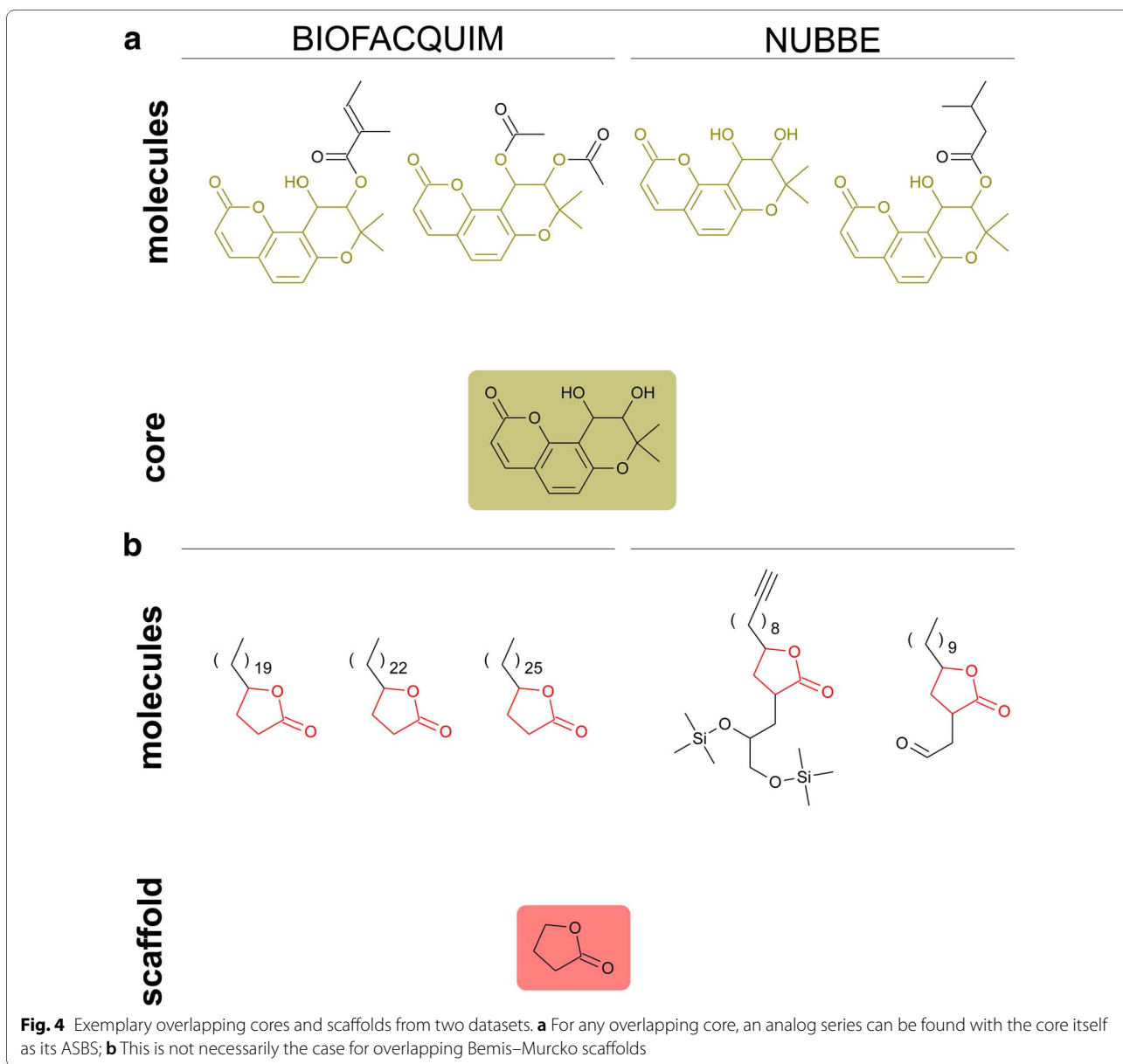
The newly introduced framework has a number of potential applications such as structural analysis of compound databases including structural diversity analysis (based on the new cores), structure–property(–activity) relationships (SP(A)R), and virtual screening [12]). In this section of the Methodology paper, we discuss selected applications of the core framework.

#### Core content analysis

##### Exemplary core overlap analysis in natural product data sets

To illustrate a core overlap analysis we present an example using two publicly available natural product datasets including NuBBE<sub>DB</sub> [20] and BIOFACQUIM [21], which contain information about Brazilian and Mexican natural products, respectively.

The motivation of pursuing a scaffold overlap analysis would be to identify common and unique chemotypes in these databases. As shown in Table 2, NuBBE<sub>DB</sub> and BIOFACQUIM share 49 (~5%) Bemis–Murcko scaffolds and around 106 (~1%) cores. By design, the number of unique Bemis–Murcko scaffolds can only be as high as the total



number of unique molecules, while this is the minimum number of cores that can be found. This explains why more cores than Bemis–Murcko scaffolds are found. Remarkably, if a core is shared between two databases, an analog series might be constructed for that core (Fig. 4a). On the other hand, a shared Bemis–Murcko scaffold, which does not consider the core-to-substituents ratio by design, might not represent a meaningful analog series (Fig. 4b).

Similar overlap analysis can be performed with other larger natural product databases such as the Dictionary of Natural Products [22], the Universal Natural Product Data Set [23] or basically any other compound collection.

Here, we illustrate the method with two natural product datasets as examples. Of note, quantitative diversity metrics remain to be developed, similar to those available to quantify scaffold diversity based on Bemis–Murcko scaffolds [24].

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

Substructure and scaffold-based representations are commonly used in many areas of chemistry. An example is R-group tables to assist in the analysis of SPRs [25, 26]. Considering cores changes the view of SPR analysis. For instance, every collection of molecules linked to a single

core can be considered an analog series, for which SPR can be conducted using an R-group table. Moreover, molecules can be assigned to more than a single core. Therefore, the progression of an analog series can be readily visualized from the core perspective (Fig. 5). Analyzing a database and identifying the most relevant analog series with a given activity, can be considered “lead discovery”. Such an approach prioritizes activity of the analog series over its size measured in the number of analogs it contains. This can be accomplished best by considering the properties in the whole molecule–core network and then selecting enriched cores. Such cores will represent an analog series where the desired property tends to appear, plus different decorations on the scaffold retain the property. Therefore, these cores could be considered leads for drug discovery programs. We call these cores “hit-to-lead cores”, as they can also resemble a hit in the sense that it can be found from exploratory and high-throughput drug discovery campaigns.

#### Exemplary CSAR analysis

Herein, we illustrate the application of CSAR analysis with a dataset of Akt2 inhibitors extracted from ChEMBL 24 [27, 28]. For preprocessing of the data, only compounds with reported  $IC_{50}$  values and standard type “=” were considered. Furthermore, duplicates were removed and the maximum ChEMBL activity values were kept. The dataset was first run through the *cores.py* script (see Additional files 1, 2) and the output was used for CSAR analysis. A Jupyter Notebook with the CSAR analysis is provided as an Additional files 1, 2 as well.

79 series had at least two compounds, and 24 series had at least five. The largest series contained 42 compounds. We analyzed the SAR of this largest series and found that only six cores were connected to more than a single compound. As shown in Fig. 5a, a bipartite network is constructed, where one part of the network is the molecules and the other their putative cores. Edges map molecules to their putative cores. In this way, for any given property, a statistical distribution can be obtained for each core through analogs mapping to the core. Also, the bipartite network allows examining the relevance of the cores. In the example shown in Fig. 5a, the core labeled **M406** represents a larger subset of molecules (represented by red dots at the top of the figure). Note that the cores labeled **M807**, **M808**, **M160**, and **M161** are mapped to the same subset of molecules (Fig. 5a).

The molecule–core bipartite network can be condensed to a core network representation. Figure 5b illustrates a molecule–core network taken the information from Fig. 5a. The network shows the relationship of the core labeled **M406** with five other cores. An edge between two cores means that they share at least one molecule. As in

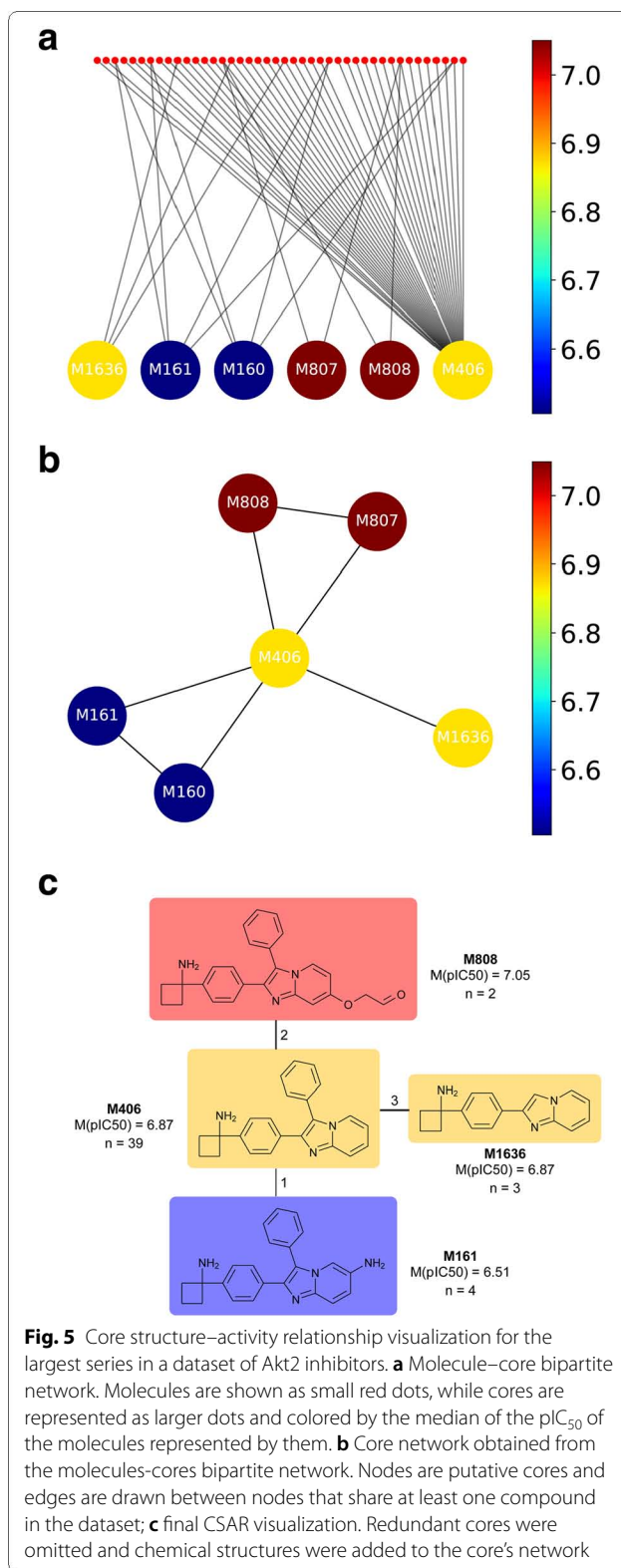




Fig. 5a, the dots in Fig. 5b are colored by the median of the  $pIC_{50}$  of the associated molecules using a continuous color scale. The core network shows that three sub-regions in the CSAR can be found. Furthermore, in this case, there is a gradient, where the most active cores (**M807** and **M808**) are connected to cores with medium activity (**M406**) but not to those with low activity (**M160** and **M161**).

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

#### Identification of analog series and corresponding scaffolds

In a recent publication, a direct application of the core framework for finding ASBS was introduced [15]. By definition, analog series must have a common scaffold and be disjoint from each other according to the paradigm of “single molecule–single scaffold” paradigm. To this end, the initial bipartite network of molecules and their putative cores can be used as a starting point. Then, the number of putative cores has to be reduced to the minimum, and subnetworks are not allowed to overlap. This can be achieved by an iterative greedy selection of cores according to which cores that are more represented in the dataset persist and disqualify secondary cores.

#### Discussion

Scaffold content and diversity analysis are common practice to explore the chemical space of compound data sets and perform classifications based on a structure representation that is highly intuitive [29–31]. There are multiple ways of defining chemical scaffolds or cores (see [32] for a comprehensive review). Of note, hierarchical scaffolds might allow each molecule to have more than a single scaffold. Nevertheless, the level a scaffold occupies in the hierarchy is arbitrary and depends on the dataset. In our general core approach, core structures are followed horizontally, rather than following a hierarchy, as they progress (see Fig. 2). A further issue that remains to be addressed is matching of cores with small chemical changes in rings.

Herein, we have introduced a novel framework for performing scaffold analysis, which is an extension and generalization of the ASBS approach. Several exemplary applications of the approach were presented. In

contrast to the generation of ASBS, where the main objective is representing analog series in a given dataset, our approach avoids any possible information loss as a consequence of not considering all possible molecule–core relationships. In consequence, the new approach generates and stores more data than required for ASBS, but this ensures consistency and interoperability among datasets. Also, for newly generated or updated chemical libraries it is possible to extend the library of cores by only processing new molecules that were added. Only in the context of a chemical dataset, cores can be chosen that represent as many molecules as possible. Reducing the number of cores might be feasible for SPR analysis, but not for comprehensively comparing core overlap between databases.

Among the limitations of the newly presented core framework is the often increased computational cost compared to chemical fingerprint methods or conventional scaffold analysis following Bemis and Murcko. Nonetheless, the off-memory and parallel nature of the scripts make it feasible to process a database as large as ChEMBL\_24 on a desktop computer in less than 24 h. Furthermore, the results depend on the definition of the retrosynthetic rules to be considered and the specific core-to-fragments ratio. We anticipate that the definition of these two parameters impacts the performance of the approach in a given project. Also, as with any approach extracting knowledge retrospectively from a dataset, data quality will obviously affect the analysis.

The method is expected to have the potential for a variety of applications. Given the scope of this Methodology paper, we present two exemplary applications in diversity and SAR analysis. Also, this new framework opens the door to new and more informative SAR visualization approaches. For instance, constellation plots have recently been proposed as a novel approach for visualizing analog series in the chemical space [33].

#### Conclusions and perspectives

In this study, a new and general method inspired by the ASBS concept is introduced. Exemplary applications are shown to establish a proof-of-concept using data from medicinal and natural product chemistry. Scaffold content and diversity analysis are fundamental to characterize compound databases. The results of the recently developed definition of ASBS have proven the chemical and biological usefulness of identifying core scaffolds through retrosynthetic rules and size restrictions. Other applications include the identification of ASBS for hit identification and structure–property analysis. Using the proposed framework, new questions can be answered when comparing datasets, such as how many molecules in a dataset match a synthetic analog in another dataset,

or how often cyclic substructures are found as substituents of a particular core in the context of a given dataset.

Going forward, the new core framework might be systematic to analog searching and core hopping.

## Supplementary information

**Supplementary information** accompanies this paper at <https://doi.org/10.1186/s13321-019-0380-5>.

**Additional file 1.** Source code for getting core data.

**Additional file 2.** A zip file containing a Jupyter Notebook with the exemplary CSAR analysis for the Akt2 dataset, as well as the data and secondary scripts required.

## Abbreviations

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

## Acknowledgements

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

## Authors' contributions

All authors participated in the conception and conceptualization of the study. JJN carried out the analysis and wrote the first draft; BAP-J participated in the scaffold overlap analyses; JLM-F and JB revised the manuscript. All authors read and approved the final manuscript.

## Funding

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

## Availability of data and materials

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

## Competing interests

The authors declare that they have no competing interests.

## Author details

<sup>1</sup> PECEM, School of Medicine, Universidad Nacional Autónoma de México, Avenida Universidad 3000, 04510 Mexico City, Mexico. <sup>2</sup> Department of Pharmacy, School of Chemistry, Universidad Nacional Autónoma de México, Avenida Universidad 3000, 04510 Mexico City, Mexico. <sup>3</sup> Department of Life Science Informatics, B-IT, LIMES Program Unit Chemical Biology and Medicinal Chemistry, Rheinische Friedrich-Wilhelms-Universität, Endenicher Allee 19c, 53115 Bonn, Germany.

Received: 4 March 2019 Accepted: 4 August 2019

Published online: 24 September 2019

## References

- Lusher SJ, McGuire R, van Schaik RC, Nicholson CD, de Vlieg J (2014) Data-driven medicinal chemistry in the era of big data. *Drug Discov Today* 19:859–868. <https://doi.org/10.1016/j.drudis.2013.12.004>
- Lavecchia A (2015) Machine-learning approaches in drug discovery: methods and applications. *Drug Discov Today* 20:318–331. <https://doi.org/10.1016/j.drudis.2014.10.012>
- Vogt M, Bajorath J (2012) Chemoinformatics: a view of the field and current trends in method development. *Bioorg Med Chem* 20:5317–5323. <https://doi.org/10.1016/j.bmc.2012.03.030>
- Lo Y-C, Rensi SE, Torng W, Altman RB (2018) Machine learning in chemoinformatics and drug discovery. *Drug Discov Today* 23:1538–1546. <https://doi.org/10.1016/j.drudis.2018.05.010>
- Bajorath J (2014) Exploring activity Cliffs from a chemoinformatics perspective. *Mol Inform* 33:438–442. <https://doi.org/10.1002/minf.201400026>
- Bajorath J (2018) Improving the utility of molecular scaffolds for medicinal and computational chemistry. *Future Med Chem* 10:1645–1648. <https://doi.org/10.4155/fmc-2018-0106>
- Schneider P, Schneider G (2017) Privileged structures revisited. *Angew Chem Int Ed Engl* 56:7971–7974. <https://doi.org/10.1002/anie.201702816>
- Hu Y, Stumpfe D, Bajorath J (2011) Lessons learned from molecular scaffold analysis. *J Chem Inf Model* 51:1742–1753. <https://doi.org/10.1021/ci200179y>
- Bemis GW, Murcko MA (1996) The properties of known drugs. 1. Molecular frameworks. *J Med Chem* 39:2887–2893. <https://doi.org/10.1021/jm9602928>
- Schuffenhauer A, Ertl P, Roggo S, Wetzel S, Koch MA, Waldmann H (2007) The scaffold tree—visualization of the scaffold universe by hierarchical scaffold classification. *J Chem Inf Model* 47:47–58. <https://doi.org/10.1021/ci600338x>
- Ertl P, Schuffenhauer A, Renner S (2011) The scaffold tree: an efficient navigation in the scaffold universe. *Methods Mol Biol* 672:245–260. [https://doi.org/10.1007/978-1-60761-839-3\\_10](https://doi.org/10.1007/978-1-60761-839-3_10)
- Schäfer T, Kriege N, Humbeck L, Klein K, Koch O, Mutzel P (2017) Scaffold Hunter: a comprehensive visual analytics framework for drug discovery. *J Cheminform* 9:28. <https://doi.org/10.1186/s13321-017-0213-3>
- Stumpfe D, Dimova D, Bajorath J (2016) Computational method for the systematic identification of analog series and key compounds representing series and their biological activity profiles. *J Med Chem* 59:7667–7676. <https://doi.org/10.1021/acs.jmedchem.6b00906>
- Dimova D, Bajorath J (2018) Collection of analog series-based scaffolds from public compound sources. *Future Sci OA* 4:FSO287. <https://doi.org/10.4155/fsoa-2017-0135>
- Naveja JJ, Vogt M, Stumpfe D, Medina-Franco JL, Bajorath J (2019) Systematic extraction of analogue series from large compound collections using a new computational compound–core relationship method. *ACS Omega* 4:1027–1032. <https://doi.org/10.1021/acsomega.8b03390>
- Dimova D, Stumpfe D, Hu Y, Bajorath J (2016) Analog series-based scaffolds: computational design and exploration of a new type of molecular scaffolds for medicinal chemistry. *Future Sci OA* 2:FSO149. <https://doi.org/10.4155/fsoa-2016-0058>
- Lewell XQ, Judd DB, Watson SP, Hann MM (1998) RECAP—retrosynthetic combinatorial analysis procedure: a powerful new technique for identifying privileged molecular fragments with useful applications in combinatorial chemistry. *J Chem Inf Comput Sci* 38:511–522. <https://doi.org/10.1021/ci970429i>
- Watson IA, Wang J, Nicolaou CA (2019) A retrosynthetic analysis algorithm implementation. *J Cheminform* 11:1. <https://doi.org/10.1186/s13321-018-0323-6>
- RDKit: Open-source cheminformatics; 2016. <http://www.rdkit.org>
- Pilon AC, Valli M, Dametto AC, Pinto MEF, Freire RT, Castro-Gamboa I et al (2017) NuBBEDB: an updated database to uncover chemical and biological information from Brazilian biodiversity. *Sci Rep* 7:7215. <https://doi.org/10.1038/s41598-017-07451-x>
- Pilón-Jiménez BA, Saldívar-González FI, Díaz-Eufracio BI, Medina-Franco JL (2019) BIOFACQUIM: a mexican compound database of natural products. *Biomolecules* <https://doi.org/10.3390/biom9010031>
- Taylor and Francis CP. Dictionary of natural products. <http://dnp.chemnetbase.com/faces/chemical/ChemicalSearch.xhtml>. Accessed 12 Feb 2019
- Gu J, Gui Y, Chen L, Yuan G, Lu H-Z, Xu X (2013) Use of natural products as chemical library for drug discovery and network pharmacology. *PLoS ONE* 8:e62839. <https://doi.org/10.1371/journal.pone.0062839>



24. González-Medina M, Prieto-Martínez FD, Owen JR, Medina-Franco JL (2016) Consensus diversity plots: a global diversity analysis of chemical libraries. *J Cheminform*. 8:63. <https://doi.org/10.1186/s13321-016-0176-9>
25. Khire UR, Bankston D, Barbosa J, Brittelli DR, Caringal Y, Carlson R et al (2004) Omega-carboxypyridyl substituted ureas as Raf kinase inhibitors. *Bioorg Med Chem Lett* 14:783–786. <https://doi.org/10.1016/j.bmcl.2003.11.041>
26. Wang M, Xu S, Wu C, Liu X, Tao H, Huang Y et al (2016) Design, synthesis and activity of novel sorafenib analogues bearing chalcone unit. *Bioorg Med Chem Lett* 26:5450–5454. <https://doi.org/10.1016/j.bmcl.2016.10.029>
27. Naveja JJ, Oviedo-Osornio CI, Trujillo-Minero NN, Medina-Franco JL (2018) Chemoinformatics: a perspective from an academic setting in Latin America. *Mol Divers*. 22:247–258. <https://doi.org/10.1007/s11030-017-9802-3>
28. Gaulton A, Hersey A, Nowotka M, Bento AP, Chambers J, Mendez D et al (2017) The ChEMBL database in 2017. *Nucleic Acids Res* 45:D945–D954. <https://doi.org/10.1093/nar/gkw1074>
29. Shang J, Sun H, Liu H, Chen F, Tian S, Pan P et al (2017) Comparative analyses of structural features and scaffold diversity for purchasable compound libraries. *J Cheminform*. 9:25. <https://doi.org/10.1186/s13321-017-0212-4>
30. Koch MA, Schuffenhauer A, Scheck M, Wetzel S, Casaulta M, Odermatt A et al (2005) Charting biologically relevant chemical space: a structural classification of natural products (SCONP). *Proc Natl Acad Sci USA* 102:17272–17277. <https://doi.org/10.1073/pnas.0503647102>
31. Medina-Franco JL, Petit J, Maggiora GM (2006) Hierarchical strategy for identifying active chemotype classes in compound databases. *Chem Biol Drug Des* 67:395–408. <https://doi.org/10.1111/j.1747-0285.2006.00397.x>
32. Langdon SR, Brown N, Blagg J (2011) Scaffold diversity of exemplified medicinal chemistry space. *J Chem Inf Model* 51:2174–2185. <https://doi.org/10.1021/ci2001428>
33. Naveja JJ, Medina-Franco JL (2019) Finding constellations in chemical space through core analysis. *Front Chem*. 7:510. <https://doi.org/10.3389/fchem.2019.00510>

### Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Ready to submit your research? Choose BMC and benefit from:**

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

**At BMC, research is always in progress.**

Learn more [biomedcentral.com/submissions](https://biomedcentral.com/submissions)





# Finding Constellations in Chemical Space Through Core Analysis

J. Jesús Naveja<sup>1,2\*</sup> and José L. Medina-Franco<sup>2\*</sup>

<sup>1</sup>PECEM, School of Medicine, Universidad Nacional Autónoma de México, Mexico City, Mexico, <sup>2</sup>Department of Pharmacy, School of Chemistry, Universidad Nacional Autónoma de México, Mexico City, Mexico

Herein we introduce the constellation plots as a general approach that merges different and complementary molecular representations to enhance the information contained in a visual representation and analysis of chemical space. The method is based on a combination of a sub-structure based representation and classification of compounds with a “classical” coordinate-based representation of chemical space. A distinctive outcome of the method is that organizing the compounds in analog series leads to the formation of groups of molecules, aka “constellations” in chemical space. The novel approach is general and can be used to rapidly identify, for instance, insightful and “bright” Structure-Activity Relationships (StARs) in chemical space that are easy to interpret. This kind of analysis is expected to be especially useful for lead identification in large datasets of unannotated molecules, such as those obtained through high-throughput screening. We demonstrate the application of the method using two datasets of focused inhibitors designed against DNMTs and AKT1.

**Keywords:** analog series, data visualization, descriptor, scaffold, structure-property relationships

## OPEN ACCESS

### Edited by:

Simona Rapposelli,  
University of Pisa, Italy

### Reviewed by:

Chanin Nantasenamat,  
Mahidol University, Thailand  
Oscar Mendez Lucio,  
Bayer, France

### \*Correspondence:

J. Jesús Naveja  
naveja@comunidad.unam.mx  
José L. Medina-Franco  
medinajl@unam.mx

### Specialty section:

This article was submitted to  
Medicinal and Pharmaceutical  
Chemistry,  
a section of the journal  
Frontiers in Chemistry

**Received:** 14 May 2019

**Accepted:** 03 July 2019

**Published:** 16 July 2019

### Citation:

Naveja JJ and Medina-Franco JL  
(2019) Finding Constellations in  
Chemical Space Through Core  
Analysis. *Front. Chem.* 7:510.  
doi: 10.3389/fchem.2019.00510

## INTRODUCTION

The concept of chemical space is broadly used in drug discovery because of its multiple potential applications; for instance, in library design, compound or dataset classification, compound selection, exploration of structure-activity relationships (SAR), and navigation through structure-property relationships (SPR) in general. However, a precise unique definition of chemical space is not simple. An even more challenging task is the visual representation of this subjective concept.

Chemical space is usually defined as the set of all possible organic compounds (Lipinski and Hopkins, 2004). It is widely recognized that the virtual chemical space is more than astronomically large, as not even all atoms in the universe would suffice to synthesize a single molecule from each of all the  $10^{63}$  possible organic compounds of a size up to 30 atoms (Clayden et al., 2012). Nevertheless, massive efforts have been undertaken to enumerate billions of hypothetical organic compounds, thus allowing large virtual screening campaigns to take place (Reymond, 2015; Lyu et al., 2019).

Along with the increasing size of the mapped chemical space, the interest of applying cartographic methods to visualize the space has expanded (Oprea and Gottfries, 2001). As a result, numerous visualization and conceptualization approaches into chemical space have emerged (Larsson et al., 2007; Osolodkin et al., 2015; Naveja and Medina-Franco, 2017). A cornerstone and key aspect of all proposed methods is the molecular representation and parameters used to define the space where the compounds will reside. Chemical space visualizations have to reduce the dimensionality of the problem of comparing molecular structures, which can be done through algorithms such as principal components analysis and t-distributed stochastic neighbor embedding (see Osolodkin et al., 2015).

In most chemical space approaches, it is desirable that chemical analogs are closer to each other than unrelated and dissimilar molecules since this allows machine learning methods to identify clusters of structurally-related molecules (Medina-Franco et al., 2008; Naveja and Medina-Franco, 2015; Naveja et al., 2016, 2018a). In addition, clustering analog series would allow, at least in principle, to map SAR/SPR into that space. However, due to the vast amplitude of the chemical space and the inevitable loss of information with an initially large space projected into lower dimensions, it is expected that non-analog compounds will end up in the same cluster. Also, when many points in the chemical space are considered at once, visualizations become harder to interpret. To address this issue, approaches such as virtual reality have emerged (Probst and Reymond, 2018).

In parallel to such chemical space approaches based on coordinates, scaffold analysis is a more consistent and chemically-intuitive approach for exploring and identifying collections of analogs (Hu et al., 2011). Ever since the pioneering work by Bemis and Murcko (1996), computational identification of chemical scaffolds has been refined. In this line, Stumpfe et al. (2016) recently introduced the analog series-based scaffold (ASBS), a revolutionary scaffold concept that is more flexible and chemically sound than its predecessors. In fact, the ASBS has proven to yield more biologically meaningful structure-activity/property relationships (SA/PR) than other scaffold definitions (Dimova et al., 2016; Kunitomo et al., 2017; Bajorath, 2018; Dimova and Bajorath, 2018).

Although the chemical space of single analog series can be effectively explored and used, for instance, to guide lead optimization programmes (Vogt et al., 2018), methods for analyzing the relationship among scaffolds of different analog series remain to be explored. Of note, a difficulty in this regard emerges as analog-series based scaffolds tend not to be as consistent as Bemis-Murcko scaffolds, since they result from the retrospective analysis of analog series (Bajorath, 2018). Accordingly, a core framework inspired in the design of the ASBS avoids the shortcoming of inconsistency by allowing molecules to be annotated with more than one putative core (Naveja et al., Submitted). Hence, large libraries containing analogs can be condensed into fewer cores. In this way, SA/PR can be preferentially analyzed for the most explored regions of the chemical space: analog series.

Herein, we present a general methodology for applying the putative core framework to produce more concise and meaningful representations of the chemical space. To our understanding, this is the first method designed for charting multiple analog series into a coordinate-based chemical space, thus combining in a single plot two general and useful approaches of molecular representation and mapping. Of note, since within this framework cores may share analogs (i.e., analog series are allowed to share compounds), such cores can be connected, thus resembling constellations in the chemical space. Therefore, we termed the resulting graphics “constellation plots.” As it will be discussed, activity data (or any property of interest) can be mapped into the constellation plot allowing to explore SA/PRs in the space and quickly identify interesting regions in the space. The rest of this methodological paper is organized

as follows: first, the concept scheme is presented and the formalism explained through a toy example; thereafter, two case studies using exemplary datasets are presented; finally, we discuss the conclusions and perspectives of this novel approach for combining the scaffold and the chemical space concepts.

## METHODS

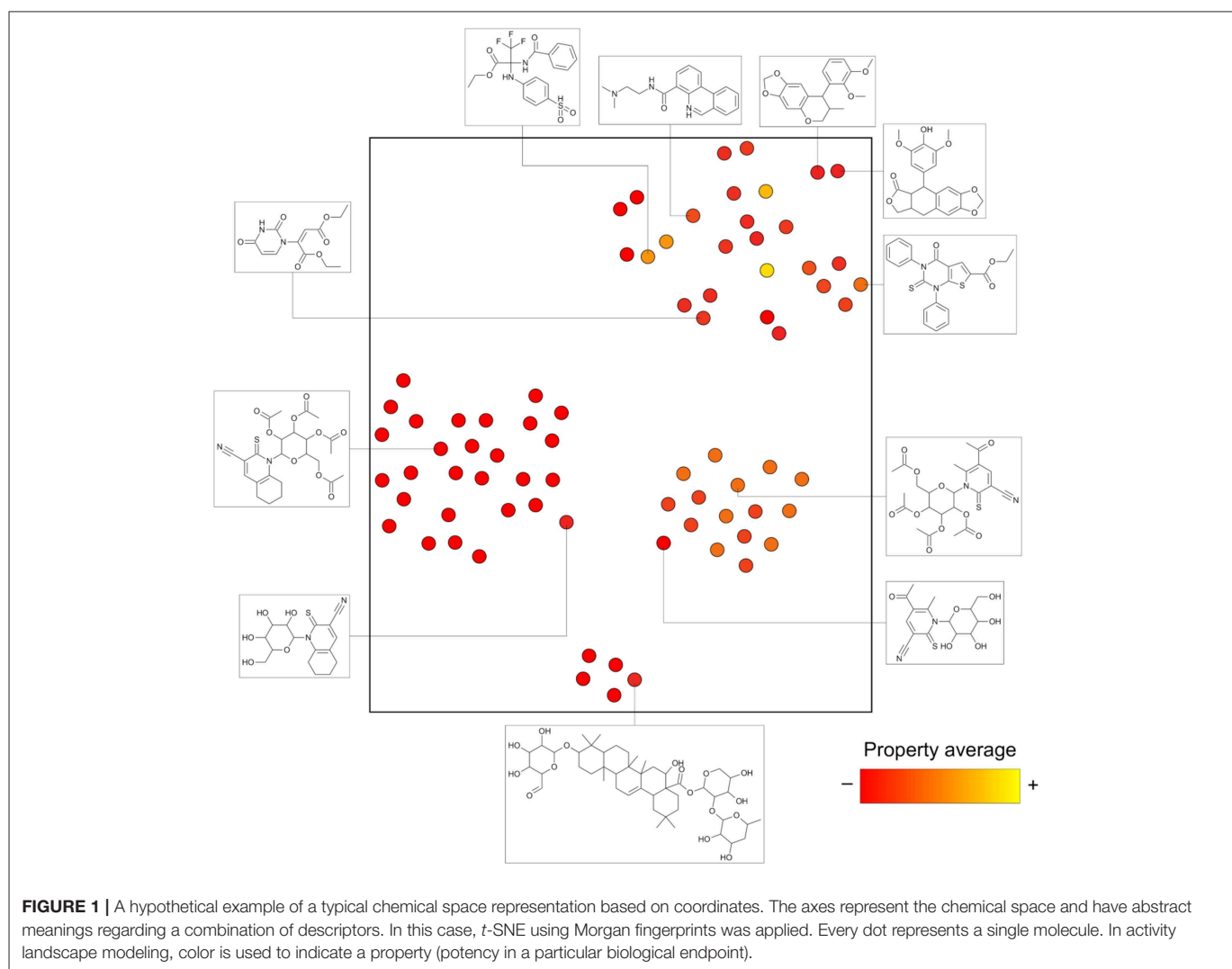
### Datasets Used in the Examples

For illustrating the application of constellation plots in two different context of analysis, we used two benchmark datasets that have been previously explored with other analysis approaches. One set was a group of 827 AKT1 inhibitors extracted and curated from ChEMBL (Gaulton et al., 2017; Naveja et al., 2018b). The second dataset was a collection of 286 compounds tested as inhibitors of DNMT (DNA methyltransferases). This second data set was integrated from multiple sources of information as described in Naveja and Medina-Franco (2018). Since this dataset integrates qualitative (such as those containing crystallographic data) and quantitative databases (such as those containing experimental determination of inhibition curves), for this dataset, we use a categorical classification of activity in “active” or “inactive.” The files of the two datasets are included as **Supplementary Information**.

### Chemical Space and Analog Series

As mentioned above, constellation plots fuse two ligand-based concepts: chemical space and core analysis. Standard chemical space analysis is carried out by computing descriptors for a collection of molecules (e.g., physicochemical properties and/or structural features) and then applying dimensionality reduction approaches (Rosén et al., 2009; Osolodkin et al., 2015; González-Medina et al., 2016; Prieto-Martínez et al., 2016; Naveja and Medina-Franco, 2017; Borrel et al., 2018). As a result, every data point represents a single molecule (see **Figure 1**). This can render many visualizations hard to read and analyze by the naked eye. Furthermore, the numerous descriptors used are combined, such that every axis in the visualization turns out to have a quite abstract meaning. Herein, for the purpose of charting chemical space, *t*-distributed stochastic neighbor embedding (*t*-SNE) is used. This methodology reduces the number of data points in the center of the map as compared to other approaches and has been used successfully in chemical space charting (Maaten and Hinton, 2008; Lewis et al., 2015). However, other coordinate-based representations of chemical space can be used in this general approach.

In contrast to chemical space, standard scaffold and analog series analysis aims toward a clear and consistent picture of the relationships among compounds. For instance, a scaffold is a substructure shared by all compounds annotated with it. A state-of-the-art approach for defining analog series-based scaffolds was proposed by Stumpfe et al. (2016). They have reasoned that for a scaffold to be relevant in medicinal chemistry, it should not only be a substructure of a molecule, but it also has to comply with three additional criteria: (i) be a major component of the whole molecule, (ii) be derived from the molecule through retrosynthetic rules, and (iii) summarize an



analog series in a particular dataset. A number of computational approaches for obtaining ASBS have been proposed (Dimova et al., 2016; Stumpfe et al., 2016; Bajorath, 2018; Naveja et al., 2019). Within these approaches, an analog series is defined as a subnetwork connected by matched molecular pairs (MMPs) (Griffen et al., 2011).

Chemical space analysis of individual analog series has been carried out to measure progression in lead optimization and saturation of analog series (Kunimoto et al., 2018; Vogt et al., 2018; Yonchev et al., 2018). Nevertheless, the fact that assumption (iii) makes analog series inconsistent in as much as the scaffold definition is dependent on the dataset used (Bajorath, 2018) is a limitation for the exploration of chemical space of multiple analog series at once. In a recent study (Naveja et al., Submitted), we discussed that by removing assumption (iii) two effects take place: first, every molecule is allowed to be annotated to more than a single core (equivalent to the term “scaffold”); and second, complete consistency is achieved as no core annotations are ever omitted for any molecule (see **Figure 2**). It is within this general core framework that we propose using constellation plots.

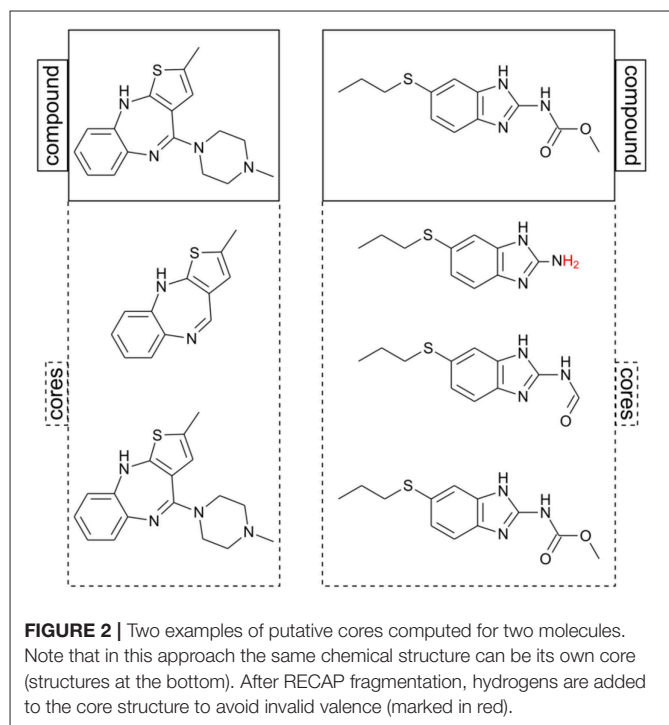
## Summarizing Analog Series Information in a Dataset Within the General Core Framework

Since the general core framework can assign multiple cores to single molecules, a useful step prior to mapping cores in the chemical space would be summarizing analog series in the smallest number of cores possible. As illustrated in **Figure 3**, in some instances it is possible to summarize a whole analog series in a single core structure, while in other cases this cannot be done without loss of information. Hence, for avoiding such situations, we did not discard cores unless only one compound mapped to it. Furthermore, if two or more cores mapped to exactly the same compounds, then only the largest core was kept and the others were disregarded from the analysis.

## Constellation Plots

After processing a collection of compounds under the general core framework, information is obtained in multiple regards, namely: (a) the chemical structure of every core; (b) the sets of

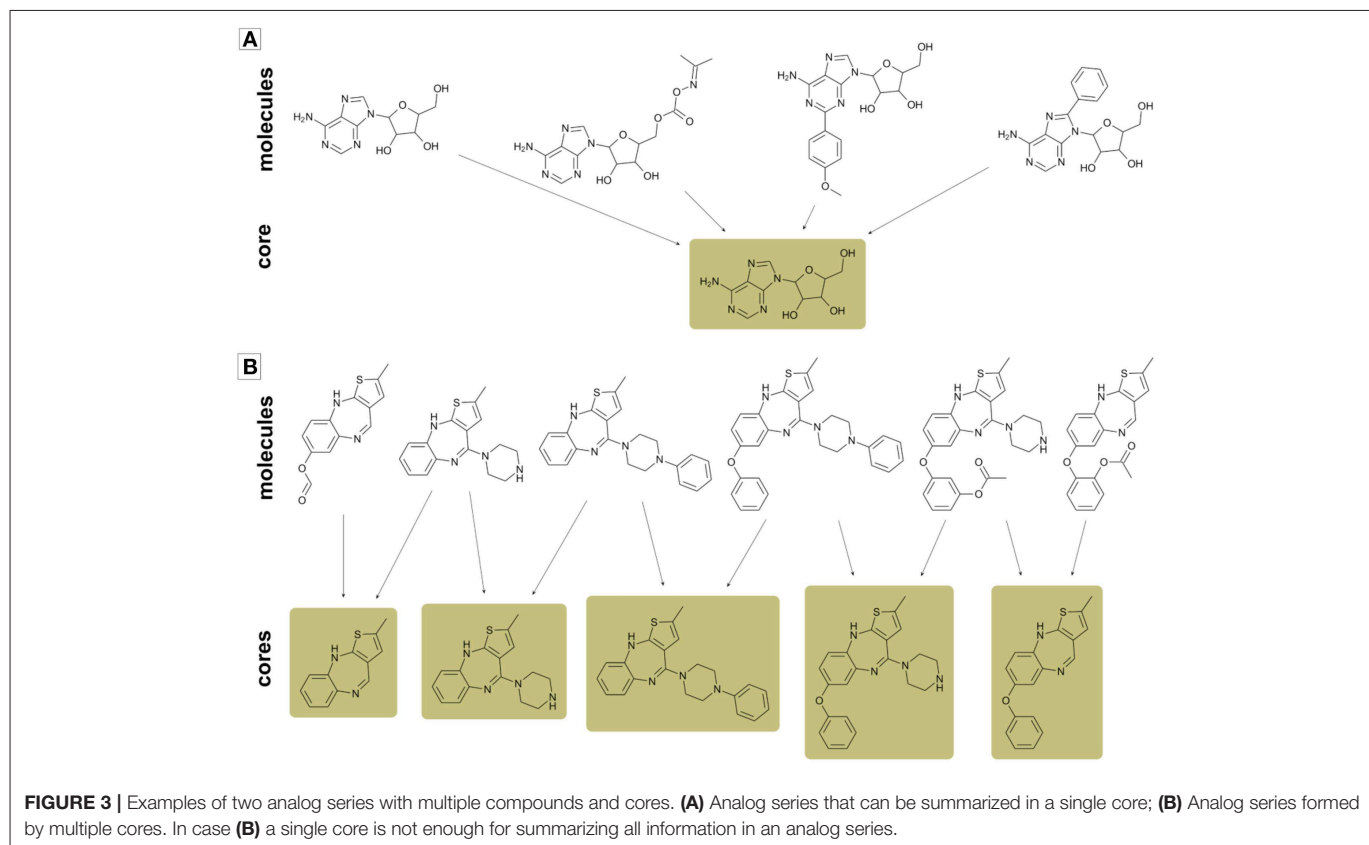
molecules mapping to each core; (c) the molecules annotated to multiple cores; and (d) the analog series to which each compound and core are annotated. We propose a visualization



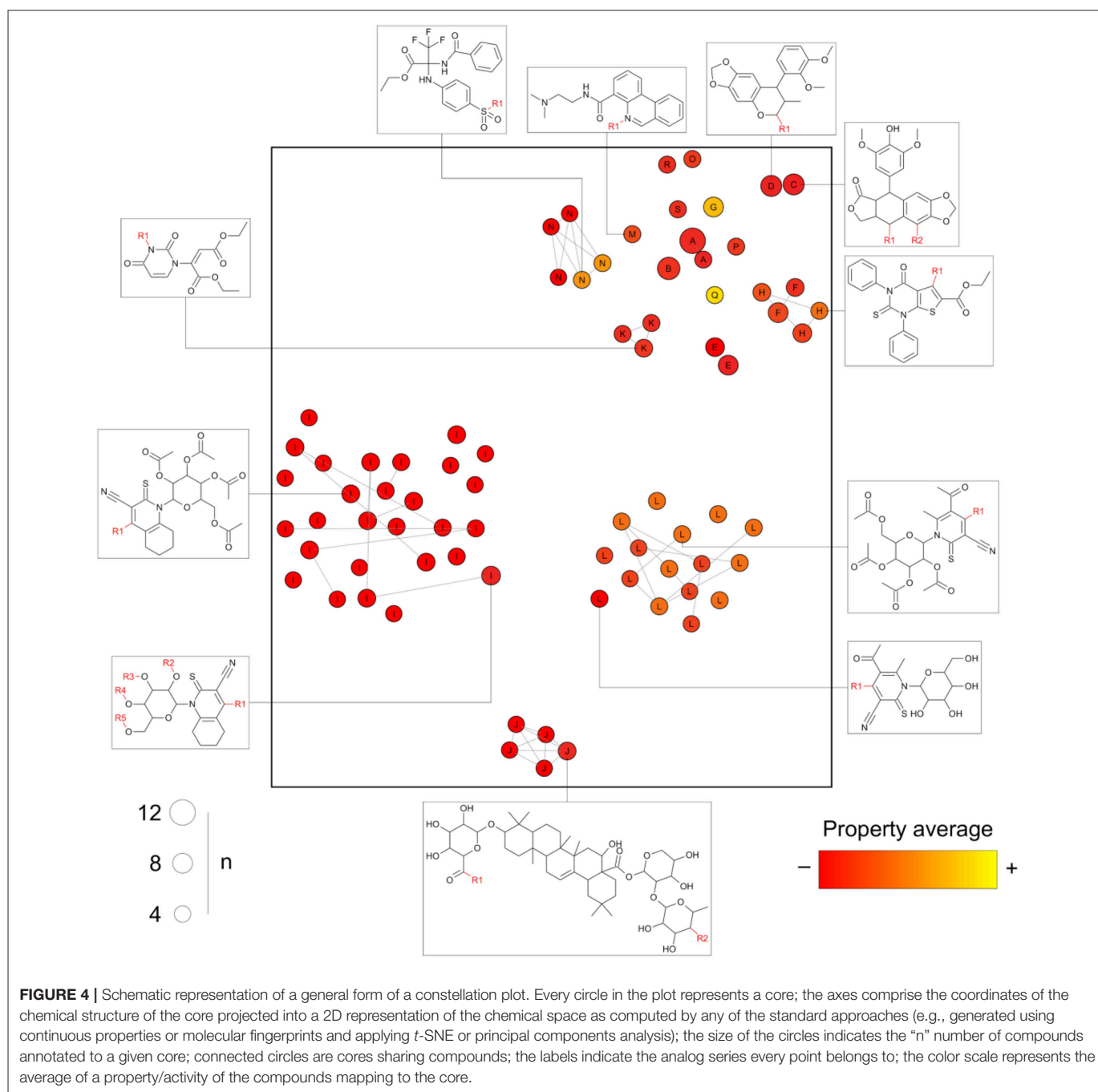
methodology summarizing these four dimensions in a single graphic: the constellation plot that is schematically illustrated in **Figure 4**.

Essentially, in a constellation plot, the chemical structure of representative cores in a database (for example, those annotated with a predefined minimum number of compounds) is used to find descriptors and map them into a chemical space as if they were single molecules. The size of the circles is used to represent the relative number of compounds annotated to each core. Cores sharing compounds are connected by lines forming “constellations” in the chemical space. Every circle is labeled with an identifier for the analog series to which each core belongs. Additionally, a color scale can be used to represent an average of a given property or activity of the compounds annotated with each core, thereby turning constellation plots useful for activity landscape modeling (Waddell and Medina-Franco, 2012). Of note, the activity can be, for instance, measured for a single molecular target. However, the property could also be a representative measure of the selectivity or promiscuity profile of all the compounds sharing a core across multiple biological endpoints (see section Conclusions and Perspectives).

**Figure 4**, as opposed to **Figure 1**, is able to summarize a larger number of compounds than points depicted and contains information about actual analogs. For instance, analog series I, J, and L form separate clusters, but the cluster top right has multiple chemotypes of distinct analog series. This could not be inferred from clustering algorithms applied to the chemical space information only.







## Implementation

All scripts required for producing the data herein reported use free Python code and are made freely available in **Supplementary Information**. RDKit was used for computing fingerprints and manipulation of chemical structures (<http://www.rdkit.org>). Scikit-learn was used for computing *t*-SNE (Pedregosa et al., 2011).

## RESULTS AND DISCUSSION

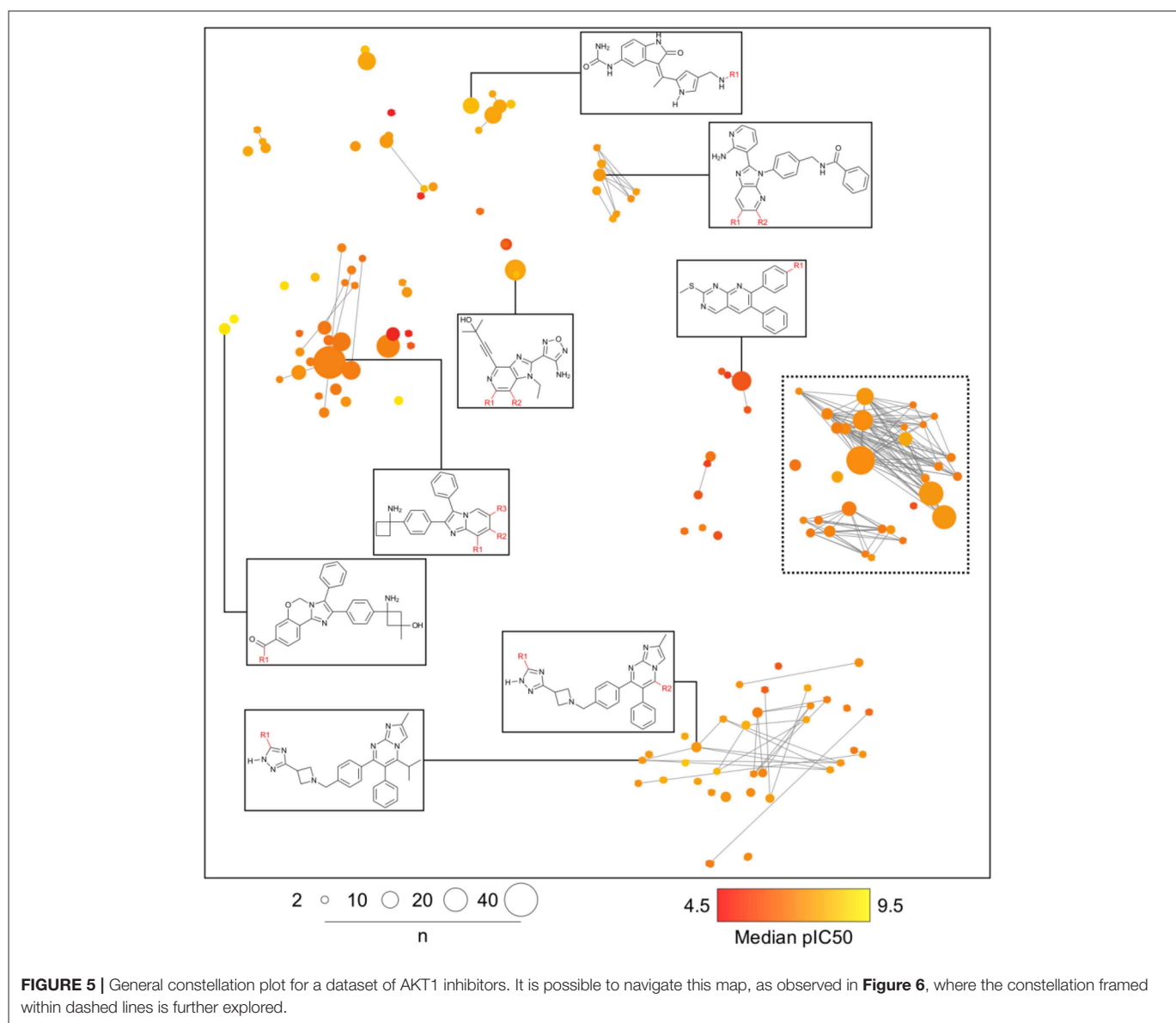
The construction of constellations plots and exemplary applications are illustrated with two case studies of general

interest in drug discovery. As mentioned in the section Methods, the first example consists of a dataset of 827 AKT1 inhibitors obtained from ChEMBL (Gaulton et al., 2017) and cheminformatically described in Naveja et al. (2018b). The second example employs a data set of 286 DNA methyltransferase (DNMT) inhibitors obtained from the integration of several databases as described in Naveja and Medina-Franco (2018).

### Case Study 1: AKT1 Inhibitors

Analogues in this library could be summarized in 144 cores as discussed in the section Methods. The cores were organized in 79 analog series and contained 440 compounds (about half of the

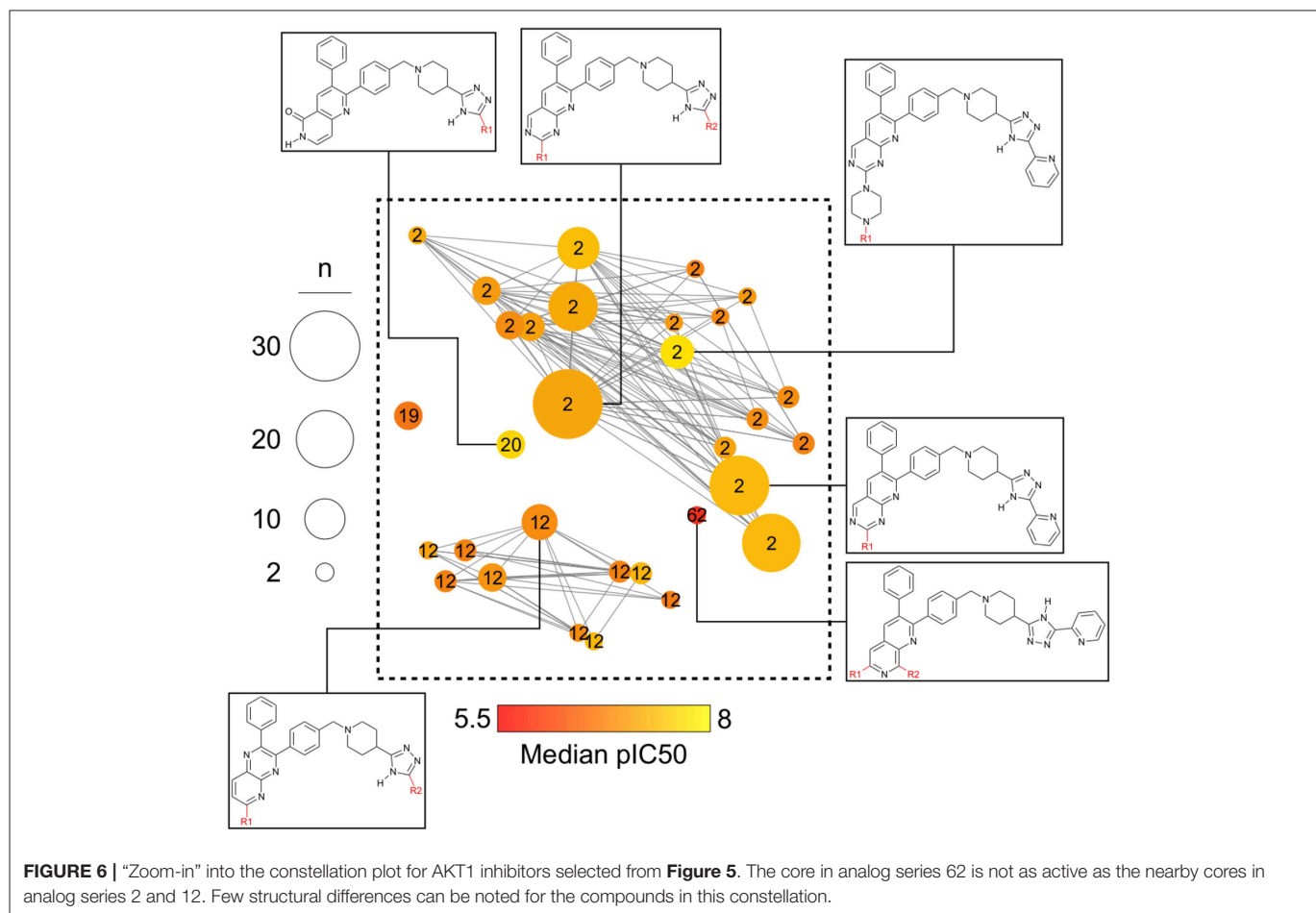




initial dataset). **Figure 5** is the constellation plot for these data, where it becomes apparent that chemical space and chemical substructure information play simultaneous roles in describing the SAR. For instance, although some inactive cores are close to active cores in chemical space, they are not usually contained in the same analog series. Therefore, these could be categorized as “scaffold cliffs” rather than simple activity cliffs conceptualized as two small molecules with similar structures and very different activities (Maggiore, 2006). In this case, collections of molecules, rather than single molecules, are being compared.

**Figure 6** is a zoomed-in picture into a single “bright” (or predominantly active) constellation comprising five analog series and 55 compounds. As it is readily observed, analog series close in the chemical space have only slight dissimilarities within their scaffolds; in this case, they all share a naphthyridine or naphthyridinone scaffold. Constellation plots allow for a more

precise visual SAR analysis and generation of hypotheses. For instance, the core associated to analog series 62 has only a different position for the nitrogens in the rings as well as where substitutions occur. Structural studies could then be conducted to elucidate which are the most relevant features for this kind of scaffolds to be active against AKT1. In this regard, a recent publication co-crystallizing 1,6-naphthyridinone derivatives similar to those in analog series 20 has shown that this scaffold is relevant in forming a  $\pi$ - $\pi$  stacking interaction with the side chain of Trp80 of the PH-domain (Uhlenbrock et al., 2019). Nonetheless, variation of the position of nitrogen atoms in the scaffold were not considered in the cited study. Indeed, previous SAR studies of these analogs have found the position of the nitrogen atoms in these scaffolds to be critical for the activity against AKT (Zhao et al., 2005; Bilodeau et al., 2008).



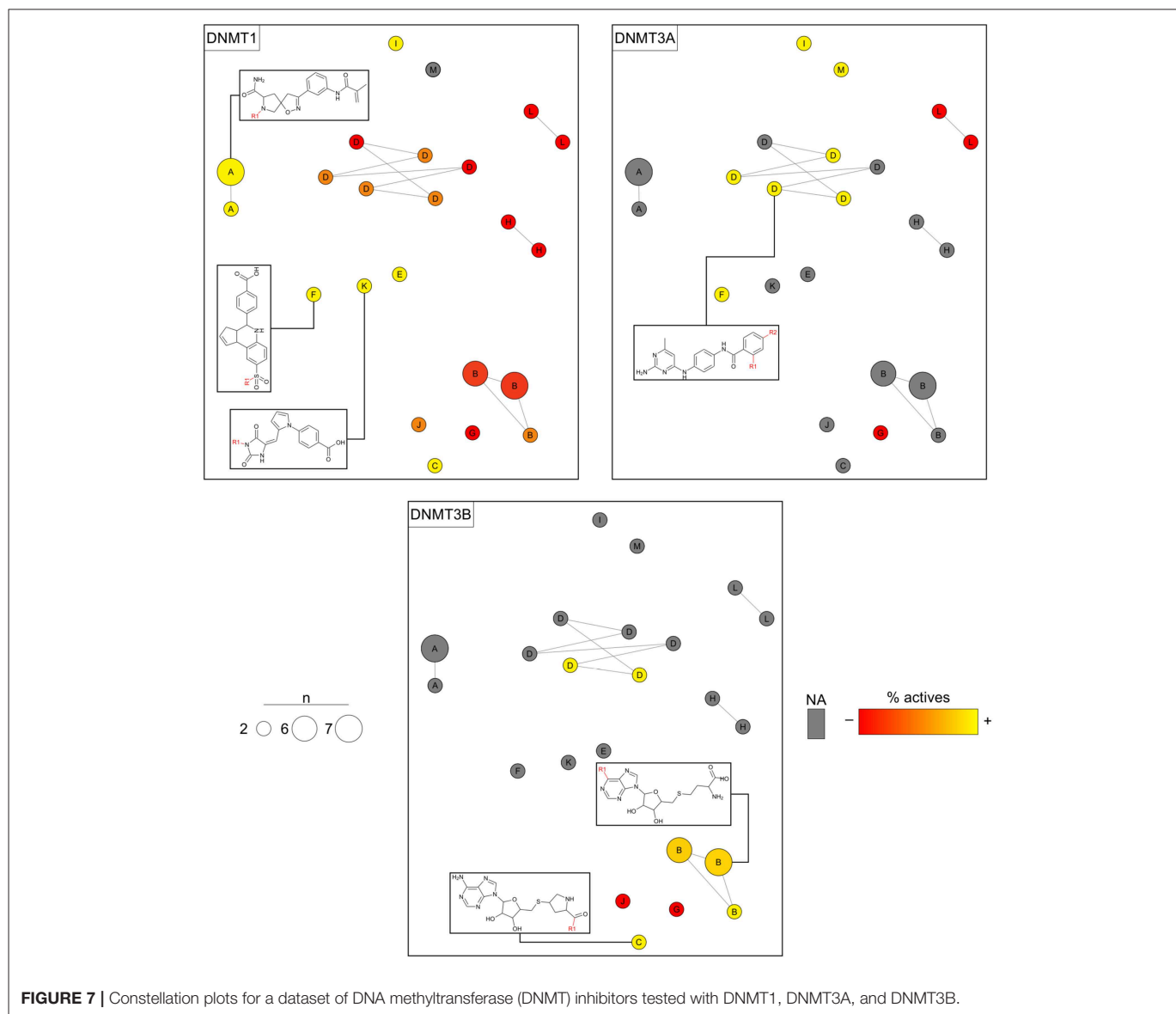
## Case Study 2: DNMT Inhibitors

Analogs in this library could be summarized in 23 cores following the procedure discussed in the section Methods. The cores were organized in 13 analog series and contained 46 compounds (about 16% of the initial dataset). Compounds in this library have annotated activity with DNMT1, DNMT3A, and/or DNMT3B. **Figure 7** shows three constellations plots, where chemical space is the same and colors change to represent the activities against each DNMT. As elaborated on the section Methods, each circle in the plot represents a core in which coordinates in the 2D graph is given by similarity measurements computed from Morgan fingerprints using *t*-SNE for dimensionality reduction. Labels indicate the analog series to which cores belong. The color represents the percentage of active compounds sharing that core using a continuous color scale from red (less active cores) to yellow (more active cores). For this example of use of the constellation plots, the definition of “active” was determined from integrating qualitative and quantitative data sources as described in Naveja and Medina-Franco (2018). Circles in gray indicate cores with no reported activity for that particular DNMT. The size of the circle indicates the number of compounds sharing the core. Connected circles are cores sharing compounds. **Figure 7** also shows the chemical structures of representative cores.

The constellation plots for DNMT inhibitors in **Figure 7** allow for rapidly getting several interesting insights of the SAR. For instance, cores at the top left part of the plot from analog series “A” are a bright constellation against DNMT1, i.e., a region in chemical space with active analogs. However, these analogs have not been tested against the other DNMT isoenzymes, which would help determine whether these inhibitors are selective.

Of note, there is a “dark” (or predominantly inactive) constellation in the chemical space of DNMT1 formed by six cores from analog series “D.” This dark constellation, however, is more active overall against DNMT3A and appears to be active against DNMT3B. Furthermore, not all cores in this constellation have been tested against DNMT3A and DNMT3B, where they have greater chances of being active.

The plot also reveals a constellation of nucleoside analogs from series “B” at the bottom-right region of the plot that is, overall, selective toward DNMT3B vs. DNMT1. This series has not been tested against DNMT3A yet. Moreover, most of the cores have been tested in DNMT1 only, thus hindering discussions on selectivity. In this regard, analysis of constellation plots is visually helpful in guiding multitarget drug discovery campaigns and in finding opportunities for selectivity.



## CONCLUSIONS AND PERSPECTIVES

We introduced a novel approach for combining chemical space and analog series methodologies into a single descriptive analysis that can be summarized in a constellation plot. Adding the analog series concept into the chemical space facilitates discussions of regions in the space, as every point summarizes a collection of analogs. A so-called “constellation in chemical space” can be conceptualized as those regions in chemical space formed by core scaffolds with similar structure (as defined by a coordinate-based projection). Mapping activity on the plot readily uncovers active and inactive zones, e.g., bright or dark regions, in chemical space. Of note, constellation plots would be useful for exploring virtually any chemical property, such as biological activity (as demonstrated with two case studies), but also physicochemical properties, complexity or selectivity statistics. In this regard, constellation plots are a flexible approach with multiple potential

applications in academia and industry, aiding in the quest of finding potential leads from large collections of screening data (e.g., such as that produced by high-throughput screening campaigns). One of the next steps of this work is the application of the constellations plots to navigate through cell selectivity data of a comprehensive screening dataset. Results will be disclosed in due course.

## DATA AVAILABILITY

The datasets generated for this study can be found in **Supplementary Information**.

## AUTHOR CONTRIBUTIONS

JN participated in the conceptualization, data gathering, data analysis, and drafted the first

version of the manuscript. JM-F participated in the conceptualization, data analysis, and revision of the manuscript.

## ACKNOWLEDGMENTS

JN was thankful to CONACyT for the granted scholarship number 622969. We also thank Consejo

Nacional de Ciencia y Tecnología (CONACyT) for grant 282785.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fchem.2019.00510/full#supplementary-material>

## REFERENCES

- Bajorath, J. (2018). Improving the utility of molecular scaffolds for medicinal and computational chemistry. *Future Med. Chem.* 10, 1645–1648. doi: 10.4155/fmc-2018-0106
- Bemis, G. W., and Murcko, M. A. (1996). The properties of known drugs. 1. Molecular frameworks. *J. Med. Chem.* 39, 2887–2893. doi: 10.1021/jm9602928
- Bilodeau, M. T., Balitz, A. E., Hoffman, J. M., Manley, P. J., Barnett, S. F., Defeo-Jones, D., et al. (2008). Allosteric inhibitors of Akt1 and Akt2: a naphthyridinone with efficacy in an A2780 tumor xenograft model. *Bioorg. Med. Chem. Lett.* 18, 3178–3182. doi: 10.1016/j.bmcl.2008.04.074
- Borrel, A., Kleinstreuer, N. C., and Fourches, D. (2018). Exploring drug space with ChemMaps.com *Bioinformatics* 34, 3773–3775. doi: 10.1093/bioinformatics/bty412
- Clayden, J., Greeves, N., and Warren, S. (2012). *Organic Chemistry*. Oxford, UK: Oxford University Press. Available online at: <http://global.oup.com/ukhe/product/organic-chemistry-9780199270293?cc=mx&lang=en&>
- Dimova, D., and Bajorath, J. (2018). Collection of analog series-based scaffolds from public compound sources. *Future Sci. OA* 4:FSO287. doi: 10.4155/fsoa-2017-0135
- Dimova, D., Stumpfe, D., Hu, Y., and Bajorath, J. (2016). Analog series-based scaffolds: computational design and exploration of a new type of molecular scaffolds for medicinal chemistry. *Future Sci. OA* 2:FSO149. doi: 10.4155/fsoa-2016-0058
- Gaulton, A., Hersey, A., Nowotka, M., Bento, A. P., Chambers, J., Mendez, D., et al. (2017). The ChEMBL database in 2017. *Nucleic Acids Res.* 45, D945–D954. doi: 10.1093/nar/gkw1074
- González-Medina, M., Prieto-Martínez, F. D., Naveja, J. J., Méndez-Lucio, O., El-Elmat, T., Pearce, C. J., et al. (2016). Chemoinformatic expedition of the chemical space of fungal products. *Future Med. Chem.* 8, 1399–1412. doi: 10.4155/fmc-2016-0079
- Griffen, E., Leach, A. G., Robb, G. R., and Warner, D. J. (2011). Matched molecular pairs as a medicinal chemistry tool. *J. Med. Chem.* 54, 7739–7750. doi: 10.1021/jm200452d
- Hu, Y., Stumpfe, D., and Bajorath, J. (2011). Lessons learned from molecular scaffold analysis. *J. Chem. Inf. Model.* 51, 1742–1753. doi: 10.1021/ci200179y
- Kunimoto, R., Dimova, D., and Bajorath, J. (2017). Application of a new scaffold concept for computational target deconvolution of chemical cancer cell line screens. *ACS Omega* 2, 1463–1468. doi: 10.1021/acsomega.7b00215
- Kunimoto, R., Miyao, T., and Bajorath, J. (2018). Computational method for estimating progression saturation of analog series. *RSC Adv.* 8, 5484–5492. doi: 10.1039/C7RA13748F
- Larsson, J., Gottfries, J., Muresan, S., and Backlund, A. (2007). ChemGPS-NP: tuned for navigation in biologically relevant chemical space. *J. Nat. Prod.* 70, 789–794. doi: 10.1021/np070002y
- Lewis, R., Guha, R., Korcsmaros, T., and Bender, A. (2015). Synergy maps: exploring compound combinations using network-based visualization. *J. Cheminform.* 7, 36. doi: 10.1186/s13321-015-0090-6
- Lipinski, C., and Hopkins, A. (2004). Navigating chemical space for biology and medicine. *Nature* 432, 855–861. doi: 10.1038/nature03193
- Lyu, J., Wang, S., Balius, T. E., Singh, I., Levit, A., Moroz, Y. S., et al. (2019). Ultra-large library docking for discovering new chemotypes. *Nature* 566, 224–229. doi: 10.1038/s41586-019-0917-9
- Maaten, L., and Hinton, G. (2008). Visualizing data using t-SNE. *J. Mach. Learn. Res.* 9, 2579–2605. Available online at: <http://www.jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf>
- Maggiora, G. M. (2006). On outliers and activity cliffs—why QSAR often disappoints. *J. Chem. Inf. Model.* 46, 1535. doi: 10.1021/ci060117s
- Medina-Franco, J., Martínez-Mayorga, K., Giulianotti, M., Houghten, R., and Pinilla, C. (2008). Visualization of the chemical space in drug discovery. *CAD* 4, 322–333. doi: 10.2174/157340908786786010
- Naveja, J. J., Cortés-Benítez, F., Bratoeff, E., and Medina-Franco, J. L. (2016). Activity landscape analysis of novel 5 $\alpha$ -reductase inhibitors. *Mol. Divers.* 20, 771–780. doi: 10.1007/s11030-016-9659-x
- Naveja, J. J., and Medina-Franco, J. L. (2015). Activity landscape sweeping: insights into the mechanism of inhibition and optimization of DNMT1 inhibitors. *RSC Adv.* 5, 63882–63895. doi: 10.1039/C5RA12339A
- Naveja, J. J., and Medina-Franco, J. L. (2017). ChemMaps: Towards an approach for visualizing the chemical space based on adaptive satellite compounds. *F1000Res.* 6:Chem Inf Sci-1134. doi: 10.12688/f1000research.12095.2
- Naveja, J. J., and Medina-Franco, J. L. (2018). Insights from pharmacological similarity of epigenetic targets in epipolypharmacology. *Drug Discov. Today* 23, 141–150. doi: 10.1016/j.drudis.2017.10.006
- Naveja, J. J., Norinder, U., Mucs, D., López-López, E., and Medina-Franco, J. L. (2018a). Chemical space, diversity and activity landscape analysis of estrogen receptor binders. *RSC Adv.* 8, 38229–38237. doi: 10.1039/C8RA07604A
- Naveja, J. J., Oviedo-Osornio, C. I., Trujillo-Minero, N. N., and Medina-Franco, J. L. (2018b). Chemoinformatics: a perspective from an academic setting in Latin America. *Mol. Divers.* 22, 247–258. doi: 10.1007/s11030-017-9802-3
- Naveja, J. J., Vogt, M., Stumpfe, D., Medina-Franco, J. L., and Bajorath, J. (2019). Systematic extraction of analogue series from large compound collections using a new computational compound–core relationship method. *ACS Omega* 4, 1027–1032. doi: 10.1021/acsomega.8b03390
- Oprea, T. I., and Gottfries, J. (2001). Chemography: the art of navigating in chemical space. *J. Comb. Chem.* 3, 157–166. doi: 10.1021/cc0000388
- Osolodkin, D. I., Radchenko, E. V., Orlov, A. A., Voronkov, A. E., Palyulin, V. A., and Zefirov, N. S. (2015). Progress in visual representations of chemical space. *Expert Opin. Drug Discov.* 10, 959–973. doi: 10.1517/17460441.2015.1060216
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* 12, 2825–2830. Available online at: <http://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf>
- Prieto-Martínez, F. D., Gortari, E. F., Méndez-Lucio, O., and Medina-Franco, J. L. (2016). A chemical space odyssey of inhibitors of histone deacetylases and bromodomains. *RSC Adv.* 6, 56225–56239. doi: 10.1039/C6RA07224K
- Probst, D., and Reymond, J.-L. (2018). Exploring drugbank in virtual reality chemical space. *J. Chem. Inf. Model.* 58, 1731–1735. doi: 10.1021/acs.jcim.8b00402
- Reymond, J.-L. (2015). The chemical space project. *Acc. Chem. Res.* 48, 722–730. doi: 10.1021/ar500432k
- Rosén, J., Lövgren, A., Kogej, T., Muresan, S., Gottfries, J., and Backlund, A. (2009). ChemGPS-NP(Web): chemical space navigation online. *J. Comput. Aided Mol. Des.* 23, 253–259. doi: 10.1007/s10822-008-9255-y
- Stumpfe, D., Dimova, D., and Bajorath, J. (2016). Computational method for the systematic identification of analog series and key compounds representing series and their biological activity profiles. *J. Med. Chem.* 59, 7667–7676. doi: 10.1021/acs.jmedchem.6b00906

- Uhlenbrock, N., Smith, S., Weisner, J., Landel, I., Lindemann, M., Le, T. A., et al. (2019). Structural and chemical insights into the covalent-allosteric inhibition of the protein kinase Akt. *Chem. Sci.* 10, 3573–3585. doi: 10.1039/c8sc05212c
- Vogt, M., Yonchev, D., and Bajorath, J. (2018). Computational method to evaluate progress in lead optimization. *J. Med. Chem.* 61, 10895–10900. doi: 10.1021/acs.jmedchem.8b01626
- Waddell, J., and Medina-Franco, J. L. (2012). Bioactivity landscape modeling: chemoinformatic characterization of structure-activity relationships of compounds tested across multiple targets. *Bioorg. Med. Chem.* 20, 5443–5452. doi: 10.1016/j.bmc.2011.11.051
- Yonchev, D., Vogt, M., Stumpfe, D., Kunitomo, R., Miyao, T., and Bajorath, J. (2018). Computational assessment of chemical saturation of analogue series under varying conditions. *ACS Omega* 3, 15799–15808. doi: 10.1021/acsomega.8b02087
- Zhao, Z., Leister, W. H., Robinson, R. G., Barnett, S. F., Defeo-Jones, D., Jones, R. E., et al. (2005). Discovery of 2,3,5-trisubstituted pyridine derivatives as potent Akt1 and Akt2 dual inhibitors. *Bioorg. Med. Chem. Lett.* 15, 905–909. doi: 10.1016/j.bmcl.2004.12.062

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

Copyright © 2019 Naveja and Medina-Franco. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



## RESEARCH NOTE

**REVISED** ChemMaps: Towards an approach for visualizing the chemical space based on adaptive satellite compounds [version 2; referees: 3 approved with reservations]

J. Jesús Naveja<sup>1,2</sup>, José L. Medina-Franco<sup>1</sup>

<sup>1</sup>Department of Pharmacy, School of Chemistry, Universidad Nacional Autónoma de México, Mexico City, 04510, Mexico

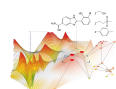
<sup>2</sup>PECEM, Faculty of Medicine, Universidad Nacional Autónoma de México, Mexico City, 04510, Mexico

**v2** First published: 17 Jul 2017, 6(CHEM INF SCI):1134 (doi: 10.12688/f1000research.12095.1)

Latest published: 04 Aug 2017, 6(CHEM INF SCI):1134 (doi: 10.12688/f1000research.12095.2)

### Abstract

We present a novel approach called ChemMaps for visualizing chemical space based on the similarity matrix of compound datasets generated with molecular fingerprints' similarity. The method uses a 'satellites' approach, where satellites are, in principle, molecules whose similarity to the rest of the molecules in the database provides sufficient information for generating a visualization of the chemical space. Such an approach could help make chemical space visualizations more efficient. We hereby describe a proof-of-principle application of the method to various databases that have different diversity measures. Unsurprisingly, we found the method works better with databases that have low 2D diversity. 3D diversity played a secondary role, although it seems to be more relevant as 2D diversity increases. For less diverse datasets, taking as few as 25% satellites seems to be sufficient for a fair depiction of the chemical space. We propose to iteratively increase the satellites number by a factor of 5% relative to the whole database, and stop when the new and the prior chemical space correlate highly. This Research Note represents a first exploratory step, prior to the full application of this method for several datasets.






This article is included in the **Chemical Information Science gateway**.

### Open Peer Review

Referee Status: ? ? ?

	Invited Referees		
	1	2	3
<b>REVISED</b>	?	?	?
<b>version 2</b>	report	report	report
published 04 Aug 2017	↑	↑	↑
<b>version 1</b>	?	?	✓
published 17 Jul 2017	report	report	report

- Gerald Maggiora** , University of Arizona, USA
- Dmitry I. Osolodkin** , Chumakov FSC R&D IBP RAS, Russian Federation Lomonosov Moscow State University, Russian Federation
- Jean-Louis Reymond** , University of Bern, Switzerland

### Discuss this article

Comments (2)



**Corresponding author:** José L. Medina-Franco ([jose.medina.franco@gmail.com](mailto:jose.medina.franco@gmail.com))

**Author roles:** **Naveja JJ:** Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – Original Draft Preparation; **Medina-Franco JL:** Conceptualization, Funding Acquisition, Methodology, Project Administration, Supervision, Writing – Review & Editing

**Competing interests:** No competing interests were disclosed.

**How to cite this article:** Naveja JJ and Medina-Franco JL. **ChemMaps: Towards an approach for visualizing the chemical space based on adaptive satellite compounds [version 2; referees: 3 approved with reservations]** *F1000Research* 2017, **6**(Chem Inf Sci):1134 (doi: [10.12688/f1000research.12095.2](https://doi.org/10.12688/f1000research.12095.2))

**Copyright:** © 2017 Naveja JJ and Medina-Franco JL. This is an open access article distributed under the terms of the [Creative Commons Attribution Licence](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. Data associated with the article are available under the terms of the [Creative Commons Zero "No rights reserved" data waiver](#) (CC0 1.0 Public domain dedication).

**Grant information:** Consejo Nacional de Tecnología (CONACyT) scholarship 622969 (JJN). Universidad Nacional Autónoma de México (UNAM), Programa de Apoyo a la Investigación y el Posgrado PAIP, grant 5000-9163 (JLMF) and Programa de Apoyo a Proyectos de Investigación e Innovación Tecnológica PAPIIT, grant IA204016 (JLMF).

*The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.*

**First published:** 17 Jul 2017, **6**(Chem Inf Sci):1134 (doi: [10.12688/f1000research.12095.1](https://doi.org/10.12688/f1000research.12095.1))

**REVISED Amendments from Version 1**

We discuss further in the Introduction, the differences of ChemMaps with other similar approaches.

We updated the [Figure 1–Figure 3](#) for better visibility. [Dataset 1](#) has been updated to also contain HDAC1 compounds used in the study.

We have expanded the perspectives of the work in the Conclusion.

The [Supplementary File](#) has been updated with [Supplementary Methods](#), [Supplementary Results](#) and [Table S1](#), containing the curation of the database and PCA details. [Supplementary Figure S1–Supplementary Figure S4](#) have been revised, and we added a new [Supplementary Figure 5](#) comparing the variance percentage contribution of the PCs for each studied database.

**See referee reports**

## Introduction

Visual representation of chemical space has multiple implications in drug discovery for virtual screening, library design and comparison of compound collections, among others<sup>1</sup>. Amongst the multiple methods to explore chemical space, principal component analysis (PCA) of pairwise similarity matrices computed with structural fingerprints has been used to analyze compound datasets<sup>2,3</sup>. A drawback of this approach is that it becomes impractical for large libraries due to the large dimension of the similarity matrix<sup>4</sup>. Other approaches use molecular representations different from structural fingerprints, such as physicochemical properties or complexity descriptors, or methods different from PCA, such as multidimensional-scaling and neural networks<sup>5,6</sup>.

In representation of the chemical space based on PCA there have been “chemical satellite” approaches, such as ChemGPS, which select satellites molecules that might not be included in the database to visualize, but have extreme features that place them as outliers, with the intention to reach as much of the chemical space as possible<sup>7–10</sup>. Also, a related and more recent approach, Similarity Mapplet, makes possible the visualization of very large chemical libraries, by considering PCA of different molecular features, including structural<sup>11</sup>.

Although we concur with the fact that not all compounds in a compound data set should be necessary to generate a meaningful chemical space, there are still obvious limitations of using a fixed set of satellites to which the user is blinded. Also, until now there was no proposal of such a method based on structural similarity.

We therefore suggest the hybrid approach, ChemMaps, in which a portion of the database to be represented is used as satellite, thereby decreasing the computational effort required to compute the similarity matrix without losing adaptability of the method to any particular database. Since it is expected that more diverse sets would require more satellites, a second goal of this study was to qualitatively explore the relationship between the internal diversity of compound datasets and the fraction of compounds required as satellites, in order to generate a good approximation of the chemical space.

## Methods

[Table 1](#) summarizes the six compound data sets considered in this study. Note that small median similarity values imply higher diversity. The datasets were selected from a large scale study of profiling epigenetic datasets (unpublished study, Naveja JJ and Medina-Franco JL) with relevance in epigenetic-drug discovery. We also included DrugBank as a control diverse dataset<sup>12</sup>. Briefly, we selected focused libraries of inhibitors of DNMT1 (a DNA-methyltransferase; library diverse 2D and 3D), L3MBTL3 (a histone methylation reader; diverse 3D and less diverse 2D), SMARCA2 (a chromatin remodeller; diverse 2D, less diverse 3D), and CREBBP (a histone acetyltransferase; less diverse both 2D and 3D). Datasets were selected based on their different internal diversity (as measured with Tanimoto index/MACCS keys for 2D measurements and Tanimoto combo/OMEGA-ROCS for 3D; see [Figure S1](#) in [Supplementary File 1](#)). Data sets in this work have approximately the same number of compounds except for HDAC1 and DrugBank, which were selected to benchmark the method in larger databases ([Table 2](#)). We evaluated 2D diversity using the median of Tanimoto/MACCS similarity measures in KNIME version 3.3.2, and 3D diversity using the median of Combo Score from the ROCS, version 3.2.2 and OMEGA, version 2.5.1, OpenEye software<sup>13–16</sup>.

**Table 1. Compound data sets used in the study.**

Dataset	Description	Size	2D similarity <sup>a</sup>	2D similarity <sup>b</sup>	3D similarity <sup>c</sup>
DNMT1 inhibitors	DNA-methyltransferase	244	0.44	0.12	0.16
SMARCA2 inhibitors	Chromatin remodeller	220	0.51	0.15	0.23
CREBBP inhibitors	Histone acetyltransferase	178	0.67	0.22	0.16
L3MBTL3 inhibitors	Histone methylation reader	115	0.77	0.41	0.03
HDAC1 inhibitors	Histone acetyltransferase	3,257	0.49	0.16	0.12
DrugBank	Approved drugs	1,900	0.35	NC	NC

<sup>a</sup>Median of Tanimoto/MACCS similarity; <sup>b</sup>Median of Tanimoto/ECFP4 similarity; <sup>c</sup>Median of OMEGA-ROCS similarity; NC: not calculated

**Table 2. Benchmark with larger databases.**

Database	Gold standard timing (s)	Satellites timing (s)	Correlation
DrugBank	162	147	0.92
HDAC1	406	287	0.99

To assess the hypothesis of this work we performed two main approaches A) *Backwards approach*: start with computing the full similarity matrix of each data set and remove compounds systematically; and B) *Forward approach*: start adding compounds to the similarity matrix until finding the reduced number of required compounds (called 'satellites') to reach a visualization of the chemical space that is very similar to computing the full similarity matrix. The second approach would be the usual and realistic approach from a user standpoint. Each method is further detailed in the next two subsections.

### Backwards approach

The following steps were implemented in an automated workflow in KNIME, version 3.3.2<sup>17</sup>:

1. For each compound in the dataset with  $N$  compounds, generate the  $N \times N$  similarity matrix using Tanimoto/extended connectivity fingerprints radius 4 (ECFP4) generated with CDK KNIME nodes.
2. Perform PCA of the similarity matrix generated in step 1 and selected the first 2 or 3 principal components (PCs).
3. Compute all pair-wise Euclidean distances based on the scores of the 2 or 3 PCs generated in step 2. The set of distances are later used as reference or 'gold standard'. It should be noted that the "real" distances or true gold standard would consider the whole distance matrix. However, for visualization purposes it is unfeasible to render more than 3 dimensions. Therefore, we selected as reference the best 2D or 3D visualization possible by means of PCA.
4. Repeat steps 1 to 3 with one compound as satellite, generating an  $N \times 1$  similarity matrix. The first compound was selected randomly. In this case, for example, it is only possible to calculate one PC, but as the number of satellites increases, we can again compute 2 or 3 PCs.
5. Calculate the correlation among the pairwise distances generated in step 2 obtained using the whole matrix (e.g., *gold standard*) and those obtained in step 4.
6. Iterate over steps 4 and 5 increasing the number of satellites one by one until  $N - 1$  satellites are reached. To select the second, third, etc. compounds, two approaches were followed: select compounds at random and select compounds with the largest diversity to the previously selected (i.e., Max-Min approach).
7. Estimate the proportion of satellite compounds required to preserve a 'high' (of at least 0.9) correlation.

8. The prior steps were repeated five times for each dataset in order to capture the stability of the method.

### Forward approach

The former approach is useful only for validation purposes of the methodology as a proof-of-principle. However, the obvious objective of a satellite-approach is to avoid the calculation of the complete similarity matrix e.g., step 1 in backwards approach. To this end, we developed a satellite-adding or forward approach, in contrast with the formerly introduced backwards approach. We started with 25% of the database as satellites and for each iteration we added 5% until the correlation of the pairwise Euclidean distances remains high (at least 0.9). A further description of the methods for standardizing the chemical data and integrating the dataset can be found in the Supplementary material, as well as a further description of the PCA analysis used.

**Dataset 1. This file contains the six compound datasets used in this work in SDF format**

<http://dx.doi.org/10.5256/f1000research.12095.d171632>

No special software is required to open the SDF files. Any commercial or free software capable of reading SDF files will open the data sets supplied.

## Results

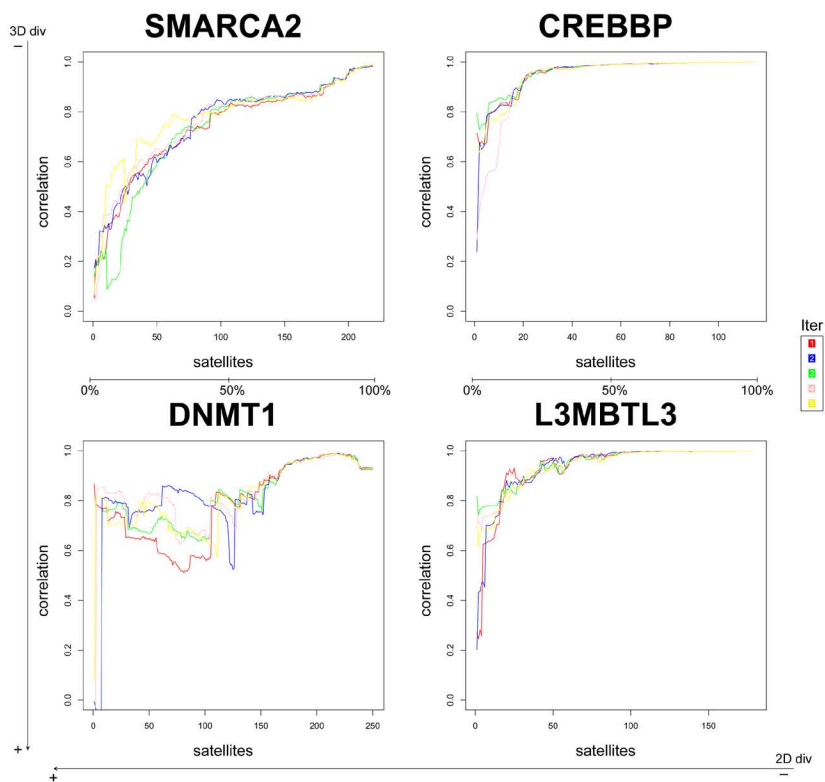
### Backwards approach

In this pilot study, we assessed a few variables to tune up the method, such as the number of PCs used (2 or 3) and the selection of satellites at random or by diversity. We found that selection at random is more stable, above all in less diverse datasets (Figure 1 and Figure 2; Figure S2 and Figure S3). Likewise, selecting 2 PCs the performance is slightly better and more stable (compare Figure 1 and Figure 2 against Figure S2 and Figure S3).

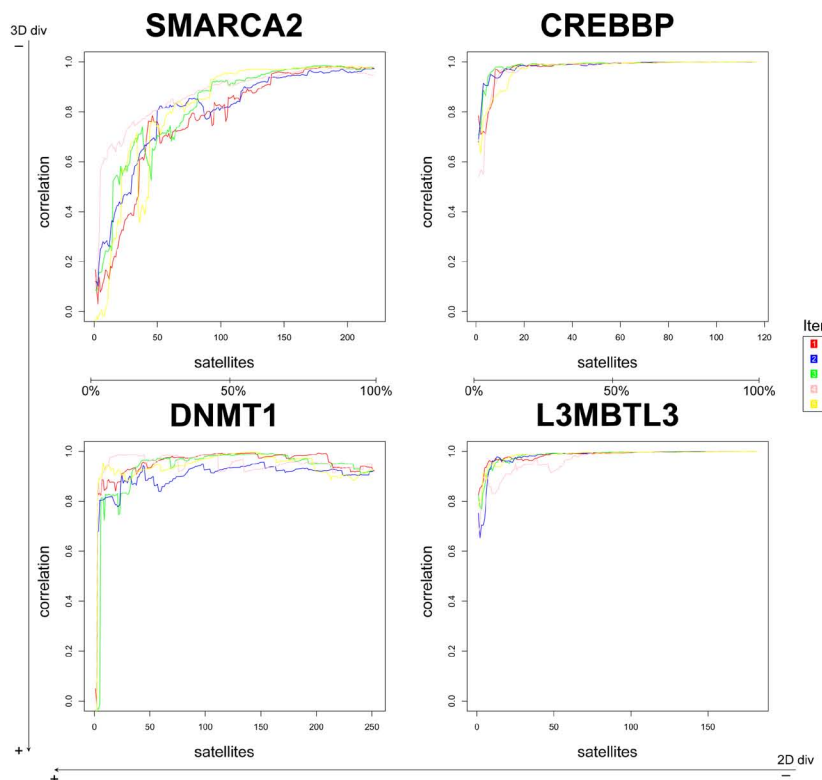
Therefore, from this point onwards we will focus on the results of the at random satellites selection and using 2 PCs (Figure 2). From the four datasets, we conclude that for datasets with lower 2D diversity (CREBBP and L3MBTL3, see Table 1), around 25% of satellite compounds are enough to obtain a high correlation ( $\geq 0.9$ ) with the gold standard (e.g., PCA on the whole matrix), whereas for 2D-diverse datasets i.e., DNMT1 and SMARCA2, up to 75% of the compounds could be needed to ensure a high correlation. Nonetheless, even for these datasets, using 25% of the compounds as satellites the correlation with the gold standard is already between 0.6 and 0.8; using 50% of the compounds as satellites the correlation is between 0.7 and 0.9. Hence, the higher the diversity of a dataset (especially 2D), the higher the number of satellites required.

### Forward approach

Evidently, a useful method for reducing computing time and disk space usage should not use the PCA on the whole similarity matrix



**Figure 1. Backwards analysis with 2PCs picking satellites by diversity.** The correlation with the results from the whole matrix was calculated with increasing numbers of satellites. Each colored line represents one of the five iterations.



**Figure 2. Backwards analysis with 2PCs picking satellites at random.** The correlation with the results from the whole matrix was calculated with increasing numbers of satellites. Each colored line represents one of the five iterations.

to determine an adequate number of satellites for each dataset. With that in mind, we decided to design a method that starts with a given percentage of the database as satellites, and then keeps adding a proportion of them until the correlation between the former and the updated data is of at least 0.9. In Figure 3 we depict this approach on the same databases in Table 1 for step sizes of 5% and starting from zero. Similarly as what we saw in the backwards method, around 5 steps (25% of the database) are usually necessary to reach a stable, high correlation between steps. Figure S4 shows that for step sizes of 10% there is no further improvement. Therefore we suggest that the method should, for default, start with 25% of compounds as satellites and then keep adding 5% until a correlation between steps of at least 0.9 is reached.

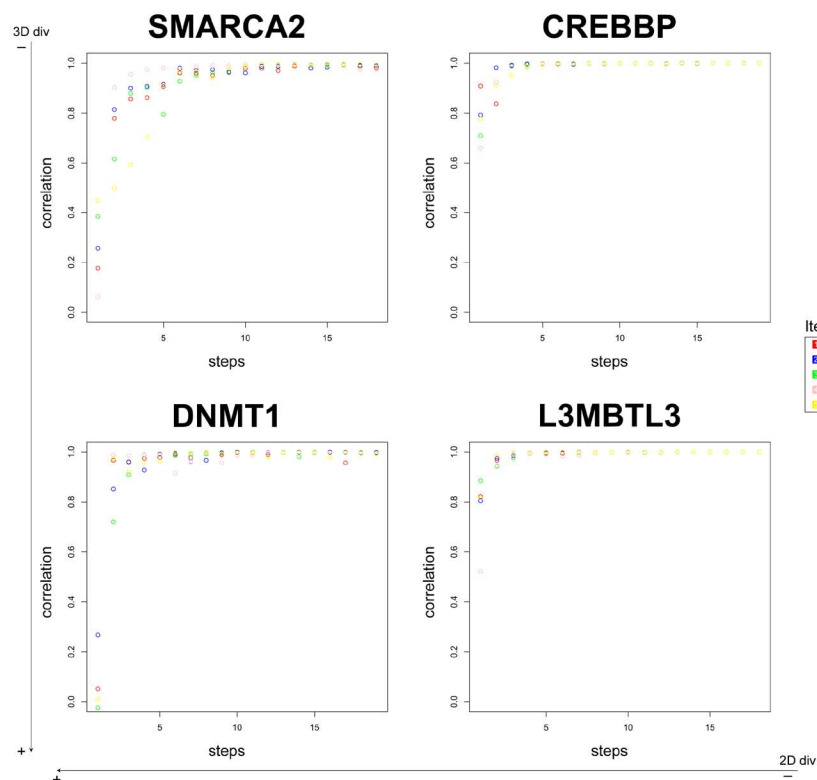
### Application

In this pilot study we applied the ChemMaps method to visualize the chemical space of two larger datasets (HDAC1 and DrugBank with 3,257 and 1,900 compounds, respectively, Table 1). As shown in Table 2, a significant reduction in time performance was achieved as compared to the gold standard, and the correlation between

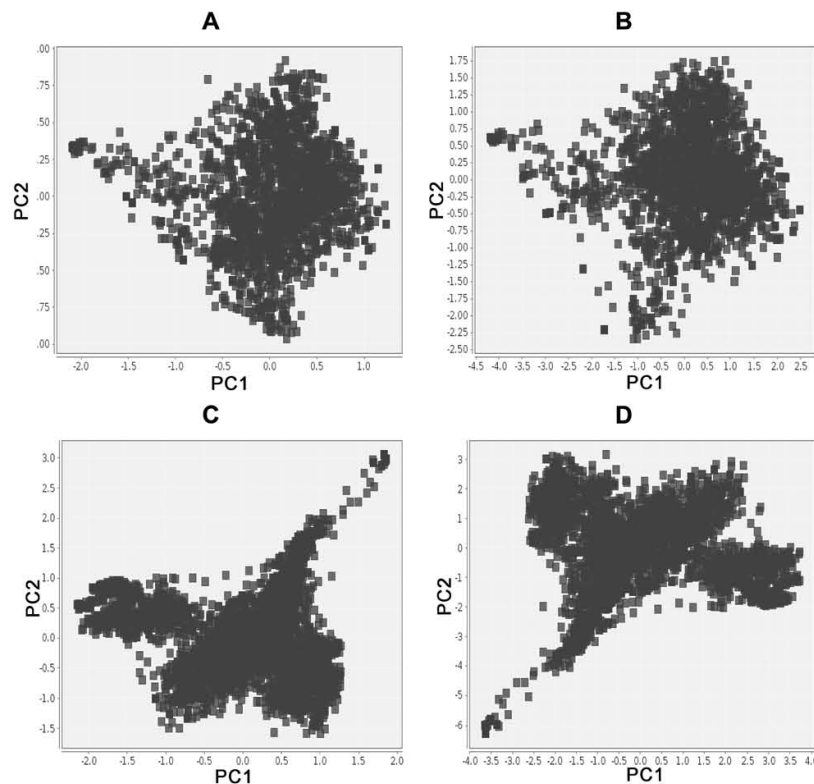
the gold standard and the satellites approach was in both cases higher than 0.9. Figure 4 depicts the chemical spaces generated in both instances. Although the orientation of the map changed for HDAC1, the shape and distances remain quite similar, which is the main objective. This preliminary work supports the hypothesis that a reduced number of compounds is sufficient to generate a visual representation of the chemical space (based on PCA of the similarity matrix) that is quite similar to the chemical space of the PCA of the full similarity matrix.

### Conclusion and future directions

This proof-of-concept study suggests that using the adaptive satellite compounds ChemMaps is a plausible approach to generate a reliable visual representation of the chemical space based on PCA of similarity matrices. The approach works better for relatively less-diverse datasets, although it seems to remain robust when applied to more diverse datasets. For datasets with small diversity, fewer satellites seem to be enough to produce a representative visual representation of the chemical space. The higher relevance of 2D diversity over 3D in this study could be importantly related to the fact that the



**Figure 3.** Forward analysis with 2PCs picking satellites at random step sizes of 5%.



**Figure 4.** Chemical space of DrugBank using (A) the adaptive satellites approach or (B) the gold standard. As well as for HDAC1 using (C) the adaptive satellites approach or (D) the gold standard.

chemical space depiction is based on 2D fingerprints. Therefore, the performance of the methods depicting the chemical space based on 3D fingerprints could also be assessed.

A major next step is to conduct a full benchmark study to assess the general applicability of the approach proposed herein, and also in larger databases, in which we anticipate this method would be even more useful. A second step is to propose a metric that determines the number of compounds required as satellites for PCA representation of the chemical space based on similarity matrices. As well, it is pending the development of quantitative metrics for assessing the stability of the satellites selection and thus conclusively establish the superiority of at random satellite selection. Finally, a more comprehensive and in-depth study of this new methodology should be addressed, in order to further characterise its applicability domain, including a dataset diversity threshold above which the confiability of the approach decreases.

#### Data availability

**Dataset 1.** This file contains the six compound datasets used in this work in SDF format. No special software is required to open the SDF files. Any commercial or free software capable of

reading SDF files will open the data sets supplied. <http://dx.doi.org/10.5256/f1000research.12095.d171632><sup>18</sup>

#### Competing interests

No competing interests were disclosed.

#### Grant information

Consejo Nacional de Tecnología (CONACyT) scholarship 622969 (JJN). Universidad Nacional Autónoma de México (UNAM), *Programa de Apoyo a la Investigación y el Posgrado PAIP*, grant 5000-9163 (JLMF) and *Programa de Apoyo a Proyectos de Investigación e Innovación Tecnológica PAPIIT*, grant IA204016 (JLMF).

*The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.*

#### Acknowledgements

Insightful discussions with Dr. Jakyung Yoo (Daewoong Life Science Research Institute) are highly appreciated. The authors thank OpenEye for the academic license granted.



## Supplementary material

**Supplementary File 1: File with supporting methods, results and five figures.** Figure S1: 3D-Consensus Diversity Plot depicting the diversity of the datasets used for the backwards approach; Figure S2: Backwards analysis with 3PCs picking satellites by diversity; Figure S3: Backwards analysis with 3PCs picking satellites at random; Figure S4: Forward analysis with 2PCs picking satellites at random with step sizes of 10%; Figure S5: Plot of the percentage of variance explained by each principal component in the studied datasets.

[Click here to access the data.](#)

## References

- Medina-Franco J, Martinez-Mayorga K, Giulianotti M, *et al.*: **Visualization of the chemical space in drug discovery.** *Curr Comput-Aided Drug Discov.* 2008; **4**(4): 322–333.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- Reymond JL: **The chemical space project.** *Acc Chem Res.* 2015; **48**(3): 722–730.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- Naveja JJ, Medina-Franco JL: **Activity landscape sweeping: insights into the mechanism of inhibition and optimization of DNMT1 inhibitors.** *RSC Adv.* 2015; **5**(78): 63882–63895.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- Maggiara GM, Bajorath J: **Chemical space networks: a powerful new paradigm for the description of chemical space.** *J Comput Aided Mol Des.* 2014; **28**(8): 795–802.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- Medina-Franco JL: **Interrogating novel areas of chemical space for drug discovery using chemoinformatics.** *Drug Dev Res.* 2012; **73**(7): 430–438.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- Osolodkin DI, Radchenko EV, Orlov AA, *et al.*: **Progress in visual representations of chemical space.** *Expert Opin Drug Discov.* 2015; **10**(9): 959–973.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- Larsson J, Gottfries J, Muresan S, *et al.*: **ChemGPS-NP: tuned for navigation in biologically relevant chemical space.** *J Nat Prod.* 2007; **70**(5): 789–794.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- Larsson J, Gottfries J, Bohlin L, *et al.*: **Expanding the ChemGPS chemical space with natural products.** *J Nat Prod.* 2005; **68**(7): 985–991.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- Rosén J, Lövgren A, Kogej T, *et al.*: **ChemGPS-NP(Web): chemical space navigation online.** *J Comput Aided Mol Des.* 2009; **23**(4): 253–259.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- Oprea TI, Gottfries J: **Chemography: the art of navigating in chemical space.** *J Comb Chem.* 2001; **3**(2): 157–166.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- Awale M, Reymond JL: **Similarity Mapplet: Interactive Visualization of the Directory of Useful Decoys and ChEMBL in High Dimensional Chemical Spaces.** *J Chem Inf Model.* 2015; **55**(8): 1509–1516.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- Wishart DS, Knox C, Guo AC, *et al.*: **DrugBank: a comprehensive resource for *in silico* drug discovery and exploration.** *Nucleic Acids Res.* 2006; **34**(Database issue): D668–72.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- OpenEye Scientific Software, Santa Fe NM: **ROCS 3.2.1.4.** 2017.  
[Reference Source](#)
- OpenEye Scientific Software, Santa Fe NM: **OMEGA 2.5.1.4.** 2017.  
[Reference Source](#)
- Hawkins PC, Skillman AG, Warren GL, *et al.*: **Conformer generation with OMEGA: algorithm and validation using high quality structures from the Protein Databank and Cambridge Structural Database.** *J Chem Inf Model.* 2010; **50**(4): 572–584.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Hawkins PC, Skillman AG, Nicholls A: **Comparison of shape-matching and docking as virtual screening tools.** *J Med Chem.* 2007; **50**(1): 74–82.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- Berthold MR, Cebron N, Dill F, *et al.*: **KNIME - the Konstanz information miner.** *SIGKDD Explor Newsl.* 2009; **11**(1): 26.  
[Publisher Full Text](#)
- Naveja JJ, Medina-Franco JL: **Dataset 1 in: ChemMaps: Towards an approach for visualizing the chemical space based on adaptive satellite compounds.** *F1000Research.* 2017.  
[Data Source](#)



Cite this: *RSC Adv.*, 2015, 5, 63882

## Activity landscape sweeping: insights into the mechanism of inhibition and optimization of DNMT1 inhibitors†‡

J. Jesús Naveja<sup>ab</sup> and José L. Medina-Franco<sup>\*a</sup>

The interest in developing inhibitors of DNA methyltransferases (DNMT) as modifiers of epigenetic features for the treatment of several chronic diseases is rapidly increasing. Herein, we present insights of a chemoinformatic characterization of DNMT focused on the analysis of the chemical space and structure–activity relationships (SAR) using activity landscape modeling (ALM). Analysis of the chemical space revealed two main groups of compounds whose chemical structures are associated with either cofactor analogs or non-nucleoside compounds. The ALM showed that non-nucleoside compounds have a continuous SAR while cofactor analogs have a rough SAR with several deep activity cliffs. Molecular modeling helped to explain the structural basis of the activity cliffs. The significance of the results is threefold: (1) the combined analysis of chemical space with activity landscape gave rise to a novel ‘activity landscape sweeping’ strategy that enabled a better structure-based interpretation of the SAR; (2) it is feasible – and advisable – to develop predictive models for non-nucleoside DNMT studied in this work, and (3) structure-based interpretation of the SAR gave clear insights into the molecular mechanism of inhibition of novel DNMT suggesting specific strategies to optimize the activity of leads compounds.

Received 25th June 2015  
Accepted 20th July 2015

DOI: 10.1039/c5ra12339a

www.rsc.org/advances

### Introduction

The term ‘Epigenetics’ was initially defined as “the interactions of genes with their environment, which brings the phenotype into being”.<sup>1</sup> Epigenetic drug discovery is an attractive research area in oncology and for the treatment of other chronic diseases associated with epigenetic alterations, particularly those influenced by the environment. There are several epigenetic targets which are broadly classified in three major groups, namely; readers, writers and erasers of the epigenetic information.<sup>2</sup>

DNA methylation is a major epigenetic change that regulates gene expression in the genome of organisms that range from viruses to humans.<sup>3</sup> DNA methylation is regulated by the family of enzymes DNA methyltransferases (DNMTs). DNMTs are responsible for the covalent addition of a methyl group from the cofactor *S*-adenosyl-L-methionine (SAM or AdoMet) (Fig. 1) to the carbon atom 5 of cytosine, preferably within CpG dinucleotides. Also, as a product of the methylation mechanism,

*S*-adenosyl-L-homocysteine (SAH) is generated.<sup>4</sup> In mammals, four DNMT enzymes have been identified: DNMT1 (the most abundant, it is a maintenance methyltransferase that acts on hemimethylated DNA); DNMT3A and DNMT3B (*de novo* methyltransferases that are capable of generating new methylation patterns in DNA), and DNMT3L that is associated with DNMT3A and DNMT3B, enhancing their activity.

The structure of DNMTs can be organized into a C-terminal catalytic domain and an N-terminal regulatory domain. The catalytic domain of all DNMTs shares a common structure called “AdoMet (SAM)-dependent Mtase fold”. The N-terminal domain is involved in distinguishing hemi- and unmethylated DNA. There are several three-dimensional (3D) structures of different domains of DNMTs, including the catalytic one.<sup>5</sup>

The role of DNMTs in carcinogenesis has been subject of intense research during the last ten years. Currently, there are two inhibitors of DNMT (IDNMT) in clinical use: 5-azacytidine and decitabine (Fig. 1) both approved by the United States Food and Drug Administration – FDA – for the treatment of myelodysplastic syndrome (MDS).<sup>6</sup> However, these two drugs are cytosine analogues that are incorporated into DNA, which implies they are unspecific and have high toxicity due to their mutagenic effects that may occur in somatic cells. Many therapies involving IDNMT are under investigation, mainly as sensitizers to therapy, since epigenetic changes may be involved in rapid adaptation of cancerous cells to therapy. In addition to cancer, DNMTs are attractive targets for the treatment of other

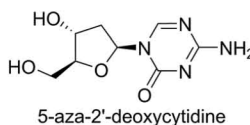
<sup>a</sup>Facultad de Química, Departamento de Farmacia, Universidad Nacional Autónoma de México, Avenida Universidad 3000, México, D.F. 04510, México. E-mail: medinajl@unam.mx; jose.medina.franco@gmail.com; Tel: +52-55-5622-3899 ext. 44458

<sup>b</sup>Facultad de Medicina, PECEM, Universidad Nacional Autónoma de México, Avenida Universidad 3000, México, D.F. 04510, México

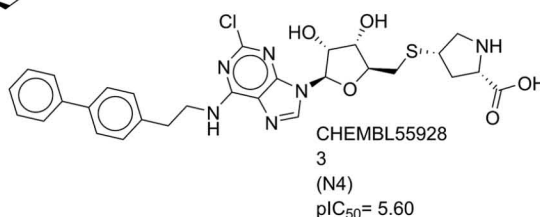
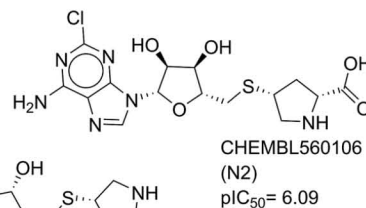
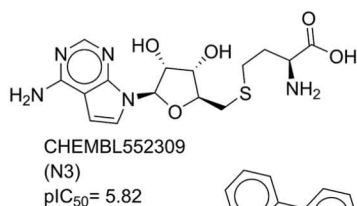
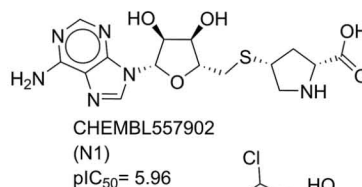
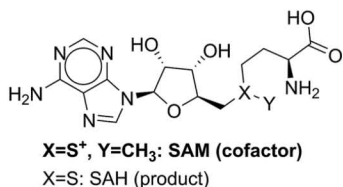
† This work is dedicated to the memory of Dr Andoni Garritz Ruiz.

‡ Electronic supplementary information (ESI) available. See DOI: 10.1039/c5ra12339a

## Cytosine analogs



## Cofactor analogs



## Non-nucleoside

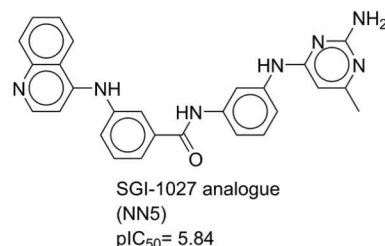
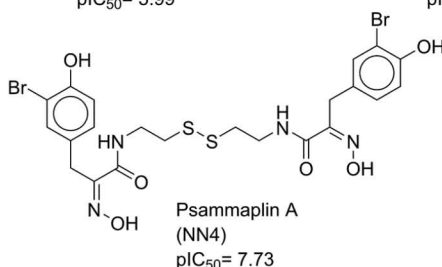
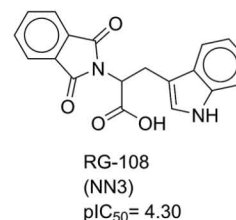
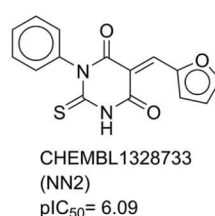
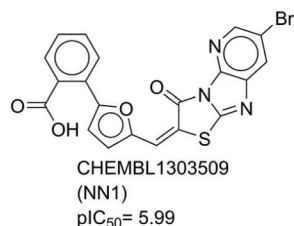


Fig. 1 Structures of representative IDNMT1. The relative position in chemical space of selected compounds: 4 SAM analogues (N) and 5 non-nucleoside (NN) compounds is shown in Fig. 2.

chronic and degenerative diseases such as Alzheimer's and psychiatric conditions. Also, DNA methylation has been involved in autoimmune diseases and inherited disorders.<sup>7</sup>

The low specificity and high toxicity of 5-azacytidine and decitabine has prompted the search for novel and specific IDNMTs. Currently there is a relatively large number of IDNMT and/or DNA demethylating compounds that have been obtained from different sources such as natural products, synthetic compounds, drugs approved for therapeutic indications other than cancer and high-throughput screening (HTS). As part of these efforts, computational analyses have been successfully implemented to model IDNMT and to identify novel inhibitors.<sup>8</sup>

Over the past few years, the structure and activity of compounds tested as IDNMT have been collected in public repositories such as ChEMBL.<sup>9</sup> The increasing amount of structure–activity data of IDNMT opens up the possibility to conduct systematic structure–activity relationships (SAR) studies, such as quantitative SAR (QSAR). Nevertheless, it has been recognized that typical QSAR studies usually assume that compounds with similar structures have similar activity *i.e.*, a 'smooth' SARs. It is well known that compounds with high structural similarity but low activity similarity *i.e.*, 'activity cliffs',<sup>10</sup> reduce the predictive ability of QSAR models.<sup>11,12</sup> Therefore, the early detection of activity cliffs is a convenient step before attempting to develop models such as QSAR.<sup>13</sup>

Similarly, it is advisable to conduct detailed descriptive analysis to understand the SAR before developing predictive models.<sup>14</sup> Thus far, limited studies have been reported to navigate and describe the SAR of a large set of IDNMT in a systematic manner.

In this work, we report a chemoinformatic-based characterization of the SAR of a dataset of 280 compounds tested as IDNMT1 and deposited in ChEMBL. The analysis had three specific aims: (a) characterization of the structural diversity and distribution in chemical space of the data set; (b) descriptive SAR analysis using the concept of activity landscape modeling and (c) structure-based interpretation of the activity cliffs. To the best of our knowledge this work represents one of the first activity landscape studies of IDNMT1. Indeed, it has been recently recognized that activity landscape modeling (ALM) is a convenient approach to explore systematically the SAR of screening data sets focused on epigenetic targets.<sup>15</sup> The characterization of the chemical space distinguished two major types of chemical structures with different activity landscape. As part of the first aim it was developed a novel 'activity landscape sweeping' approach, that is, a dissection of the global activity landscape (global SAR) into smaller but more structural interpretable local landscapes (local SARs). The structure-based interpretation of the SAR of the activity cliffs gave key insights into the molecular mechanism of inhibition of active molecules. This analysis also prompted for structural modifications to lead compounds to continue developing IDNMT as potential epi-drugs or epi-probes.

## Methods

### Dataset

A data set of 280 compounds with different (no duplicate) chemical structures and activity against DNMT1 was obtained from ChEMBL (version 20)<sup>9</sup> and recent literature. Only compounds with reported IC<sub>50</sub> values obtained in enzymatic inhibitory assay were included in the analysis. The activity range for the compounds in the dataset was 18.6–1 600 000 nM (pIC<sub>50</sub> range 7.73–2.80). Molecules were pre-processed with the 'washing' workflow implemented in Molecular Operating Environment (MOE) (version 2010.10,<sup>16</sup>). During the washing procedure, only the largest molecular structure was retained; counter ions, if present, were removed and protonation states were set to neutral. Visualization of the chemical structures was performed with MOE and Data Warrior (version 4.1.1).<sup>17</sup>

### Structural similarity

In order to measure the structural similarity for each pair of compounds in the dataset (39 060 pair-wise comparisons) we employed two structural fingerprints of different design, namely, Molecular Access System (MACCS) 166 bits (dictionary based fingerprints)<sup>16</sup> as implemented in MOE, and extended connectivity fingerprints (ECFP; radial based fingerprints),<sup>18</sup> with neighborhood radius of 2 as implemented in MayaChemTools (<http://www.mayachemtools.org>). The structural similarity was computed with the Tanimoto coefficient.<sup>19,20</sup>

### Data fusion

In order to explore the effect of data fusion in this study, two approaches were implemented to combine the similarity values computed with MACCS keys and ECFP: (a) fusion mean *i.e.*, calculation of the mean values<sup>21</sup> and (b) Z-fusion *i.e.*, addition of the Z-transformed values of both fingerprints.<sup>15</sup>

### Visual representation of chemical space

To obtain a visual representation of the chemical space<sup>22</sup> we conducted a principal components analysis (PCA) on the similarity matrices computed for the 280 molecules with the two fingerprints and the two fusion approaches. This method has been broadly used to obtain visual representations of the chemical space.<sup>20,23</sup> The PCA was performed with the FactoMineR R package version 1.29. For visualization, the ggplot2 R package was used (<http://www.R-project.org/>).<sup>24</sup> K-means method was also conducted with R using in-house scripts to perform clustering of the PCA's output. Further details of the PCA and K-means analysis employed are in the ESI.†

### SAS maps

The activity landscape of IDNMT was explored using Structure–Activity Similarity (SAS) maps.<sup>25</sup> SAS and related 2D- and 3D-SAS maps have been extensively employed to describe the SAR of a large number of data sets.<sup>26–28</sup> Features of SAS maps, including their advantages and disadvantages, are elaborated elsewhere.<sup>29</sup> Briefly, a typical SAS map is a 2D plot of the structural similarity *vs.* the potency difference of all possible pairs of compounds in a data set. The structural similarity can be computed with any similarity method. Aggregation of similarity values using data fusion may be implemented.<sup>26,30</sup> To facilitate the visual interpretation of the SAS maps, 'density SAS maps' were used in this work. A density SAS map represents the frequency of data points usually with a continuous color scale.<sup>15</sup> Density SAS maps were generated for the entire data set (*e.g.*, analysis of the 'global activity landscape') and for subsets of compounds that emerged from the analysis of the compounds in chemical space (*e.g.*, analysis of the 'local activity landscape').

### Activity cliffs generators

'Activity cliffs generators'<sup>31</sup> were defined as active compounds recurrent (frequency >1) in the activity cliff region of the activity landscape. In turn, the 'activity cliff' region of the landscape was defined as the quadrant in the SAS map that contains pairs of compounds with high structure similarity and high potency difference. A quantitative definition of 'high' structure similarity is not straightforward. Herein, we considered high values those with two standard deviations above the mean similarity of the entire data set. Two values to define 'very high' and 'high' potency difference were used to distinguish 'deep' and 'shallow' activity cliffs, respectively *i.e.*,  $\Delta\text{pIC}_{50} > 2$  standard deviations above the mean (2SD) and  $2\text{SD} > \Delta\text{pIC}_{50} > 1$ .



### Structure-based interpretation of activity cliffs

In order to provide a structure-based rationalization of the activity cliffs that emerged from the activity landscape analysis, we conducted computational studies with the crystallographic structure of the catalytic domain of human DNMT1 co-crystallized with SAH (PDB ID: 3PTA).<sup>32</sup> Notably, the conformation of the catalytic domain of human DNMT1 (Protein Data Bank (PDB) ID: 3PTA) shows the prevention of the *de novo* methylation mechanism by an auto-inhibitory linker that blocks DNA to reach the catalytic site. For docking studies we employed MOE 2010 using default parameters. The binding cavity was defined differently for nucleoside and non-nucleoside structures (*vide infra*). The docking protocol was validated by re-docking the co-crystal structure with a root-mean-square deviation (RMSD) of: 1.22 Å<sup>2</sup> for the best scored pose. The docking poses were further analyzed using Protein Ligand Interaction Fingerprints (PLIFs)<sup>33</sup> implemented in MOE as detailed below. PLIFs, also called structural interaction fingerprints, capture key 3D interactions between a ligand and a protein in 1D. PLIFs were recently used in activity landscape studies.<sup>34</sup>

**Molecular modeling of nucleoside activity cliffs.** In order to propose a structure-based explanation of the activity cliffs of the cofactor analogues, we worked under the hypothesis that these compounds bind in the cofactor binding site. We also assumed that, in general, the compounds have a binding orientation comparable to that of SAH. Therefore, we conducted docking using pharmacophoric constraints that were obtained from the crystallographic binding mode of SAH. The pharmacophore had four points: hydrogen bond donor with Cys1191, hydrogen bond acceptor with Met1169, anion and hydrogen bond acceptor with Gly1150 and aromatic ring interacting with Phe1145. During docking in MOE it was enabled the partial homology criterion with the formation of at least three of the pharmacophoric constraints, the rest of the options remained as default (Fig. S2 in the ESI†). The docked poses were post-processed with PLIFs available in MOE to identify the most relevant interactions of the activity cliffs *i.e.*, potential hot spots.<sup>35</sup>

**Molecular modeling of non-nucleoside activity cliffs.** Since there is no experimental evidence of the binding site for most of the non-nucleoside DNMT1, we searched for potential binding sites in the catalytic domain of DNMT1 using site finder in MOE with default parameters. Then, the most active compound forming activity cliffs was docked with MOE in the absence of the co-factor considering all putative binding sites. The geometry of the docking pose with the best docking score was minimized with the cofactor present using the MMFF94x force field as implemented in MOE. To conduct the minimization, the ligand and nearby residues of the binding pocket (with atoms within a radius of 4.5 Å) were selected. Default parameters were used. In order to detect putative ‘interaction cliffs’ (*i.e.*, ligand–target complexes with high structural and interaction similarity but a large potency difference of the ligands),<sup>34</sup> the optimized docked pose of the most active compound was used as a template to conduct flexible alignment of the other cliff

forming compounds with which it formed activity cliffs. The flexible alignment was done in MOE using default parameters. For comparison, both regular and pharmacophore-constrained docking (see pharmacophore in Fig. S3†) were performed in the binding site proposed for the most active molecule (with the energy-minimized conformation of the protein).

## Results and discussion

### Structural diversity analysis

A structural diversity analysis of the 280 compounds was conducted using two molecular fingerprints of different design.<sup>30</sup> The distribution of the similarity values (Fig. S1 in the ESI†) showed that, in general, this is a relatively diverse set with, for example, mean Tanimoto similarity values of 0.63 (MACCS keys) and 0.11 (ECFP). This diversity is comparable to that reported for other sets of compounds tested for other therapeutic indications.<sup>36</sup>

### Visualization of chemical space

Activity landscapes have been defined as methods that find the association between structure similarity and activity similarity.<sup>37</sup> Therefore, the next step in this work was to explore the distribution of the 280 compounds in chemical space. Fig. 2 shows a visual representation of the chemical space obtained with PCA of the similarity matrix computed with ECFP and the Tanimoto coefficient. Data points are colored by the pIC<sub>50</sub>

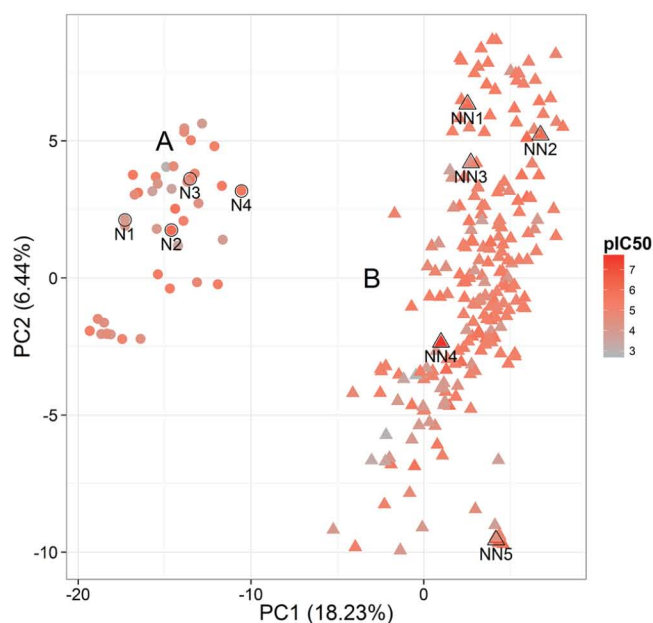


Fig. 2 Visual representation of the chemical space of the 280 compounds in the data set. The visualization was obtained by principal component analysis of the similarity matrices computed with ECFP. The percentage of variance explained by each PC is indicated in the corresponding axis. Data points are colored by the pIC<sub>50</sub> values in a continuous scale. Two main clusters (A: circles, B: triangles) are readily distinguished. Nine selected compounds are identified as SAM analogues (N) and non-nucleosides (NN) compounds. The chemical structures are shown in Fig. 1.

values using a continuous color scale from red (more active) to gray (less active).

Fig. 2 shows two major clusters in chemical space herein labeled as cluster A (45 compounds) and cluster B (235 compounds), respectively. Both groups of compounds have active and inactive molecules *e.g.*, red and gray points. Furthermore, the active compounds in each cluster are not further grouped suggesting that they are structurally diverse.

Visual inspection of all compounds in each cluster revealed that all the chemical structures in cluster A have a purine ring in their structure and are structurally related to the co-factor SAM. In contrast, molecules in cluster B are non-nucleoside.

Representative structures from each cluster are depicted in Fig. 1 and are mapped into the visual representation of the chemical space of Fig. 2. The visual representation of the chemical space in Fig. 2 also suggested that molecules in cluster B (non-nucleoside) are structurally more diverse than the molecules in cluster A. Not surprisingly, the distribution of the similarity values (Fig. S1 in the ESI<sup>†</sup>) confirmed that the non-nucleoside set has a higher structural diversity than the SAM-related compounds. This is because no further distinction is made on the type of chemical structures. In contrast, all compounds in cluster A are chemically related to SAM.

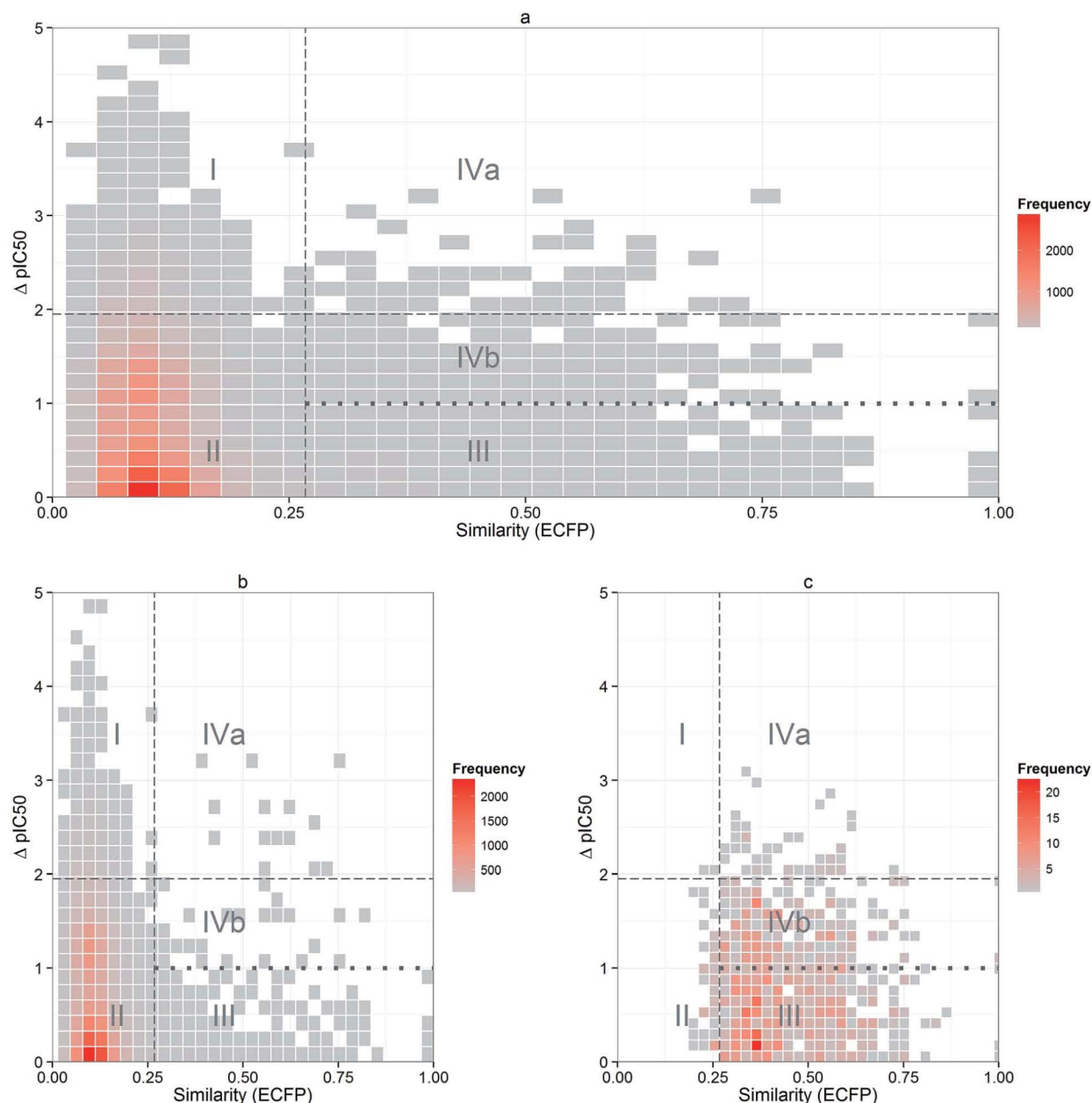


Fig. 3 Density SAS maps of the global and local activity landscapes. The 2D plots are colored by the frequency of data points in the coordinates given. Dashed lines divide the maps into the four quadrants labeled I–IV. The dotted line further divides the activity cliff quadrant (IV) in two regions (IVa and IVb) to distinguish shallow and deep cliffs (see text for details). (a) IDNMT1 SAS map for the entire set with 280 compounds (39 060 pairwise comparisons). (b) SAS map for 235 non-nucleoside compounds (cluster B in Fig. 2) (27 495 pairwise comparisons). (c) SAS map for 45 SAM analogues (cluster A in Fig. 2) (990 pairwise comparisons).



It is possible to further divide the non-nucleosides in smaller sub-sets chemically related. For instance, K-means clustering shows that 3–6 subgroups would provide an efficient clustering in terms of number of clusters and within group's sums of squares (see the ESI† for a detailed explanation on K-means methodology followed). However, clustering in two groups already diminished by more than 40% of the within groups sums of squares (see Fig. S4†). Herein, we analyze the activity landscape of two clusters to discuss local SAR as general as possible. Undoubtedly, additional studies can be extended to analyze smaller clusters and provide information of more local SARs.

Equivalent clusters A and B were identified in the PCA of the combined ECFPs and MACCS keys similarity matrices using the fusion approaches detailed in the Methods section (Fig. S5 in the ESI†). Interestingly, MACCS keys alone did not lead to the identification of the two clusters (Fig. S5a†); this can be attributed to the low resolution of this fingerprint.<sup>30</sup>

### Overview of activity landscape

**Global activity landscape (global SAR).** Fig. 3a shows a density SAS map generated with ECFP and Tanimoto for the entire data set with 280 compounds. The four major quadrants (I–IV) are distinguished in the figure. The activity cliff zone (region IV) is further divided in two sub-regions (IVa and IVb) that distinguish the shallow from the deep activity cliffs, depending on the potency difference (1 vs. 2 log units; see the Methods section). The amount of data points in each different region of the SAS map is visually represented with a continuous color scale from red (more data points) to gray (fewer). Table 1 summarizes the fraction of data points in each region (I–IV) of the SAS map.

Fig. 3a and Table 1 indicate that, overall, IDNMT1 have a heterogeneous SAR with data points in the continuous and discontinuous regions of the SAR (zones III and IV).<sup>15</sup> Noteworthy, the scaffold hop, more recently called 'similarity cliffs'<sup>38</sup> region has the highest density of data points (92.6%). This indicates that there are quite different chemical structures with similar activity. Note however that both compounds in the pair may be either active or inactive. Fig. 3a and Table 1 also shows the presence of shallow and deep activity cliffs with a relatively small fraction of the entire data set (0.79 and 0.16%, respectively). The overall low frequency of activity cliffs is in agreement

with the low frequency of activity cliffs observed for data sets for other molecular targets.<sup>26–28,30</sup>

The high density of data points in the similarity cliff region of the SAS maps and the two main clusters of compounds distinguished in the chemical space analysis, prompted us to conduct analysis of local activity landscapes of clusters A and B. As discussed in the next section, the chemical structures of compounds in each cluster, plus the knowledge of the mechanism of DNA methylation, led to an interpretable SAR.

**Local activity landscapes (local SARs).** Fig. 3b and c show the density SAS maps generated for the 235 non-nucleosides and 45 SAM analogues identified in the analysis of the chemical space (clusters B and A in Fig. 2, respectively). Table 1 summarizes the number and percentage of data points in the four major regions of the local SAS map. The number and fraction of the deep and shallow cliffs (regions IVa and IVb, respectively) are also summarized in the same table.

The lower fraction of similarity cliffs for SAM-related analogues (4.6%) vs. the fraction of similarity cliffs for the non-nucleoside analogues (94.8%) is in agreement with the type of structures and molecular diversity in each cluster. Indeed, the visual representation of the chemical space (Fig. 2) and distribution of ECFP/Tanimoto similarity values for the compounds in each cluster (Fig. S1†) yield consistent results. Similarly, the higher percentage of compounds in the smooth SAR region (III) for SAM analogues (62%) as compared to the percentage of compounds for non-nucleoside analogues (1.2%) (Table 1) is in line with the structural diversity of the chemical structures of each type of compounds.

As mentioned above, the distribution of data points in the similarity cliff and smooth regions of the SAS maps are expected from the type of chemical structures. But surprisingly, for SAM related analogues there is a larger fraction of deep and shallow activity cliffs as compared to the fraction of cliffs in the entire data set (4.9% and 28.2% vs. 0.16% and 0.79%, respectively; Table 1). In sharp contrast, the fraction of activity cliffs for the non-nucleosides is lower (0.05% and 0.11%, respectively, Table 1). These results indicate that SAM related analogues may be enriched with activity cliff generators.<sup>31</sup> The next sections discuss the activity landscapes of each set of compounds, *i.e.*, local activity landscapes. A brief analysis of the activity landscape of SAM-related compound is mentioned first followed by a more extensive discussion of the landscape of the non-

**Table 1** Number and proportion of pairs of compounds into the four different regions of the global and two local SAS maps

Quadrant	Region	Entire dataset <sup>a</sup>	SAM analogues (cluster A) <sup>b</sup>	Non-nucleosides (cluster B) <sup>c</sup>
I	Uncertainty	1571 (4.02%)	2 (0.20%)	1066 (3.88%)
II	Similarity cliff (scaffold hop)	36 177 (92.61%)	46 (4.64%)	26 059 (94.78%)
III	Smooth SAR	939 (2.40%)	614 (62.02%)	325 (1.18%)
IVa	Deep activity cliffs	64 (0.16%)	49 (4.95%)	15 (0.05%)
IVb	Shallow activity cliffs	309 (0.79%)	279 (28.18%)	30 (0.11%)
Total		39 060 (100%)	990 (100%)	27 495 (100%)

<sup>a</sup> 280 compounds. <sup>b</sup> 45 compounds in cluster A of Fig. 2. <sup>c</sup> 235 compounds in cluster B of Fig. 2.

nucleosides. We elaborated more on the non-nucleosides since they are currently more attractive as IDNMT1.<sup>39</sup>

### Activity landscape of SAM-related compounds

As discussed above, SAM-related compounds have a discontinuous SAR with several (nearly 5%) of activity cliffs. For

comparison, the non-nucleosides have 0.05% of activity cliffs. Despite the fact these proportions are dependent of the current contents of ChEMBL *i.e.*, the numbers may change as more activity data is published, this is a clear indication of the rough nature of the SAR of SAM-related compounds. This observation highlights the challenge to conduct lead optimization of IDNMT1 using SAM-related compounds

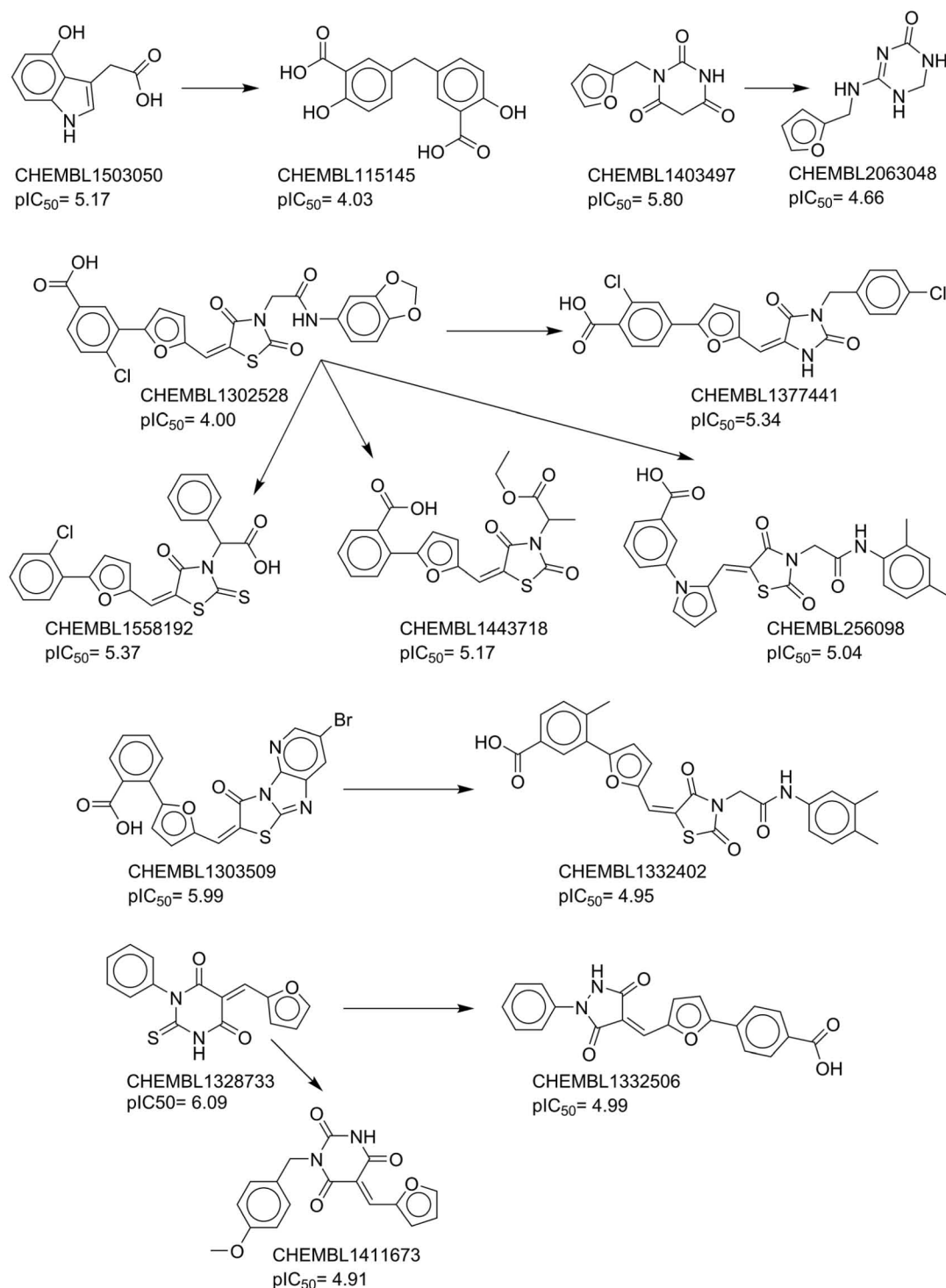


Fig. 4 Structures of activity cliffs of non-nucleoside compounds identified by high-throughput screening. Table 2 summarizes the potency difference and structure similarity for each compound pair associated with an arrow.

besides the risk of hitting other methyltransferases. Deep activity cliffs generators of SAM-related compounds (cluster A) are shown in Fig. 1 (N1–N4). The compound pairs with whom they form activity cliffs are illustrated in Fig. S7–S10 in the ESI.†

### Activity landscape of non-nucleoside compounds

The activity landscape of the non-nucleoside compounds is more continuous than the landscape of SAM-related molecules. The landscape of the non-nucleosides is characterized by a small fraction of activity cliffs of which a small number are deep activity cliffs (Table 1).

As discussed in the literature, activity cliffs are rich in SAR information since they point to specific structural changes that have a large impact in the biological activity. In an activity landscape study based on structural fingerprints, the interpretability of the activity cliffs is a key component.<sup>37</sup> In other words, the SAR of the activity cliffs should be easily translated in terms of specific structural changes. In the local activity landscape of non-nucleoside molecules we identified two major types of compounds with high ECFP/Tanimoto similarity whose chemical structures are structurally related, namely: compounds identified by HTS and structures related to SGI-1027.<sup>40</sup> All pairs of compounds from HTS are shallow cliffs and are shown in Fig. 4 and S11 of the ESI.† From the 30 shallow activity cliffs found in the SAS map for non-nucleoside compounds, 16 (53%) compounds were found to be from HTS assays (Fig. 4 and Table 2). A considerable number of screenings and confirmatory assays were performed for these compounds, as found in PubChem.

In the activity landscape of non-nucleoside molecules the deepest activity cliffs as well as the most relevant in medicinal chemistry were the structures related to the quinolone-based inhibitor SGI-1027. This compound is one of the most promising IDNMT1 that has been recently subject of a medicinal chemistry optimization program (*vide infra*).

Therefore, in the next section we describe studies focused on the interpretation at a molecular level of activity cliffs related to SGI-1027.

### Deep and shallow cliffs for compounds related to SGI-1027.

Systematic analysis of all pairwise comparisons of the structure and activity of the 235 non-nucleoside compounds (39 060 comparisons), readily uncovered that analogues of SGI-1027 are the compounds with the most dramatic changes in activity associated with a small change in the structure. In fact, 14/30 (47%) of the shallow and 15/15 (100%) of the deep activity cliffs found in the non-nucleoside database were found to be related to the compounds recently synthesized by Valente *et al.*<sup>41</sup> The chemical structures of the nine activity cliff-forming compounds are presented in Fig. 5. The enzymatic inhibitory activity of the nine compounds *vs.* DNMT1 was recently reported using the same assay conditions.<sup>41</sup> These molecules were synthesized as part of a hit-to-lead optimization program of SGI-1027 which showed high potency in enzyme and cell assays.<sup>40</sup> Compounds in Fig. 5 are regioisomers of SGI-1027.

In order to describe the analogues of the lead compound, Valente *et al.* considered that SGI-1027 is composed of four fragments (4-aminoquinoline + 4-aminobenzoic acid + 1,4-phenylenediamine + 2,4-diamino-6-methylpyrimidine) linked in sequence with *para/para* orientation.<sup>41</sup> The most active compound in this series was CHEMBL3126646 which can be regarded as the *meta/meta* regioisomer of SGI-1027 (CHEMBL2336409).<sup>41</sup> Table 3 summarizes the deep activity cliffs that form CHEMBL3126646. It must be noted that this compound is the most important activity cliff generator in the database *i.e.*, it is the most prevalent compound within the activity cliff region of the SAS map.<sup>31</sup> The deepest activity cliffs of the *meta/meta* regioisomer are formed with *ortho* regioisomers such as CHEMBL3126644, 3126647, 3126648, 3126649 with potency differences of two or more logarithmic units (Fig. 5 and Table 3).

Table 2 Shallow activity cliffs formed by non-nucleoside compounds that are not SGI-1027 regioisomers

Compound pair	Activity of most active compound in the pair (pIC <sub>50</sub> )	ΔpIC <sub>50</sub>	ECFP/Tanimoto
CHEMBL115145, CHEMBL1503050	5.17	1.14	0.28
CHEMBL1302528, CHEMBL1377441	5.34	1.34	0.3
CHEMBL1302528, CHEMBL1443718	5.17	1.17	0.3
CHEMBL1302528, CHEMBL1558192	5.37	1.37	0.28
CHEMBL1302528, CHEMBL256098	5.04	1.03	0.3
CHEMBL1303509, CHEMBL1332402	5.99	1.04	0.27
CHEMBL1328733, CHEMBL1332506	6.09	1.1	0.27
CHEMBL1328733, CHEMBL1411673	6.09	1.18	0.37
CHEMBL1379120, CHEMBL592316	5.91	1.9	0.28
CHEMBL1403497, CHEMBL2063048	5.8	1.14	0.28
CHEMBL1564869, CHEMBL3109084	4.7	1.29	0.39
CHEMBL1607517, CHEMBL1704614	5.87	1.49	0.36
CHEMBL1607517, CHEMBL1988862	5.99	1.6	0.46
CHEMBL1916517, CHEMBL1916672	3.82	1.03	0.55
CHEMBL1978925, CHEMBL1990599	5.27	1.26	0.32
CHEMBL1983083, CHEMBL1990599	5.07	1.07	0.38

Valente *et al.* reported docking models of CHEMBL3126646 with crystallographic structures of DNMT1. It was concluded from that studies that this molecule could interact with the CXXC auto-inhibitory domain of DNMT1 and be close to SAM but without making interactions with the cofactor or competing with any of the interactions that SAM makes.<sup>41</sup> However, no structure-based explanation of the large potency difference of the significantly less active SGI-1027 analogues (*e.g.*, *ortho*

regioisomers) was explored. A structure-based interpretation of the activity cliffs that form the most active compound is elaborated in the next section.

### Structural interpretation of representative activity cliffs

Structure-based interpretation of the activity cliffs can help to understand the SAR of data sets at the molecular level and provide insights to optimize the activity.<sup>31,35,42</sup> In this study, the

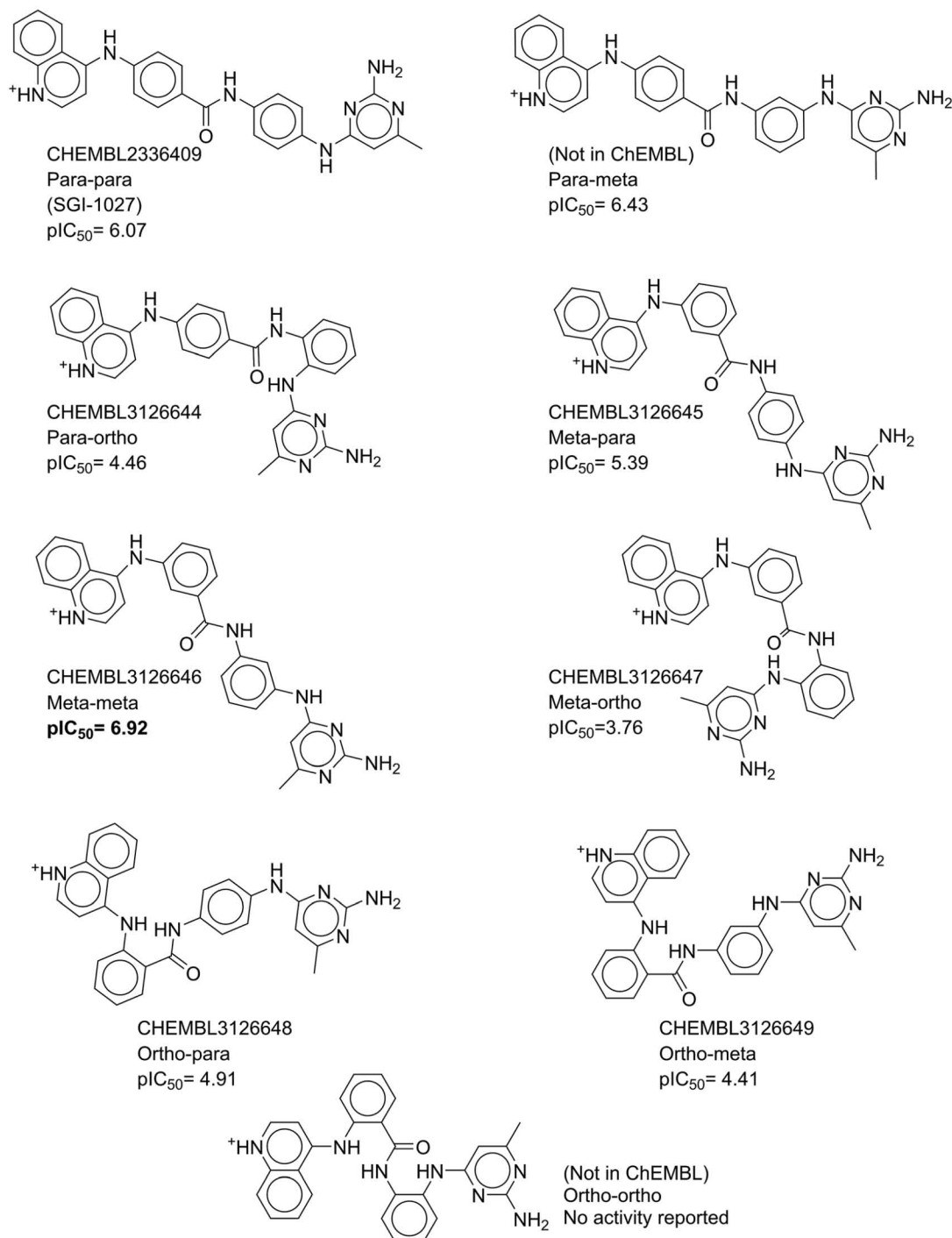


Fig. 5 Chemical structures of non-nucleoside activity cliffs related to regioisomers of SGI-1027. Table 3 summarizes the potency difference and structure similarity for each compound in this figure and the lead molecule CHEMBL3126646.



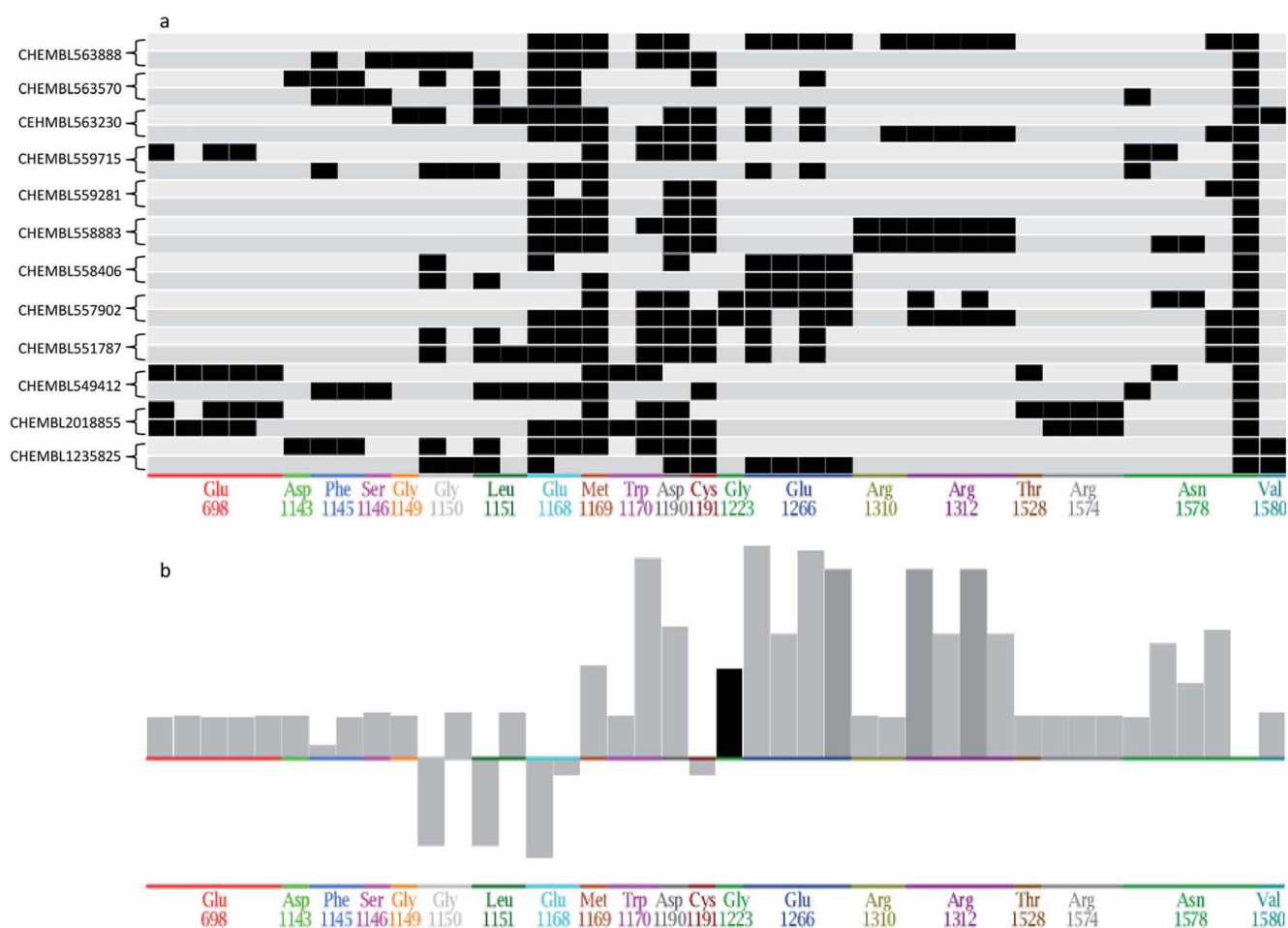
**Table 3** Activity cliffs formed by CHEMBL3126646 (*meta/meta* SGI-1027 regioisomer)

Compound	$\Delta\text{pIC}_{50}$	ECFP/Tanimoto
CHEMBL3126647	3.16	0.75
<i>ortho/ortho</i> SGI-1027 regioisomer	3.16	0.53
CHEMBL3126654	3.16	0.38
CHEMBL3126649	2.51	0.69
CHEMBL3126644	2.46	0.60
CHEMBL3126653	2.37	0.42
CHEMBL3126648	2.01	0.57

availability of structure information of the 3D coordinates of DNMT1 enabled a structure-based interpretation of the activity cliffs using molecular modeling. Of note, despite the fact docking studies of all compounds reported in ChEMBL as IDNMT1 is warranted, this is out of the scope of this work.

Herein, we focus on the structure-based analysis of the most representative activity cliffs. As detailed in the Methods section, we employed different modeling strategies to study the activity cliffs related to SAM analogues and to CHEMBL3126646 based on the structural information available for each type of compounds.

**SAM-related activity cliffs.** The experimental co-crystal structure of SAH bound in the co-factor site of DNMT1 was the starting point of the structure-based studies of relevant activity cliffs related to SAM. As described in the Methods section, we worked under the assumption that SAM-related activity cliffs bind in the co-factor binding site. We conducted docking studies using pharmacophoric constrains of the compounds forming activity cliffs with the four most prominent activity cliff generators related to SAM-analogues: CHEMBL557902, CHEMBL560106, CHEMBL552309, and CHEMBL559283 (labeled N1–4 in Fig. 1 and 2). The binding poses were analyzed using PLIFs.



**Fig. 6** Summary of protein–ligand interaction fingerprint (PLIFs) analysis of the activity cliff generator CHEMBL552309 (compound N3 in Fig. 1) and 11 SAM-analogues that form activity cliffs with this compound (the chemical structures of the 11 molecules are shown in Fig. S8 of the ESI†). For each compound the best two docking poses are represented. (a) Data matrix summarizing the protein–ligand contacts between the best two poses of 12 docked molecules and DNMT1. In this matrix, the rows represent the docked poses. The columns are the fingerprint bits indicating the amino acid residues that make at least one contact with one of the compounds. A black cell in the matrix indicates that a contact is present between the intersecting compound and amino acid residue *i.e.*, fingerprint bit turned ‘on’. In contrast, a white cell means that there is no contact *i.e.*, fingerprint bit turned ‘off’. (b) The statistically more significant PLIFs. A darker color means that the interaction is more associated to the active compound.

Results of the PLIFs for the activity cliff generator CHEMBL557902 plus 11 related (paired) compounds are shown in Fig. 6. The chemical structures are shown in Fig. S8 of the ESI.† The data matrix in Fig. 6a summarizes the protein–ligand contacts between the best two poses of 12 docked molecules and DNMT1. In this matrix, the rows represent the docked poses of the 12 molecules. The columns are the fingerprint bits indicating the amino acid residues that make at least one contact with one of the compounds. A black cell in the matrix indicates that a contact is present between the intersecting compound and amino acid residue *i.e.*, fingerprint bit turned ‘on’. In contrast, a white cell means that there is no contact *i.e.*, fingerprint bit turned ‘off’. Fig. 6 revealed that interactions with Gly1223 (backbone hydrogen bond donor), Glu1266 (ionic attraction) and Arg1312 (both side chain hydrogen bond acceptor and ionic attraction) were found in the active SAM-analogue (CHEMBL557902) but not in the compounds with much lower pIC<sub>50</sub> values. Similar analyses were performed with the three remaining activity cliff generators related to SAM (Fig. S12–S14†). It was concluded that the loss of a hydrogen bond donor that could interact with Asp1190 is generating cliffs for CHEMBL557902, CHEMBL560106, and CHEMBL559283.

**Non-nucleoside activity cliffs.** As stated above, in this study we focused on the structure-based interpretation of the most significant activity cliffs of the non-nucleoside molecules, *i.e.*, structural analogues of SGI-1027. In particular, we focus on the analysis of the deep activity cliffs formed with CHEMBL3126646 (Fig. 5). As explained above, these cliffs have large potency differences (>2 standard deviations above the dataset’s mean) and the high ECFP/Tanimoto similarity (ranging from 0.38 to 0.75, see Table 3) of these activity cliffs is structurally interpretable.

There is no co-crystallized structure available for the most active compound CHEMBL3126646 with DNMT1 (this is the

case for every non-nucleoside IDNMT1). Therefore, its precise binding region is unknown. In order to explore the putative binding zone, before docking all activity cliff forming compounds, CHEMBL3126646 was docked with DNMT1 as detailed in the Methods section. Results were compared with the experimental biochemical results and docking studies recently published for this molecule. Fig. 7 shows the optimized docking model. In this model, CHEMBL3126646 is close to but does not occupy the binding region of the co-cofactor (as predicted for other type of IDNMT1 (ref. 43 and 44)). Remarkably, a potential hydrogen bond interaction was found between the carbonyl oxygen of CHEMBL3126646 and the O<sub>2</sub> oxygen atom of the co-crystal SAH. The molecule is able to make hydrogen bond contacts with the backbone of Ala647, and  $\pi$ – $\pi$  interactions (T-shape) with the side chain of Phe648 of the CXXC domain. In addition, CHEMBL3126646 makes hydrophobic interactions with the side chains of Met696, Glu698, and Ala699 of the CXXC domain (see Fig. 7 and S15 in the ESI† for a 3D and 2D ligand interactions representation, respectively). The possibility of this inhibitor or making ‘sandwich’ interactions with both the CXXC domain and the co-factor in DNMT1 is in agreement with the docking study reported by Valente *et al.*<sup>41</sup> Therefore, it is plausible that CHEMBL3126646 inhibits DNMT1 by a mechanism we previously proposed for SGI-1027 *i.e.*, stabilization of the autoinhibitory linker.<sup>45</sup> This hypothesis is further supported by the experimental evidence that CHEMBL3126646 seems to do not compete with the co-factor.

The binding mode for the most active compound proposed herein also explains the activity cliffs to a large extent. The most pronounced *e.g.*, deepest activity cliffs with compound CHEMBL3126646 (Table 3) are regioisomers with at least one *ortho* substitution: CHEMBL3126644, 3126647, 3126648, 3126649 and *ortho/ortho* regioisomer (compound 9, as numbered in the Valente *et al.*<sup>41</sup> paper). In agreement with Valente *et al.*,<sup>41</sup> the shape of the *ortho* regioisomers may not

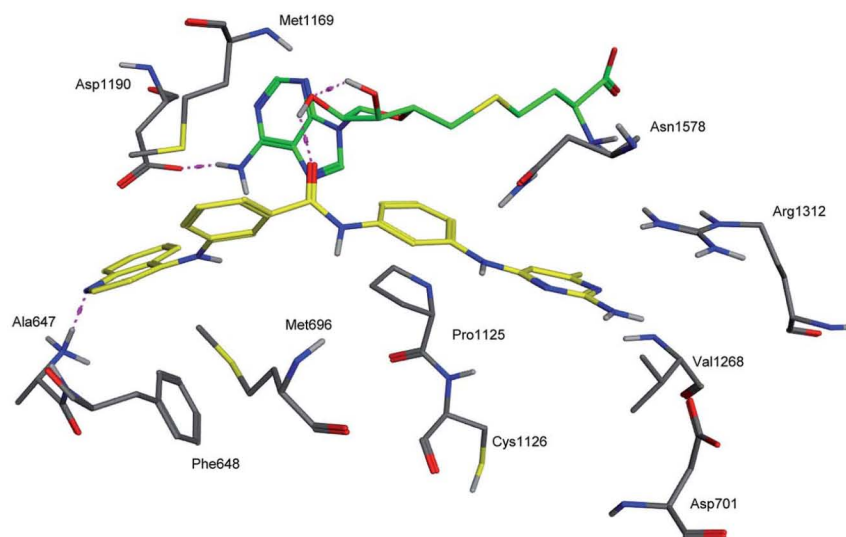


Fig. 7 Docking model of CHEMBL3126646 (carbon atoms in yellow) with DNMT1. The position of the co-crystal SAH is displayed (carbon atoms in green). Selected residues of the binding pocket are shown. Hydrogen bond interactions are in dashed lines. Note the predicted hydrogen bond interaction between the carbonyl oxygen of CHEMBL3126646 and the O<sub>2</sub> oxygen of SAH. Non-polar hydrogens are hidden for clarity.



adopt the extended conformation required to stabilize inhibitory linker domain. Flexible alignment of SGI-1027 analogues with the most active compound CHEMBL3126646 (Fig. 8) clearly shows the very different shape of the more active *meta/meta* and other non-*ortho* regioisomers (Fig. 8A) as compared to the inactive *ortho* regioisomers of SGI-1027 (Fig. 8B). Docking of the *ortho* containing compounds with DNMT1 (data not shown) showed the loss of the interaction with the co-factor also highlighting this key interaction of CHEMBL3126646.

Preliminary regular and pharmacophore-constrained docking studies of the eight compounds related to CHEMBL3126646 (Fig. 5) were conducted with a crystallographic structure of DNMT1. The docking poses were post-processed with PLIFs as detailed in the Methods section. Results are summarized in Fig. S16a.† In order to explore the protein–ligand contacts that may differentiate ‘active’ from ‘inactive’ compounds, the significance analysis implemented in MOE was performed. For this analysis we considered as “active” a compound with  $\text{pIC}_{50} > 5$ . Fig. S16b† shows that there are not statistically significant differences that might distinguish active from inactive molecules. This reflects the fact that *ortho* regioisomers are not unable of stretching to the required extent, but the energy

necessary to do so is higher, mainly due to their intermolecular interactions. Further computational analyses are required to test this hypothesis (see below section of Future directions).

### Insights into the structure-based optimization of lead molecules

The structure-based interpretation of the activity cliff generators associated with CHEMBL3126646 leads to strategies to further optimize the affinity with DNMT1 and possible the biological activity. For example, addition of cationic moieties at both sides of the molecule would provide the structure with stronger ionic interactions. In addition, a hydrogen bond may be more easily formed with Asp701 in the CXXC domain if a further small elongation of the molecule is produced by adding a carbon or an aromatic ring into the structure. It remains to conduct additional molecular modeling analysis of the designed structures to further guide the structure-based optimization of quinolone-based inhibitors.

## Conclusions

Analysis of the distribution in chemical space of 280 compounds tested as IDNMT1 readily revealed two well-defined groups of structures: SAM-analogues and non-nucleoside compounds. Local SAR analysis showed that the two clusters have different activity landscapes. Molecules similar to the cofactor SAM have a heterogeneous landscape with the presence of deep activity cliffs *i.e.*, similar molecules with large potency difference. In sharp contrast, non-nucleoside compounds have a smoother SAR with few shallow activity cliffs and fewer deep activity cliffs. The significance of this observation is that, at least in principle, almost any active small non-SAM-like molecule in this data set can be used as a query in similarity-based virtual screening. Also, in general, the non-nucleoside data set can be the starting point to develop predictive models. Of course, these conclusions depend on the current contents of ChEMBL. As the coverage of the chemical space of non-SAM-like compounds increases the corresponding landscape may change and more activity cliffs may emerge.

The structural interpretation of the activity cliffs indicated that SAM-related analogues contain several pharmacophoric interactions that are substantial to determining its potency. Therefore, even small changes in its structure may produce deep activity cliffs. Hence, SAM-analogues may not be suitable for classical predictive approaches that assume linear relationships.

Structure-based analysis of the most relevant non-nucleoside activity cliff generator, a regioisomer of SGI-1027 developed recently, supported the hypothesis that this type of molecules may act through a stabilization of the auto-inhibitory linker domain of DNMT1. Results of the docking model are in agreement with the SAR of the deepest activity cliffs involving CHEMBL3126646. Results are also in agreement with the biochemical analysis showing that CHEMBL3126646 is not a competitive inhibitor of the co-factor.

During the course of this work we concluded that density SAS maps are convenient graphical representations that enhance

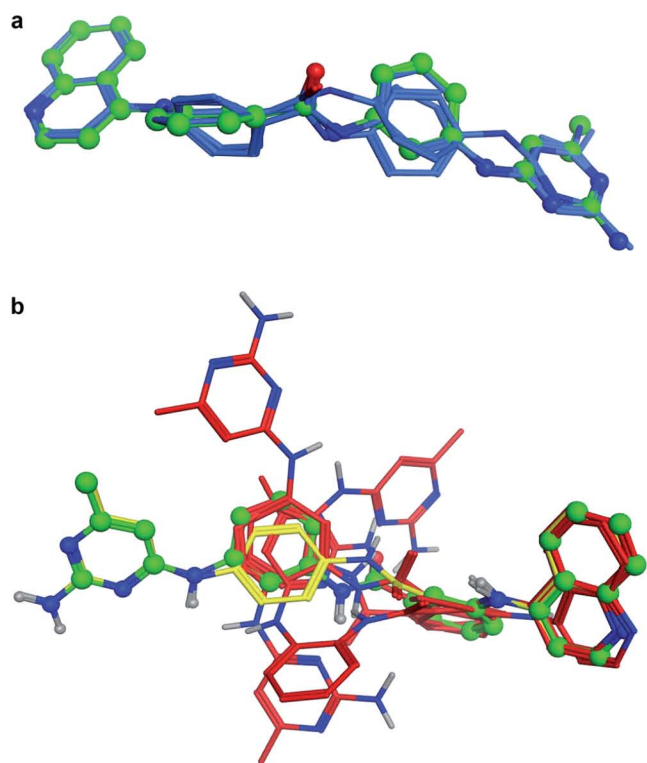


Fig. 8 Flexible alignment of regioisomers of SGI-1027 (chemical structures are shown in Fig. 5) with the best docked pose of CHEMBL3126646 (balls and sticks and carbon atoms in green). (a) Non-*ortho* regioisomers (carbon atoms in blue). (b) *ortho* regioisomers [CHEMBL3126644, CHEMBL3126647, CHEMBL3126649 and *ortho/ortho* SGI-1027 analog (not registered in ChEMBL) (carbon atoms in red) and CHEMBL3126648 (*ortho/para*) (carbon atoms in yellow)]. Note the alignment different in the red molecules and the different orientation of the carbonyl oxygen in the both the red and the yellow molecules, which is not the case in (a).

the interpretation of the SAS maps. It was also highlighted the convenience of performing 'activity landscape sweeping' before the analysis of the activity landscape of a data set. The activity landscape sweeping presented in this work led to the exploration of local activity landscapes that provided interpretable SAR results and provided insights for the structure-based optimization of lead compounds as IDNMT1.

### Future directions

As part of this chemoinformatics work, we focused on the initial docking and molecular modeling of active compounds forming the most representative activity cliffs. A next logical step of this study is to conduct the molecular modeling of all the active molecules including those with a smooth SAR. Similarly, comprehensive molecular modeling studies should be conducted to explain, at the molecular level, other activity cliffs (e.g., non-quinolone based) identified in this work. As part of these studies, induced-fit docking and/or other methods that consider protein flexibility should be used. These studies are ongoing in our group and will be reported in due course. It remains to explore the similarity cliffs (scaffold hops) that emerged from this work. Finally, other perspective is to develop predictive models (such as QSAR) for non-nucleoside compounds. During writing of this manuscript, a paper reporting QSAR models of IDNMT1 was published.<sup>46</sup>

## List of abbreviations

3D	Three-dimensional
ALM	Activity landscape modeling
DNMT	DNA methyltransferases
ECFP	Extended connectivity fingerprints
HTS	High-throughput screening
IDNMT	Inhibitor of DNA methyltransferase
MDS	Myelodysplastic syndrome
MOE	Molecular operating environment
PCA	Principal component analysis
PDB	Protein data bank
PLIF	Protein ligand interaction fingerprint
QSAR	Quantitative structure–activity relationships
RMSD	Root-mean-square deviation
SAH	S-Adenosyl-L-homocysteine
SAM	S-Adenosyl-L-methionine
SAR	Structure–activity relationships
SAS maps	Structure–activity-similarity maps

## Acknowledgements

This work was funded by the National Autonomous University of Mexico (UNAM), grant PAIP 5000-9163 to JLMF.

## References

- C. H. Waddington, *Int. J. Epidemiol.*, 2012, **41**, 10–13.
- S. Knapp and H. Weinmann, *ChemMedChem*, 2013, **8**, 1885–1891.

- K. D. Robertson, *Oncogene*, 2001, **20**, 3139–3155.
- A. Jeltsch, *ChemBioChem*, 2002, **3**, 274–293.
- J. L. Medina-Franco, J. Yoo and A. Dueñas-Gonzalez, in *Epigenetic Technological Applications*, ed. Y. G. Zheng, Elsevier, 2015, ch. 13, pp. 265–290.
- E. J. B. Derissen, J. H. Beijnen and J. H. M. Schellens, *Oncologist*, 2013, **18**, 619–624.
- C. Gros, J. Fahy, L. Halby, I. Dufau, A. Erdmann, J.-M. Gregoire, F. Ausseil, S. Vispé and P. B. Arimondo, *Biochimie*, 2012, **94**, 2280–2296.
- J. L. Medina-Franco, O. Méndez-Lucio, J. Yoo and A. Dueñas, *Drug Discovery Today*, 2015, **20**, 569–577.
- 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.
- G. M. Maggiora, *J. Chem. Inf. Model.*, 2006, **46**, 1535.
- R. Guha and J. H. van Drie, *J. Chem. Inf. Model.*, 2008, **48**, 1716–1728.
- A. Golbraikh, E. Muratov, D. Fourches and A. Tropsha, *J. Chem. Inf. Model.*, 2014, **54**, 1–4.
- M. Cruz-Monteagudo, J. L. Medina-Franco, Y. Pérez-Castillo, O. Nicolotti, M. N. D. S. Cordeiro and F. Borges, *Drug Discovery Today*, 2014, **19**, 1069–1080.
- J. Medina-Franco, G. Navarrete-Vázquez and O. Méndez-Lucio, *Future Med. Chem.*, 2015, **7**, 1197–1211.
- J. J. Naveja and J. L. Medina-Franco, *Expert Opin. Drug Discovery*, 2015, DOI: 10.1517/17460441.2015.1073257, in press.
- Molecular Operating Environment (MOE), version 2010.10*, Chemical Computing Group Inc., Montreal, Quebec, Canada, available at <http://www.chemcomp.com>.
- T. Sander, J. Freyss, M. von Korff and C. Rufener, *J. Chem. Inf. Model.*, 2015, **55**, 460–473.
- D. Rogers, R. D. Brown and M. Hahn, *J. Biomol. Screening*, 2005, **10**, 682–686.
- P. Jaccard, *Bull. Soc. Vaudoise Sci. Nat.*, 1901, **37**, 547–579.
- J. L. Medina-Franco and G. M. Maggiora, in *Chemoinformatics for Drug Discovery*, ed. J. Bajorath, John Wiley & Sons, Inc., 2014, ch. 15, pp. 343–399.
- P. Willett, *J. Chem. Inf. Model.*, 2013, **53**, 1–10.
- J. L. Medina-Franco, K. Martínez-Mayorga, M. A. Giulianotti, R. A. Houghten and C. Pinilla, *Curr. Comput.-Aided Drug Des.*, 2008, **4**, 322–333.
- R. Kraft, A. Kahn, J. L. Medina-Franco, M. L. Orłowski, C. Baynes, F. Lopez-Vallejo, K. Barnard, G. M. Maggiora and L. L. Restifo, *Dis. Models & Mech.*, 2013, **6**, 217–235.
- H. Wickham, *ggplot2: Elegant Graphics for Data Analysis*, Springer, New York, 2009.
- V. Shanmugasundaram and G. M. Maggiora, Presented in part at the 222nd ACS National Meeting, Chicago, IL, United States, August 26–30, 2001.
- J. Pérez-Villanueva, R. Santos, A. Hernández-Campos, M. A. Giulianotti, R. Castillo and J. L. Medina-Franco, *Bioorg. Med. Chem.*, 2010, **18**, 7380–7391.

- 27 A. Yongye, K. Byler, R. Santos, K. Martínez-Mayorga, G. M. Maggiora and J. L. Medina-Franco, *J. Chem. Inf. Model.*, 2011, **51**, 1259–1270.
- 28 F. Renee, L. M. Travis, R. G. Santos, A. Morales, A. Nefzi, G. S. Welmaker, J. L. Medina-Franco, M. A. Giulianotti, R. A. Houghten and L. N. Shaw, *J. Med. Chem.*, 2015, **58**, 3340–3355.
- 29 J. L. Medina-Franco, *J. Chem. Inf. Model.*, 2012, **52**, 2485–2493.
- 30 J. L. Medina-Franco, K. Martínez-Mayorga, A. Bender, R. M. Marín, M. A. Giulianotti, C. Pinilla and R. A. Houghten, *J. Chem. Inf. Model.*, 2009, **49**, 477–491.
- 31 O. Mendez-Lucio, J. Perez-Villanueva, R. Castillo and J. L. Medina-Franco, *Mol. Inf.*, 2012, **31**, 837–846.
- 32 J. Song, O. Rechkoblit, T. H. Bestor and D. J. Patel, *Science*, 2011, **331**, 1036–1040.
- 33 S. C. Brewerton, *Curr. Opin. Drug Discovery Dev.*, 2008, **11**, 356–364.
- 34 O. Méndez-Lucio, A. J. Kooistra, C. D. Graaf, A. Bender and J. L. Medina-Franco, *J. Chem. Inf. Model.*, 2015, **55**, 251–262.
- 35 B. Seebeck, M. Wagener and M. Rarey, *ChemMedChem*, 2011, **6**, 1630–1639.
- 36 J. L. Medina-Franco, K. Martínez-Mayorga, A. Bender and T. Scior, *QSAR Comb. Sci.*, 2009, **28**, 1551–1560.
- 37 D. Stumpfe, Y. Hu, D. Dimova and J. Bajorath, *J. Med. Chem.*, 2014, **57**, 18–28.
- 38 P. Iyer, D. Stumpfe, M. Vogt, J. Bajorath and G. M. Maggiora, *Mol. Inf.*, 2013, **32**, 421–430.
- 39 A. Erdmann, L. Halby, J. Fahy and P. B. Arimondo, *J. Med. Chem.*, 2014, **58**, 2569–2583.
- 40 J. Datta, K. Ghoshal, W. A. Denny, S. A. Gamage, D. G. Brooke, P. Phiasivongsa, S. Redkar and S. T. Jacob, *Cancer Res.*, 2009, **69**, 4277–4285.
- 41 S. Valente, Y. W. Liu, M. Schnekenburger, C. Zwergel, S. Cosconati, C. Gros, M. Tardugno, D. Labella, C. Florean, S. Minden, H. Hashimoto, Y. Q. Chan, X. Zhang, G. Kirsch, E. Novellino, P. B. Arimondo, E. Miele, E. Ferretti, A. Gulino, M. Diederich, X. D. Cheng and A. Mai, *J. Med. Chem.*, 2014, **57**, 701–713.
- 42 J. Husby, G. Bottegoni, I. Kufareva, R. Abagyan and A. Cavalli, *J. Chem. Inf. Model.*, 2015, **55**, 1062–1076.
- 43 A. Kabro, H. Lachance, I. Marcoux-Archambault, V. Perrier, V. Dore, C. Gros, V. Masson, J. M. Gregoire, F. Ausseil, D. Cheishvili, N. B. Laulan, Y. St-Pierre, M. Szyf, P. B. Arimondo and A. Gagnon, *MedChemComm*, 2013, **4**, 1562–1570.
- 44 S. Castellano, D. Kuck, M. Viviano, J. Yoo, F. López-Vallejo, P. Conti, L. Tamborini, A. Pinto, J. L. Medina-Franco and G. Sbardella, *J. Med. Chem.*, 2011, **54**, 7663–7677.
- 45 J. Yoo, S. Choi and J. L. Medina-Franco, *PLoS One*, 2013, **8**, e62152.
- 46 W. Maldonado-Rojas, J. Olivero-Verbel and Y. Marrero-Ponce, *J. Mol. Graphics Modell.*, 2015, **60**, 43–54.

# Chemical space, diversity and activity landscape analysis of estrogen receptor binders†

J. Jesús Naveja,<sup>id abc</sup> Ulf Norinder,<sup>de</sup> Daniel Mucs,<sup>df</sup> Edgar López-López<sup>ag</sup> and José L. Medina-Franco<sup>id \*a</sup>

Understanding the structure–activity relationships (SAR) of endocrine-disrupting chemicals has a major importance in toxicology. Despite the fact that classifiers and predictive models have been developed for estrogens for the past 20 years, to the best of our knowledge, there are no studies of their activity landscape or the identification of activity cliffs. Herein, we report the first SAR of a public dataset of 121 chemicals with reported estrogen receptor binding affinities using activity landscape modeling. To this end, we conducted a systematic quantitative and visual analysis of the chemical space of the 121 chemicals. The global diversity of the dataset was characterized by means of Consensus Diversity Plot, a recently developed method. Adding pairwise activity difference information to the chemical space gave rise to the activity landscape of the data set uncovering a heterogeneous SAR, in particular for some structural classes. At least eight compounds were identified with high propensity to form activity cliffs. The findings of this work further expand the current knowledge of the underlying SAR of estrogenic compounds and can be the starting point to develop novel and potentially improved predictive models.

Received 12th September 2018  
 Accepted 5th November 2018

DOI: 10.1039/c8ra07604a

[rsc.li/rsc-advances](http://rsc.li/rsc-advances)

## 1. Introduction

Endocrine disrupting chemicals (EDCs) affect normal hormonal action related to the endocrine system of humans and other organisms.<sup>1,2</sup> These chemicals can produce a vast range of adverse effects including developmental, reproductive, neurological, and immune system related effects. EDCs act through endocrine system pathways, including those related to estrogens, androgens, and thyroid hormones. Many investigations to derive robust and predictive quantitative structure–activity relationship (QSAR) models for EDCs interacting with endocrine hormone receptors, and in particular the estrogen receptor (ER), have been performed over the past 15 years.<sup>3–13</sup> Xenoestrogens are known to have large chemical

diversity including, for instance, estrogen diethylstilbestrol, polychlorinated biphenyls, alkylphenols, phthalates, and parabens, among others.<sup>14</sup> Several structure–activity relationship (SAR) analysis and predictive models of estrogens have been developed over the past years and commented on extensively.<sup>14</sup> However, there are no reports on the activity landscape of the EDCs.

One of the consistent manners to characterize the SAR of compound data sets is through the systematic pairwise comparison of the structure with the activity. This approach termed “activity landscape modeling”<sup>15–17</sup> is based upon the similarity principle of chemical data sets, *i.e.*, structurally similar compounds have similar activity values. Activity landscape modelling identifies activity cliffs *i.e.*, pairs of compounds with high structure similarity but large potency difference.<sup>18</sup> Depending on the scope, activity cliffs can have beneficial or detrimental consequences in many cases of study because they are major exceptions to the similarity principle. On one hand, activity cliffs challenge the development of many predictive models founded on the similarity principle. On the other hand, activity cliffs lead directly to key structural information that influence the property.<sup>19</sup> Over the past few years, several quantitative and/or visual approaches have been published to get the profile of the activity landscape of compounds with one<sup>20</sup> or several endpoints.<sup>21</sup> Of note, to the best of our knowledge, these approaches have not been used to explore the property landscapes of estrogenic binding compounds despite their major importance.

<sup>a</sup>Department of Pharmacy, School of Chemistry, Universidad Nacional Autónoma de México, Mexico City, 04510, Mexico. E-mail: [medinajl@unam.mx](mailto:medinajl@unam.mx); [jose.medina.franco@gmail.com](mailto:jose.medina.franco@gmail.com); Tel: +52-55-5622-3899 ext. 44458

<sup>b</sup>PECEM, Faculty of Medicine, Universidad Nacional Autónoma de México, Mexico City, 04510, Mexico

<sup>c</sup>Department of Life Science Informatics, Bonn-Aachen International Center for Information Technology, University of Bonn, Bonn, 53113, Germany

<sup>d</sup>Swetox, Karolinska Institutet, Unit of Toxicology Sciences, SE-151 36 Södertälje, Sweden

<sup>e</sup>Department of Computer and Systems Sciences, Stockholm University, Box 7003, SE-164 07 Kista, Sweden

<sup>f</sup>Unit of Work Environment Toxicology, Institute of Environmental Medicine, Karolinska Institutet, Stockholm, Sweden

<sup>g</sup>Medicinal Chemistry Laboratory, University of Veracruz, Veracruz, Mexico

† Electronic supplementary information (ESI) available. See DOI: 10.1039/c8ra07604a





Because all pairwise comparisons can lead to large amounts of structure–activity information difficult to mine and visualize, an approach called ‘activity landscape sweeping’ was developed. This is a dissection of the global activity landscape *i.e.*, global SAR, into smaller but more structural interpretable local landscapes *i.e.*, local SARs.<sup>22</sup>

Herein we report an activity landscape study of 121 chemicals with measured ER binding affinities. One of the main goals was to identify activity cliffs and ‘activity cliff generators’,<sup>23</sup> *i.e.*, compounds that are frequently associated with cliffs. The activity landscape sweeping approach was implemented to further understanding the activity landscape of particular groups of compounds. To this end, an analysis of the chemical space, diversity and clustering of the compounds was conducted before doing the activity landscape modeling.

## 2. Materials and Methods

### 2.1. Data sets

We focused the study on a set of 121 molecules with published values of measured binding affinities.<sup>14</sup> This is a set of experimentally active estrogens of different structural families including steroids, DES-like, phytoestrogens, diphenylmethanes, biphenyls and phenols. The chemical structures were prepared and standardized with MOE 2016, including manual curation to avoid duplicate entries and structural errors, as well as salt removal, charges neutralization and keeping only the largest fragment if more than one molecule was present.

### 2.2. Molecular representations

Standard 2D chemical features were studied to characterize the chemical space. The analysis focused on molecular fingerprints (ECFP4, *i.e.*, Extended Connectivity Fingerprints diameter 4),<sup>24</sup> molecular scaffold (as computed using the Bemis and Murcko approach<sup>25</sup>), and six physicochemical properties (PCP) of pharmaceutical relevance, namely: octanol/water partition coefficient (Slog *P*), molecular weight (MW), topological polar surface area (TPSA), number of rotatable bonds (RB), number of hydrogen bond donors and number of hydrogen bond acceptors (HBD/HBA). The molecular fingerprints, scaffolds and properties were computed with KNIME<sup>26</sup> RDkit and CDK nodes.<sup>27</sup>

### 2.3. Chemical space and clustering

In order to aid the activity landscape modeling of the 121 chemicals and explore local SARs, we conducted an analysis of the chemical space. It has been previously shown that principal component analysis (PCA) and *k*-means clustering applied to structural similarity data using ECFP4 is a useful approach for finding and visualizing different subsets of compounds that are structurally related, for which it is feasible to find local SAR differences.<sup>22</sup> Herein this approach was followed, and by direct inspection of the first 3 principal components (55.7% of variance) we concluded that at least four clusters could be defined. Clustering was performed with

*k*-means on the first 7 principal components (72.7% of the variance). To further characterize these subsets, we analyzed their structural diversity through the molecular scaffolds (computed as described in Section 2.1).

### 2.4. Global diversity

The ‘global’ or total diversity of the entire compound data set and each individual cluster was evaluated using Consensus Diversity Plots.<sup>28</sup> Briefly, these are low dimensional graphs that are aimed to integrate different but complementary measures of diversity of databases. Typically, Consensus Diversity Plots represent fingerprint, scaffold, property diversity and size *i.e.*, number of compounds in different datasets. The position of the data points in the plot, the color and size provide a quick assessment of the relative diversity of data sets. Further details of these plots and their use are elaborated elsewhere<sup>28,29</sup> As discussed in the Results and discussion section, it would be expected that the clusters tend towards lower fingerprint-based diversity than the original data, given that they are being put together by this very criterion.

### 2.5. Activity landscape modeling

Activity landscape analysis was done for the data set with all the 121 compounds and for each of the clusters (4 in total) identified during the analysis of the chemical space (Section 2.3). The activity landscape analysis was performed using Structure–Activity Similarity (SAS) map which is one of the first approaches in order to perform activity landscape modeling and identify activity cliffs.<sup>30</sup> A schematic representation of a SAS map is presented in Fig. 1. Briefly, a SAS map is a two-dimensional graph where pairwise structure and activity similarity of usually all pairwise comparisons of a data set are plotted. The structure similarity is represented on the X-axis and the activity difference (or activity similarity) is plotted on the Y-axis. In this work, the structure similarity was computed using ECFP4 fingerprints and the Tanimoto coefficient. The activity difference was computed as the absolute value of the activity difference initially expressed in relative binding affinity units (RBA), obtained by means of dividing the determined potency (IC<sub>50</sub>) by the IC<sub>50</sub> of 17β-estradiol.<sup>14</sup> Information from the activity landscape was contrasted with the diversity analysis, to find whether some areas of the chemical space are more susceptible to form activity cliffs. As presented in Fig. 1A, activity cliffs are identified in the top-right quadrant of the SAS map that identifies pairs of molecules with high structure similarity but large activity difference.

### 2.6. Activity cliffs and generators

As mentioned in the Introduction, activity cliff generators are molecules frequently identified in the activity cliff region of the activity landscape.<sup>23</sup> In other words, activity cliff generators are molecules that are commonly found in activity cliff pairs. In this work, compounds involved in at least five activity cliffs were selected as activity cliff generators and subject to further analysis. Direct analysis and interpretation of these activity cliffs generators is expected to yield insights into the relevant



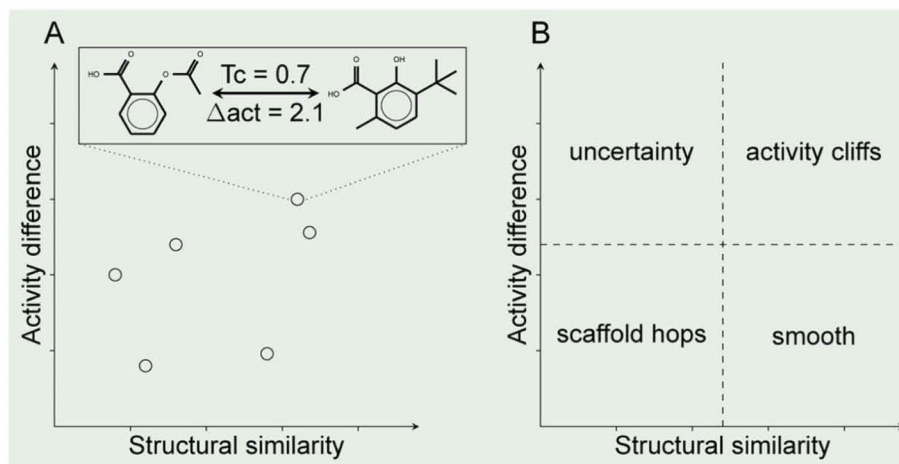


Fig. 1 General form of a Structure–Activity Similarity (SAS) map. (A) Each data point represents a pair-wise comparison. Hypothetical distribution five pairs of compounds. The two example chemical structures illustrate an activity cliff: compounds with similar chemical structures but large activity difference *e.g.*, larger than two potency units. (B) Four major regions that can be roughly identified in a SAS map. Each quadrant is labeled with the overall type of landscape.

features providing estrogenic activity. All analyses were done using KNIME version 3.5.3 and its corresponding RDkit and CDK nodes.

### 3. Results and discussion

Results are presented and discussed in two major parts. In the first part an analysis of the chemical space diversity and content of the data set of the 121 compounds is described (Subsections 3.1 and 3.2). The second part (Subsection 3.3) addresses the activity landscape analysis that was developed based on the analysis of the chemical space.

#### 3.1. Chemical space and clustering

Fig. 2 shows a visual representation of the chemical space of the 121 compounds using PCA based on six drug-like properties of

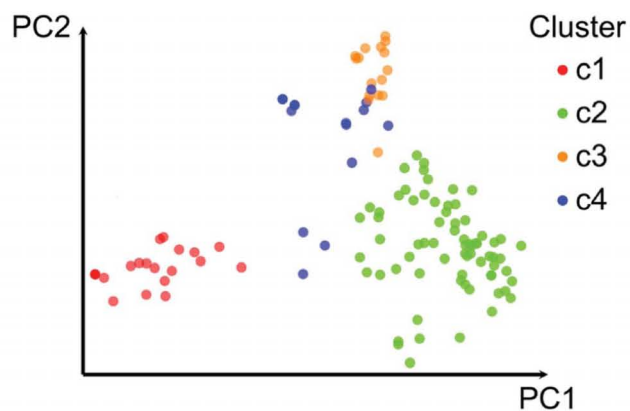


Fig. 2 Visual representation of the chemical space of the data set with 121 compounds. The visual representation was generated with principal component analysis of six drug-like physicochemical properties. The first two principal components account for 43.7% of the variance. Data points are color-coded by the cluster each compound belongs based on pairwise structure similarity computed with ECFP4/Tanimoto. Clustering was performed with *k*-means on the first 7 principal components (72.7% of the variance).

pharmaceutical relevance. The first three principal components captured 55.7% of the variance. As described on the Methods section, the 121 compounds were further clustered into four groups based on the pairwise structure similarity computed with ECFP4 fingerprints and the Tanimoto coefficient. In Fig. 2 compounds (data points) are color-coded by the cluster number of each compound. Table 1 summarizes the number of compounds in each cluster. Overall, Fig. 2 shows a reasonable good qualitative relationship between the PCP and fingerprint-based similarity. In other words, compounds with similar PCP also have similar chemical structures as captured by the ECFP4/Tanimoto combination.

In order to further interpret the type of compounds present in each cluster, the main chemical scaffolds (computed as described in the Methods section) present in each cluster were identified. Fig. 3 shows representative Bemis and Murcko scaffolds. Cluster 1 with 20 (17%) compounds is characterized by the presence of steroidal scaffolds. Cluster 2 with 70 (58%) compounds is the largest group: it contains 20 molecules that share the ubiquitous benzene scaffold, compounds related to the DES, hexestrol and tetraphenylethylene derivatives. Cluster 3 with 16 (13%) compounds contain flavones. Finally, cluster 4 has 15 (12%) compounds containing flavanones,

Table 1 Total diversity profile of compounds in each of the four clusters (sub sets of compounds; local SAR) and for ALL compounds (global SAR)<sup>a</sup>

Cluster	No. cpds	Median MACCS keys/Tanimoto	AUC	Median PCP
1	20	0.37	0.64	2.99
2	70	0.42	0.72	3.12
3	16	0.48	0.72	2.99
4	15	0.83	0.71	3.18
ALL	121	0.40	0.77	2.75

<sup>a</sup> AUC: area under the curve. PCP: physicochemical properties.





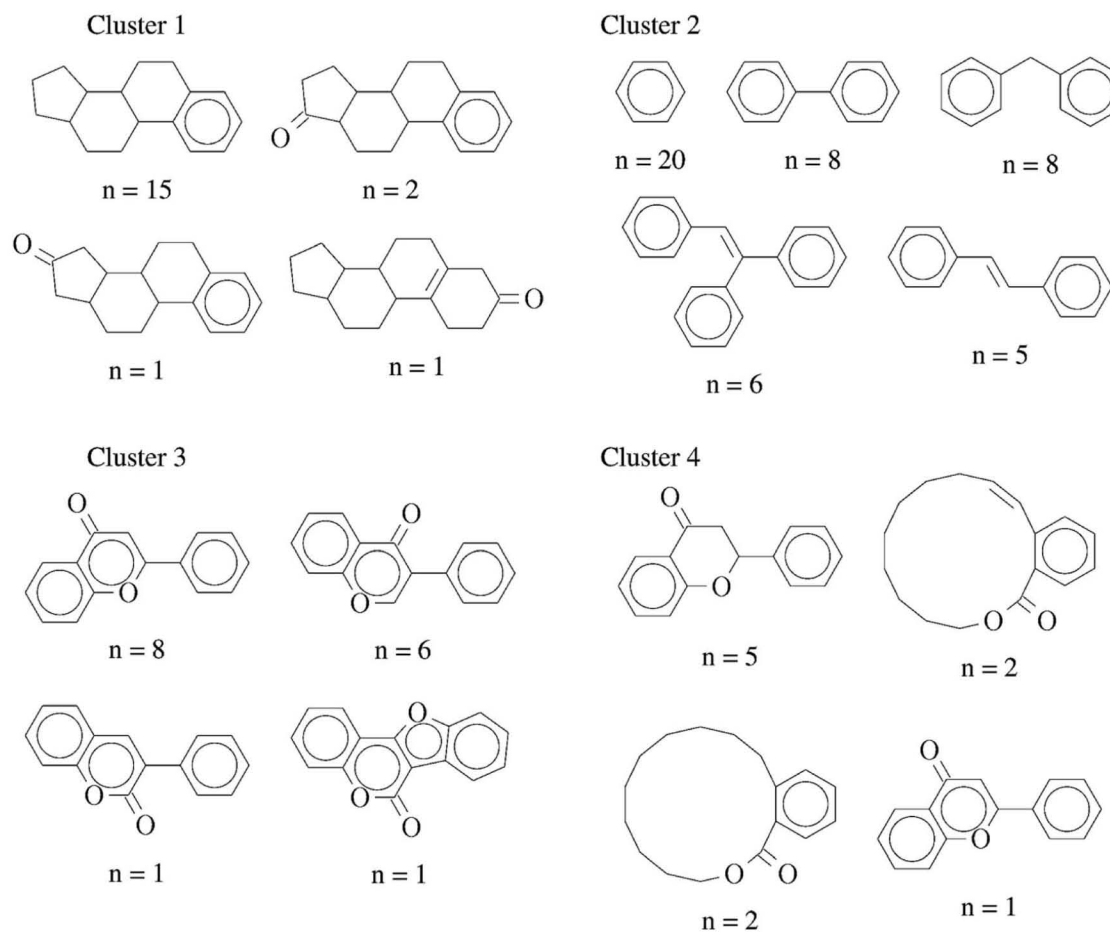


Fig. 3 Representative chemical scaffolds found in each of the four clusters. The number of compounds ( $n$ ) containing each cluster is indicated.

mycoestrogens and other scaffolds. We want to emphasize that the clustering was performed based on molecular fingerprints considering the entire chemical structures.

### 3.2. Global diversity

Fig. 4 shows the Consensus Diversity Plot comparing the relative global diversity of each cluster (or subset described in Section 3.1) as compared to the diversity of the entire data set. In this plot, each data point represents one compound cluster. As described in the Methods section, the fingerprint-based diversity of each cluster is represented on the X-axis, in this case measured as the median MACCS keys (166 bits) and Tanimoto similarity of the cluster. Hence, data points to the left have, in general, lower molecular similarity *e.g.*, larger diversity. The scaffold diversity is represented on the Y-axis as measured by the area under the curve (AUC) of the scaffold recovery curve. Thus, clusters at the bottom of the plot with lower AUC values have higher scaffold diversity. Of note, as described in detail elsewhere, in a scaffold recovery curve the minimum value of AUC is 0.5 that means that a compound data set has the largest scaffold diversity: each molecule would have their own scaffold.<sup>31</sup> The diversity based on PCP is represented with a continuous color scale from less diverse (red) to most diverse (green). Finally, the size of the data point is a relative measure of the

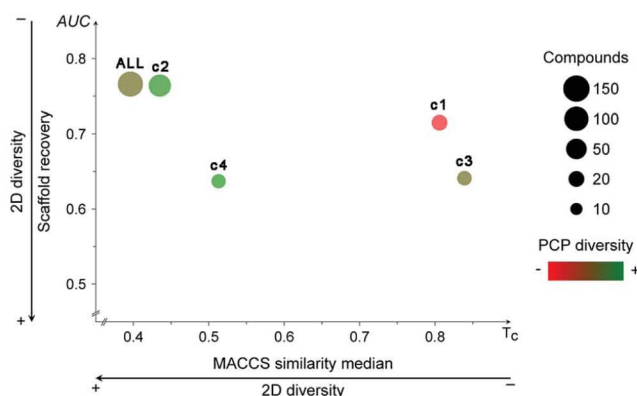


Fig. 4 Consensus Diversity Plot comparing the global diversity of the four different clusters and the entire data set (ALL). Each cluster is represented with a data point. The structural diversity (X-axis) is defined as the median Tanimoto coefficient of MACCS keys fingerprints. The scaffold diversity (Y-axis) is defined as the area under the corresponding scaffold recovery curve. The diversity based on physicochemical properties (PCP) was defined as the Euclidean distance of six auto-scaled properties (Slog  $P$ , TPSA, AMW, RB, HBD, and HBA) and is shown as the filling of the data points using a continuous color scale. The relative number of compounds is represented with a different size of the data points (larger clusters are represented with larger data points).



number of compounds in each cluster *e.g.*, smaller clusters have fewer number of molecules.

Fig. 4 indicates that the entire data set (labeled as “ALL”) has a relative large fingerprint diversity but a low scaffold diversity. Cluster 2 (58% of compounds) is almost as diverse as the entire data set in terms of fingerprints and scaffolds. In contrast, cluster 4 (12% of compounds) has the relative largest combined scaffold and fingerprint diversity while cluster 1 is the least diverse with the overall lowest scaffold and fingerprint diversity. This observation is consistent with the type of molecules present in cluster 1, most of them have a steroid scaffold (a relative large scaffold that should be related to the entire diversity-*vide supra*). Also in contrast, compounds in cluster 2 have a small core scaffold and it would be expected that the fingerprint-diversity is influenced by the side chains. Regarding the diversity in terms of PCP, the Consensus Diversity Plot in Fig. 4 also highlights the opposite diversity of compounds in clusters 2 and 3.

### 3.3. Activity landscape analysis

Following the concept of activity landscape sweeping described in the Introduction and Methods, herein we analyzed the landscape for all compounds in the data set and activity landscapes for each of the four clusters. Fig. 5 shows the SAS maps for all compounds and for each of the four clusters. Thus, Fig. 5 represents the “global” and “local” activity landscapes. The SAS maps are colored coded by the density of the data points *i.e.*, density SAS maps. Overall, most of the data points, in particular for ALL compounds and for compounds in cluster 2 are located

in the lower left region of the SAS map *e.g.*, compounds with low molecular similarity (*e.g.*, high diversity), and low activity difference. In general, this result is consistent with the known observation that there are a large number of chemicals with diverse chemical structures but with small variations in ER binding affinity properties. Visual inspection of Fig. 5 also suggests that the activity landscape of compounds in cluster 2 resemble the landscape of the entire data set (ALL). However, a quantitative analysis would provide more insights.

Table 2 summarizes a quantitative characterization of the activity landscape based on the contents of the SAS maps. A key point in the quantitative analysis of the SAS maps is setting the thresholds that define the four major quadrants of the plots *i.e.*, the thresholds used in this study to define high/low/structural similarity (along the X-axis) and high/low activity difference. Several valid approaches have been used to define such thresholds in the SAS maps.<sup>32</sup> Herein, we used a potency difference of two log units in potency difference along the Y-axis. This criterion has been adopted in several studies as a reasonable large potency difference. To define high/low structure similarity we used the median of the distribution of the pairwise similarity values of the 121 compounds plus two standard deviations *i.e.*, the threshold was set to 0.424. Again, another criterion could have been used. Table 2 indicates the total number of pairwise comparisons for ALL and each of the four sets, *i.e.*, the number of data points in the plots. Table 2 also summarize the percentage of compounds in each quadrant (major region of the SAS map as defined in Fig. 1) after setting up the thresholds.

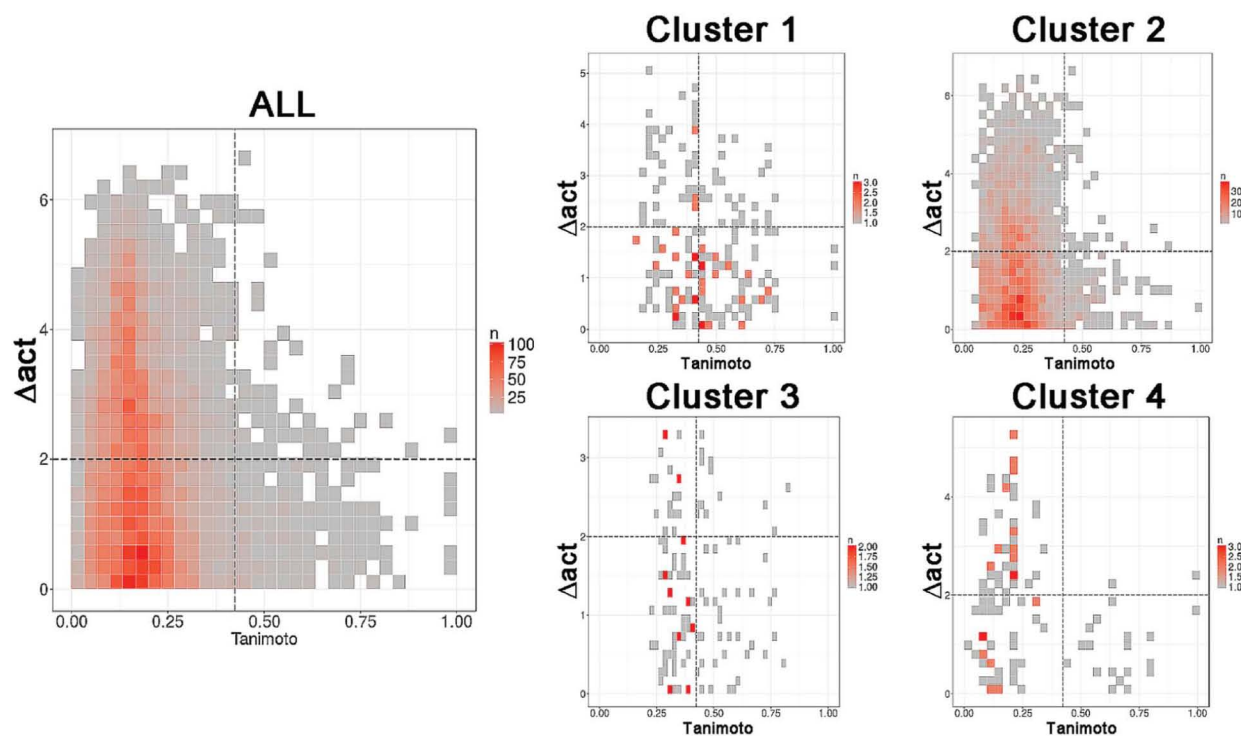


Fig. 5 Density Structure–Activity Similarity (SAS) maps for the entire set (ALL, 121 compounds) *i.e.*, global activity landscape and for each of the four individual clusters *i.e.*, local landscapes. More red areas contain more data points. A quantitative description of the SAS maps is summarized in Table 2.



Table 2 Quantitative analysis of the SAS maps and further analysis of the compounds in each cluster

Cluster	Uncertain <sup>a</sup>	Hops <sup>a</sup>	Cliffs <sup>a</sup>	Smooth <sup>a</sup>	Cliffs/smooth <sup>b</sup>	<i>n</i> <sup>c</sup>	Pairs <sup>d</sup>	<i>X</i> <sub>sim</sub> <sup>e</sup>	<i>n</i> scaff <sup>f</sup>
ALL	41%	54.5%	1%	3.5%	0.286	121	7260	0.192	39
1	21%	28%	13%	38%	0.342	20	190	0.451	5
2	34%	60%	1.2%	4.8%	0.250	70	2415	0.239	21
3	17.5%	44%	11.5%	27%	0.426	16	120	0.417	4
4	44%	37%	1.9%	17.1%	0.111	15	105	0.267	9

<sup>a</sup> Percentage of pairs of compounds in each of the four regions of the SAS map. <sup>b</sup> Ratio of the number of pairs of compounds in the activity cliff/smooth region of the SAS map. <sup>c</sup> Number of compounds in the set (*n*). <sup>d</sup> Number of pairwise comparisons. <sup>e</sup> Median similarity of the compounds in each cluster (*X*<sub>sim</sub>). <sup>f</sup> Number of different Bemis–Murcko scaffolds in each cluster.

The quantitative analysis indicates that compounds in clusters 1 and 3 have the largest proportion of activity cliffs (13% and 11%, respectively). This can also be seen in the SAS maps (Fig. 5) with a relative larger number of data points in the top right region of the plots. In contrast, cluster 2 has the lowest proportion of activity cliffs (1.2%), followed by cluster 4, comparable to the proportion of activity cliffs in the entire data set (1.0%). Interestingly compounds in cluster 1 (with steroid-type scaffolds) and cluster 3 (with several flavones) also have the largest proportion of data points in the smooth region of the landscape (38% and 27%, respectively). Since cluster 1 and 3 have the largest proportion of compounds in both, smooth and activity cliff regions, clusters 1 and 3 have the relative most rough or heterogeneous landscape. Table 2 also indicates that the more diverse compounds (*i.e.*, in cluster 4) have an activity profile similar to the entire dataset (ALL).

**3.3.1. Activity cliff generators and interpretation of the SAR.** In this work we consider an activity generator a molecule found in at least five activity cliff pairs. Based on this criterion, eight compounds were identified as activity cliff generators. Fig. 6 shows the chemical structures of three representative cliff generators: 16beta-ol-16alfa-methyl-3-methyl-estradiol, diethylstilbestrol, and genistein. Examples of activity cliffs pairs for each activity cliff generator are illustrated.

Activity cliffs associated with 16beta-ol-16alfa-methyl-3-methyl-estradiol (Fig. 6A) highlights the relevance and sensitivity of the hydroxyl groups around the estradiol molecule for binding. Of note, all activity cliff pairs in Fig. 6A are steroids with a phenolic ring. The cliffs in the figure points to the high relevance of both hydroxyl groups the 3- and 17beta positions of the molecule as discussed by,<sup>14</sup> a crystallographic structure of the estrogen receptor with 17beta-estradiol indicate that the two

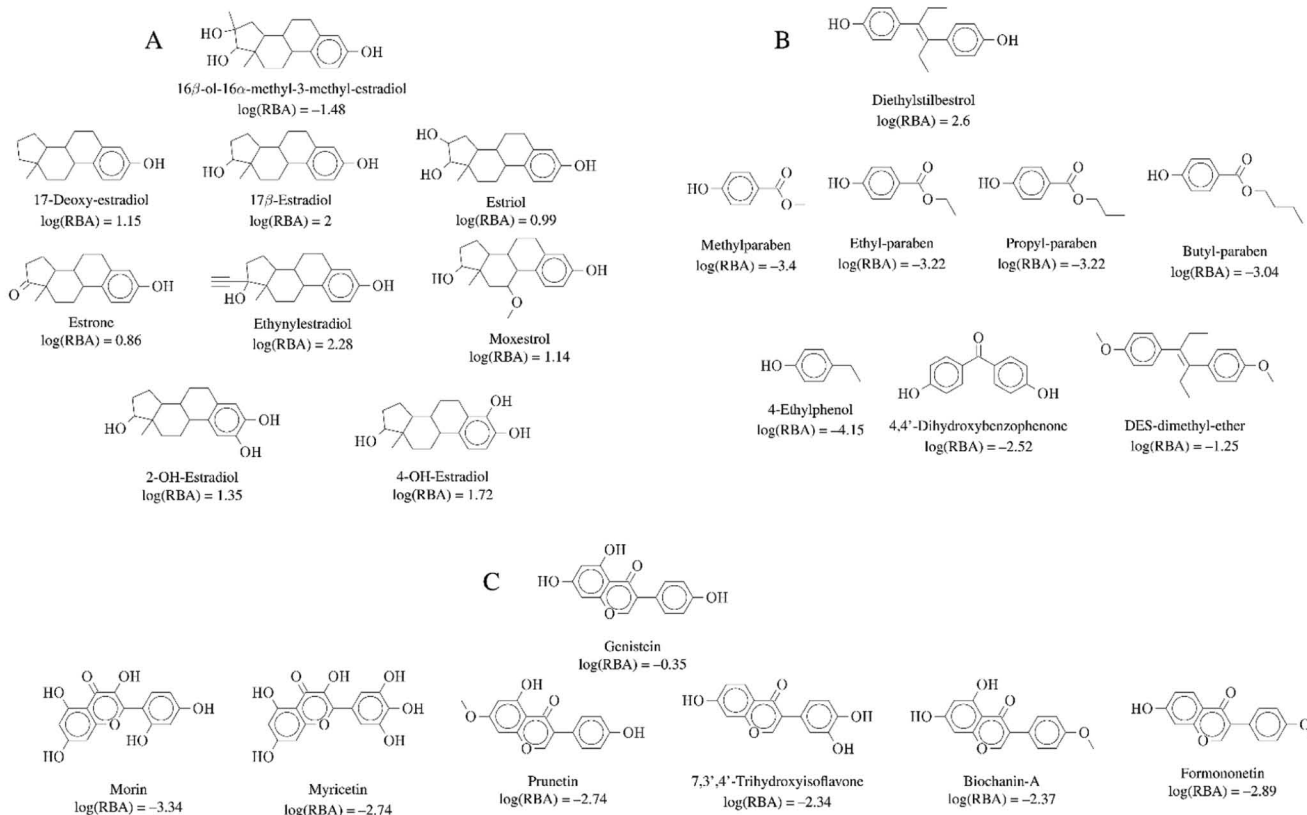


Fig. 6 Representative activity cliff generators and selected pairs of compounds formed with the generators (A) 16beta-ol-16alfa-methyl-3-methyl-estradiol, (B) diethylstilbestrol and (C) genistein. The figure includes the value of the relative binding affinity (RBA) as reported by.<sup>14</sup>



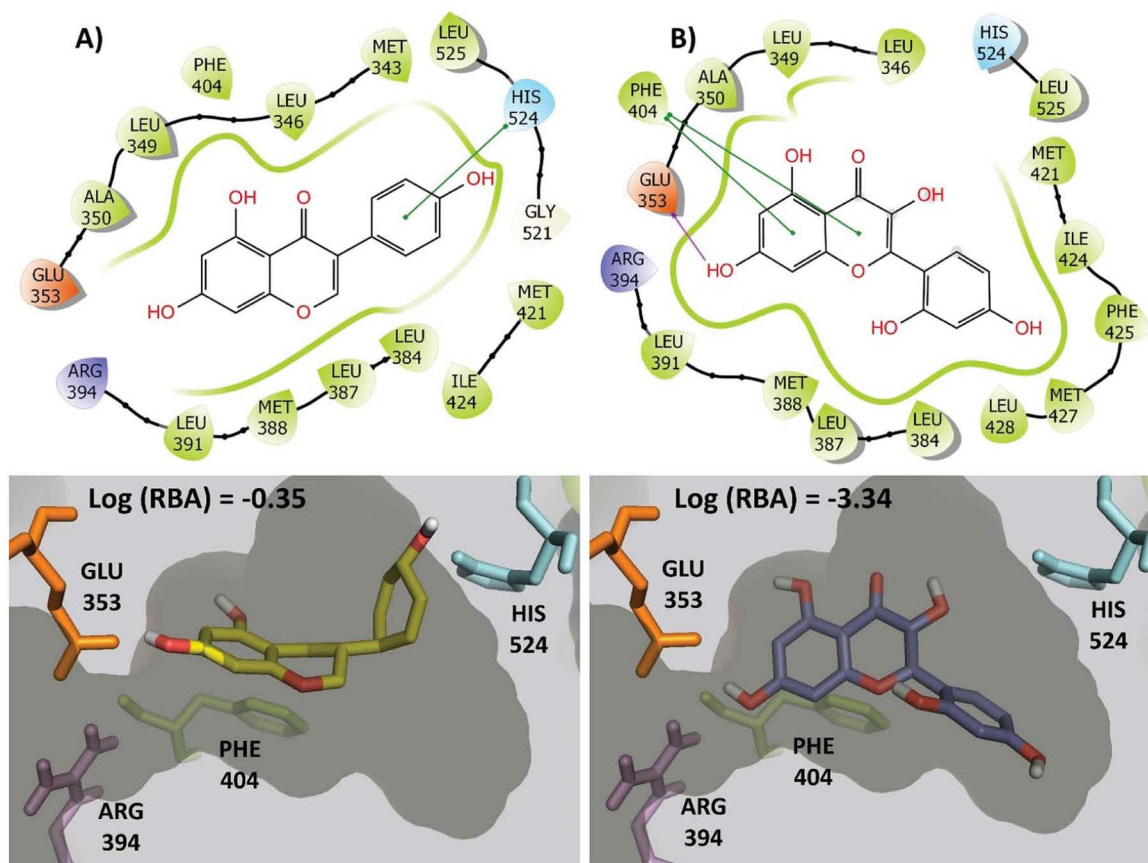


Fig. 7 2D and 3D representation of representative activity cliff generators and selected pairs of compounds with greater difference in activity. (A) Genistein and (B) morin. The figure includes the value of the relative binding affinity (RBA) as reported by.<sup>14</sup>

hydroxyl groups serve as H-bond donors and acceptors at the binding site. The hydroxyl group at the 3-position is more crucial. Similarly, the activity cliffs formed with the generator diethylstilbestrol *e.g.*, which is one of the highest-affinity synthetic estrogens (Fig. 6B), also indicates the critical role of the two symmetrical position of the hydroxyl groups of diethylstilbestrol. The distance of these two groups and rigidity of the molecule (due to the double bond) facilitates the formation of hydrophobic and hydrogen bond interactions of diethylstilbestrol. Finally, activity cliffs formed with the isoflavone genistein (Fig. 6C) further highlights the key position and distance of the two hydroxyl groups at positions 7 and 4' around the isoflavone scaffold that mimic the 4 and 4' hydroxyl groups of diethylstilbestrol.

The large changes in activity can be rationalized from a molecular perspective. This is illustrated in Fig. 7 for the activity cliff generator, genistein and morin (chemical structures also in Fig. 6C). Both compounds have interactions with the side chain of Glu353 through its hydroxyl group at the position 4' of the isoflavone scaffold. In addition, both compounds have conserved pi-pi interactions with the side chain of Phe404. However, genistein makes additional key interactions between a hydroxyl group of the position 7 of the isoflavone scaffold with His524. This key interaction is not formed by morin. Similar conclusions can be reached by two- and three-dimensional representations of the protein-ligand

contacts of the pairs of activity cliffs 16beta-ol-16alpha-methyl-3-methyl-estradiol and estrone (Fig. S1 in the ESI†) and diethylstilbestrol and 4-ethylphenol (Fig. S2 in the ESI†).

As discussed in detail elsewhere,<sup>16</sup> the detection of activity cliffs in compound data sets can be crucial to guide the development of predictive models. Specifically, it is hypothesized that removing activity cliffs from compounds data sets would increase the performance of predictive models that are specially based on the similarity principle, for instance, classical QSAR approaches. For compound data set studied in this work, it would remain to develop and test different predictive models with and without the activity cliffs and assess quantitatively the predictive power.

## 4. Conclusions

Activity landscape analysis of a diverse set of 121 compounds with ER binding affinities revealed an overall heterogeneous SAR with the presence of compounds with high propensity to form activity cliffs. Distinct activity cliff generators are 16beta-ol-16alpha-methyl-3-methyl-estradiol, diethylstilbestrol, and genistein, that represent major structural classes with known ER affinity, namely; a steroid, a DES-like chemical and a phytoestrogen. SAR analysis around these compounds enabled to identify specific structural features associated with a large difference in the ER binding affinities further highlighting the





critical role of two hydroxyl groups for binding recognition to the binding site of the ER. Reported crystallographic structures provide a structure-based context of these cliffs. Chemical space and diversity analysis of the entire data set helped to identify four major groups of compounds, each with a distinct activity landscape *e.g.*, local SAR. Thus, compounds with the more rigid steroid-like scaffold and molecules with a flavone-type scaffold have the most heterogeneous SAR. Global and local activity landscape regions identified in this work with a smooth SAR could be more amenable for developing predictive models. To the best of our knowledge, this is the first activity landscape analysis of compounds with ER binding affinities.

## Conflicts of interest

There are no conflicts to declare.

## Abbreviations

AUC	Area under the curve
DES	Diethylstilbestrol
ECPF4	Extended connectivity fingerprints diameter 4
EDCs	Endocrine-disrupting chemicals
HBA	Hydrogen bond acceptors
HBD	Hydrogen bond donors
MW	Molecular weight
PCA	Principal component analysis
PCP	Physicochemical properties
RB	Number of rotatable bonds
RBA	Relative binding affinity
SAR	Structure–activity relationships
SAS	Structure–activity similarity
Slog <i>P</i>	Octanol/water partition coefficient
TPSA	Topological polar surface area

## Acknowledgements

The research at Swetox (UN, DM) was supported by Knut and Alice Wallenberg Foundation [2013.0253] and Swedish Research Council FORMAS [2016-02031]. Jesus is grateful to Consejo Nacional de Ciencia y Tecnología (CONACyT, Mexico) scholarship number 622969, and DAAD programme number 57378443 for funding. Authors also acknowledge the support of the School of Chemistry of the Universidad Nacional Autónoma de México (UNAM), grant PAIP 5000-9163 and the Programa de Apoyo a Proyectos para la Innovación y Mejoramiento de la Enseñanza (PAPIME) grant PE200118, UNAM.

## References

- 1 E. Diamanti-Kandarakis, J.-P. Bourguignon, L. C. Giudice, R. Hauser, G. S. Prins, A. M. Soto, R. T. Zoeller and A. C. Gore, *Endocr. Rev.*, 2009, **30**, 293–342.
- 2 WHO/UNEP report, *State of the Science of Endocrine Disrupting Chemicals – 2012*, <http://www.who.int/ceh/publications/endocrine/en/>.

- 3 D. Ding, L. Xu, H. Fang, H. Hong, R. Perkins, S. Harris, E. D. Bearden, L. Shi and W. Tong, *BMC Bioinf.*, 2010, **11**, S5.
- 4 G. Klopman and S. K. Chakravarti, *Chemosphere*, 2003, **51**, 445–459.
- 5 H. Hong, W. Tong, Q. Xie, H. Fang and R. Perkins, *SAR QSAR Environ. Res.*, 2005, **16**, 339–347.
- 6 W. Tong, Q. Xie, H. Hong, L. Shi, H. Fang and R. Perkins, *Environ. Health Perspect.*, 2004, **112**, 1249–1254.
- 7 S.-P. Korhonen, K. Tuppurainen, R. Laatikainen and M. Peräkylä, *J. Chem. Inf. Model.*, 2005, **45**, 1874–1883.
- 8 T. Ghafourian and M. T. D. Cronin, *QSAR Comb. Sci.*, 2006, **25**, 824–835.
- 9 H. Liu, E. Papa and P. Gramatica, *Chem. Res. Toxicol.*, 2006, **19**, 1540–1548.
- 10 L. Ji, X. Wang, S. Luo, L. Qin, X. Yang, S. Liu and L. Wang, *Sci. China, Ser. B: Chem.*, 2008, **51**, 677.
- 11 L. Ji, X. Wang, X. Yang, S. Liu and L. Wang, *Chin. Sci. Bull.*, 2008, **53**, 33–39.
- 12 N. Stojić, S. Erić and I. Kuzmanovski, *J. Mol. Graphics Modell.*, 2010, **29**, 450–460.
- 13 L. Zhang, A. Sedykh, A. Tripathi, H. Zhu, A. Afantitis, V. D. Mouchlis, G. Melagraki, I. Rusyn and A. Tropsha, *Toxicol. Appl. Pharmacol.*, 2013, **272**, 67–76.
- 14 H. Fang, W. Tong, L. M. Shi, R. Blair, R. Perkins, W. Branham, B. S. Hass, Q. Xie, S. L. Dial, C. L. Moland and D. M. Sheehan, *Chem. Res. Toxicol.*, 2001, **14**, 280–294.
- 15 J. Bajorath, L. Peltason, M. Wawer, R. Guha, M. S. Lajiness and J. H. Van Drie, *Drug Discovery Today*, 2009, **14**, 698–705.
- 16 L. Peltason and J. Bajorath, *Chem. Biol.*, 2007, **14**, 489–497.
- 17 M. Reutlinger, W. Guba, R. E. Martin, A. I. Alanine, T. Hoffmann, A. Klenner, J. A. Hiss, P. Schneider and G. Schneider, *Angew. Chem., Int. Ed.*, 2011, **50**, 11633–11636.
- 18 G. M. Maggiora, *J. Chem. Inf. Model.*, 2006, **46**, 1535.
- 19 M. Cruz-Monteagudo, J. L. Medina-Franco, Y. Pérez-Castillo, O. Nicolotti, M. N. D. S. Cordeiro and F. Borges, *Drug Discovery Today*, 2014, **19**, 1069–1080.
- 20 D. Stumpfe, A. de la Vega de León, D. Dimova and J. Bajorath, *F1000Research*, 2014, **3**, 75.
- 21 F. I. Saldívar-González, J. J. Naveja, O. Palomino-Hernández and J. L. Medina-Franco, *RSC Adv.*, 2017, **7**, 632–641.
- 22 J. J. Naveja and J. L. Medina-Franco, *RSC Adv.*, 2015, **5**, 63882–63895.
- 23 O. Mendez-Lucio, J. Perez-Villanueva, R. Castillo and J. L. Medina-Franco, *Mol. Inf.*, 2012, **31**, 837–846.
- 24 D. Rogers and M. Hahn, *J. Chem. Inf. Model.*, 2010, **50**, 742–754.
- 25 G. W. Bemis and M. A. Murcko, *J. Med. Chem.*, 1996, **39**, 2887–2893.
- 26 M. R. Berthold, N. Cebron, F. Dill, T. R. Gabriel, T. Kötter, T. Meinl, P. Ohl, C. Sieb, K. Thiel and B. Wiswedel, in *Data Analysis, Machine Learning and Applications: Proceedings of the 31st Annual Conference of the Gesellschaft für Klassifikation e.V., Albert-Ludwigs-Universität Freiburg, March 7–9, 2007*, ed. C. Preisach, H. Burkhardt, L. Schmidt-Thieme and R. Decker, Springer, Berlin, Heidelberg, 2008, pp. 319–326, DOI: 10.1007/978-3-540-78246-9\_38.





- 27 S. Beisken, T. Meinel, B. Wiswedel, L. F. de Figueiredo, M. Berthold and C. Steinbeck, *BMC Bioinf.*, 2013, **14**, 257.
- 28 M. González-Medina, F. D. Prieto-Martínez and J. L. Medina-Franco, *J. Cheminf.*, 2016, **8**, 63.
- 29 J. Naveja, M. Rico-Hidalgo and J. Medina-Franco, *F1000Research*, 2018, **7**, 993.
- 30 V. Shanmugasundaram and G. M. Maggiora, *Presented in part at the 222nd ACS National Meeting*, Chicago, IL, United States, August 26–30, 2001.
- 31 J. L. Medina-Franco, K. Martínez-Mayorga, A. Bender and T. Scior, *QSAR Comb. Sci.*, 2009, **28**, 1551–1560.
- 32 J. L. Medina-Franco, *J. Chem. Inf. Model.*, 2012, **52**, 2485–2493.





## RESEARCH ARTICLE

**REVISED** Analysis of a large food chemical database: chemical space, diversity, and complexity [version 2; referees: 3 approved]J. Jesús Naveja <sup>1,2</sup>, Mariel P. Rico-Hidalgo <sup>2</sup>, José L. Medina-Franco <sup>2</sup><sup>1</sup>PECEM, Faculty of Medicine, Universidad Nacional Autónoma de México, Mexico City, 04510, Mexico<sup>2</sup>Department of Pharmacy, School of Chemistry, Universidad Nacional Autónoma de México, Mexico City, 04510, Mexico**v2** First published: 03 Jul 2018, 7(CHEM INF SCI):993 (doi: 10.12688/f1000research.15440.1)

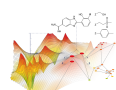
Latest published: 10 Aug 2018, 7(CHEM INF SCI):993 (doi: 10.12688/f1000research.15440.2)

**Abstract****Background:** Food chemicals are a cornerstone in the food industry. However, its chemical diversity has been explored on a limited basis, for instance, previous analysis of food-related databases were done up to 2,200 molecules.

The goal of this work was to quantify the chemical diversity of chemical compounds stored in FooDB, a database with nearly 24,000 food chemicals.

**Methods:** The visual representation of the chemical space of FooDB was done with ChemMaps, a novel approach based on the concept of chemical satellites. The large food chemical database was profiled based on physicochemical properties, molecular complexity and scaffold content. The global diversity of FooDB was characterized using Consensus Diversity Plots.**Results:** It was found that compounds in FooDB are very diverse in terms of properties and structure, with a large structural complexity. It was also found that one third of the food chemicals are acyclic molecules and ring-containing molecules are mostly monocyclic, with several scaffolds common to natural products in other databases.**Conclusions:** To the best of our knowledge, this is the first analysis of the chemical diversity and complexity of FooDB. This study represents a step further to the emerging field of "Food Informatics". Future study should compare directly the chemical structures of the molecules in FooDB with other compound databases, for instance, drug-like databases and natural products collections. An additional future direction of this work is to use the list of 3,228 polyphenolic compounds identified in this work to enhance the on-going polyphenol-protein interactome studies.**Keywords**

ChemMaps, chemical space, chemoinformatics, consensus diversity plots, diversity, FooDB, Foodinformatics, in silico

This article is included in the **Chemical Information Science gateway**.**Open Peer Review**

Referee Status:

	Invited Referees		
	1	2	3
<b>REVISED</b>			
<b>version 2</b> published 10 Aug 2018			
	↑		
<b>version 1</b> published 03 Jul 2018			

- 1 **Piotr Minkiewicz** , University of Warmia and Mazury in Olsztyn, Poland
- 2 **Khushbu Shah** , Duquesne University, USA  
Kramer Levin Naftalis Frankel LLP, USA
- 3 **Rachelle J. Bienstock**, RJB  
Computational Modeling LLC, USA

**Discuss this article**

Comments (0)

**Corresponding author:** José L. Medina-Franco ([medinajl@unam.mx](mailto:medinajl@unam.mx))

**Author roles:** **Naveja JJ:** Conceptualization, Formal Analysis, Investigation, Methodology, Writing – Original Draft Preparation, Writing – Review & Editing; **Rico-Hidalgo MP:** Formal Analysis, Investigation, Writing – Original Draft Preparation, Writing – Review & Editing; **Medina-Franco JL:** Conceptualization, Methodology, Project Administration, Resources, Supervision, Writing – Review & Editing

**Competing interests:** No competing interests were disclosed.

**Grant information:** This work was supported by a Consejo Nacional de Tecnología (CONACyT) scholarship [622969] (JJN). Programa de Apoyo a Proyectos de Investigación e Innovación Tecnológica (PAPIIT) Grant [IA203018] from the Universidad Nacional Autónoma de México (JLMF). *The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.*

**Copyright:** © 2018 Naveja JJ *et al.* This is an open access article distributed under the terms of the [Creative Commons Attribution Licence](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. Data associated with the article are available under the terms of the [Creative Commons Zero "No rights reserved" data waiver](#) (CC0 1.0 Public domain dedication).

**How to cite this article:** Naveja JJ, Rico-Hidalgo MP and Medina-Franco JL. **Analysis of a large food chemical database: chemical space, diversity, and complexity [version 2; referees: 3 approved]** *F1000Research* 2018, 7(Chem Inf Sci):993 (doi: [10.12688/f1000research.15440.2](https://doi.org/10.12688/f1000research.15440.2))

**First published:** 03 Jul 2018, 7(Chem Inf Sci):993 (doi: [10.12688/f1000research.15440.1](https://doi.org/10.12688/f1000research.15440.1))

**REVISED** Amendments from Version 1

We thank the reviewers for the valuable comments and suggestions. We addressed all the comments of Piotr Minkiewicz emphasizing on the novelty, implications and future directions of this work. In the revised version of the manuscript the three suggested references were added and discussed accordingly. It is now mentioned that the findings of this work agree with the results of Lacroix S. *et al.* and the list of polyphenolic compounds made available in this work can further complement the works of Jensen K. *et al.* (2014 and 2015). In the revised manuscript we also acknowledged the optional suggestions of Khushbu Shah. The rationale behind the selection of the three version of the data sets was added. It was also acknowledged as a future work, the suggestion of conducting a systematic analysis of the functional groups in the acyclic compounds of FooDB.

See referee reports

## Introduction

Despite the high relevance of food chemicals in many areas including nutrition, disease prevention, and broad impact in the food industry, the chemical space and diversity of food chemical databases (Minkiewicz *et al.*, 2016) has been quantified on a limited basis. Previous efforts include the analysis and comparison of about 2,200 Generally Recognized as Safe (GRAS) flavoring substances (discrete chemical entities only) with compound databases relevant in drug discovery and natural product research e.g., drugs approved for clinical use, compounds in the ZINC database, and natural products from different sources (Burdock & Carabin, 2004; González-Medina *et al.*, 2016; González-Medina *et al.*, 2017; Martínez-Mayorga *et al.*, 2013; Medina-Franco *et al.*, 2012; Peña-Castillo *et al.*, 2018). Other food-related chemical databases, comprising around 900 compounds, were analyzed by Ruddigkeit and J.-L. Reymond (Ruddigkeit & Reymond, 2014). The limited quantitative analysis of food chemicals has been in part due to the scarce availability of food chemical databases in the public domain. A major exception, however, is FooDB a large database with more than 20,000 food chemicals (The Metabolomics Innovation Centre, 2017). To date, it is the most informative public repository of food compounds.

As part of a continued effort to characterize the chemical contents and diversity of food chemicals (González-Medina *et al.*, 2016; Martínez-Mayorga & Medina-Franco, 2009; Medina-Franco *et al.*, 2012), herein we report a quantitative analysis of the chemical space and chemical diversity of FooDB. Widely characterized compound databases such as GRAS, approved drugs and screening compounds used in drug discovery projects were employed as references. We used well-established and novel (but validated) chemoinformatic methods to analyze compound collections. Although most of these approaches are commonly used in drug discovery, this and previous works show they can be readily applied for food chemicals (Peña-Castillo *et al.*, 2018). Thereby this study represents a contribution to further advance the emerging field of Foodinformatics (Martínez-Mayorga & Medina-Franco, 2014).

## Methods

### Databases and data curation

Four chemical databases were homogeneously curated and analyzed, namely: FooDB version 1.0 (accessed November, 2017) (The Metabolomics Innovation Centre, 2017), drugs approved for clinical use available in DrugBank 5.0.2. (Law *et al.*, 2014), GRAS (Burdock & Carabin, 2004), and a random subset of drug-like natural products from ZINC 12 (Irwin & Shoichet, 2005), of a size comparable to FooDB. The GRAS and DrugBank sets used in this work also have been used as reference in other comparative studies (Medina-Franco *et al.*, 2012). The random set from ZINC was employed just as reference and other random sets from ZINC could be used. Compounds from all databases were washed and prepared using Wash MOE 2017 node in KNIME version 3.5.3 (Berthold *et al.*, 2008). Briefly, the washing protocol implemented in MOE included removing salts and neutralizing the charges in the molecules. The largest fragments were kept and duplicates in each dataset deleted. Table 1 summarizes the databases and sizes after data preprocessing.

### Chemical space visualization

The visual representation was generated with ChemMaps, a novel method for large chemical space visualizations (Naveja & Medina-Franco, 2017). Briefly, ChemMaps is able to generate two- and three-dimensional representations of the chemical space based. It uses as input the pairwise chemical similarity computed using fingerprints data. This approach exploits the 'chemical satellites' concept (Oprea & Gottfries, 2001), i.e., molecules whose similarity to the rest of the molecules in the database yield sufficient information for generating a visualization of the chemical space. Further details of ChemMaps are described elsewhere (Naveja & Medina-Franco, 2017).

### Physicochemical properties

Six physicochemical properties (PCP) were calculated with RDKit KNIME nodes version 3.4, namely: SlogP (partition coefficient), TPSA (topological polar surface area), AMW (atomic mass weight), RB (rotatable bonds), HBD (hydrogen bond donors) and HBA (hydrogen bond acceptors). For the analysis reported in this short communication, these properties were selected based on their broadly extended use for cross-comparison

**Table 1. Compound databases analyzed in this work.**

Database	Size <sup>a</sup>
FooDB	23,883
GRAS	2,244
DrugBank	8,748
Natural products in ZINC (drug-like random subset)	24,000

<sup>a</sup>Number of compounds after data curation

GRAS: Generally Recognized as Safe

of compound databases of biological relevance. However, additional properties can be calculated.

### Molecular complexity

Fraction of  $sp^3$  carbons and number of stereocenters were computed for FooDB as measures of structural complexity. Despite the fact that there are several other measures, these two are straightforward to interpret, easy to calculate and are becoming standard to make cross comparisons among databases (Méndez-Lucio & Medina-Franco, 2017). As described in the Results and Discussion section, the computed values for FooDB were compared to literature data already reported for the reference data sets.

### Scaffold content

The term “molecular scaffold” is employed to describe the core structure of a molecule (Brown & Jacoby, 2006). Different approaches have been proposed to consistently obtain a molecule’s scaffold *in silico*. In this work, scaffolds were generated under the Bemis-Murcko definition using the RDKit nodes available in KNIME (Bemis & Murcko, 1996). Bemis and Murcko define a scaffold as “the union of ring systems and linkers in a molecule”, i.e., all side chains of a molecule are removed.

### Global diversity

The so-called “global diversity” (or total diversity) of FooDB was assessed and compared to other reference collections using a consensus diversity plot (González-Medina *et al.*, 2016). As described recently, a consensus diversity plot simultaneously represents, in two-dimensions, four diversity criteria: structural (based on pairwise molecular fingerprint similarity values), scaffolds (using Murcko scaffolds computed as described in the Scaffold content section), physicochemical properties (based on the six properties described in Physicochemical properties section), and database size (the number of compounds) (González-Medina *et al.*, 2016). The structural diversity of each data set is represented on the X-axis and was defined as the median Tanimoto coefficient of MACCS keys fingerprints. The scaffold diversity of each database is represented on the Y-axis and was defined as the area under the corresponding scaffold recovery curve, a well-established metric to measure scaffold diversity (Medina-Franco *et al.*, 2009). The diversity based on PCP was defined as the Euclidean distance of six auto-scaled properties (SlogP, TPSA, AMW, RB, HBD, and HBA - *vide supra*) and is shown as the filling of the data points using a continuous color scale. The relative number of compounds in the data set is represented with a different size of the data points (smaller data sets are represented with smaller data points).

## Results and discussion

### Visual representation of the chemical space

Chemical space of FooDB in comparison with the compounds of the three reference databases is visualized in Figure 1. The figure also shows the individual comparisons of FooDB with GRAS, DrugBank and natural products subset from ZINC, respectively. As shown in Figure 1a, the coverage of chemical space of FooDB is quite large as compared to other datasets.

Most GRAS compounds lie within the chemical space framed by FooDB (Figure 1b): indeed, 1,193 compounds (53% of GRAS) are structurally identical between the two databases. Hence, FooDB largely contains and upgrades structural information from GRAS. There is significant overlap with approved drugs (Figure 1c) and natural products from ZINC with FooDB (Figure 1d).

### Distribution of physicochemical properties

Figure 2 shows the boxplots for the distribution of PCP in all the four databases. For better visualization, the outliers above or below the median  $\pm 1.5$  interquartile range are omitted. As expected, due to the large structural diversity, distribution of PCP in FooDB is broad, in many cases overcoming even approved drugs. For most properties, except RB, several compounds in FooDB share the properties of drugs, and drug-like natural products in ZINC. The comparable physicochemical properties between compounds from FooDB and DrugBank encourages additional systematic investigations for bioactivity of food components. Of course, during this search one needs to consider that compounds with similar properties may have different activity profile. In turn, GRAS consists mostly of small-sized compounds. Table S1 (Supplementary File 1) summarizes the statistics for FooDB and other reference collections.

### Molecular complexity

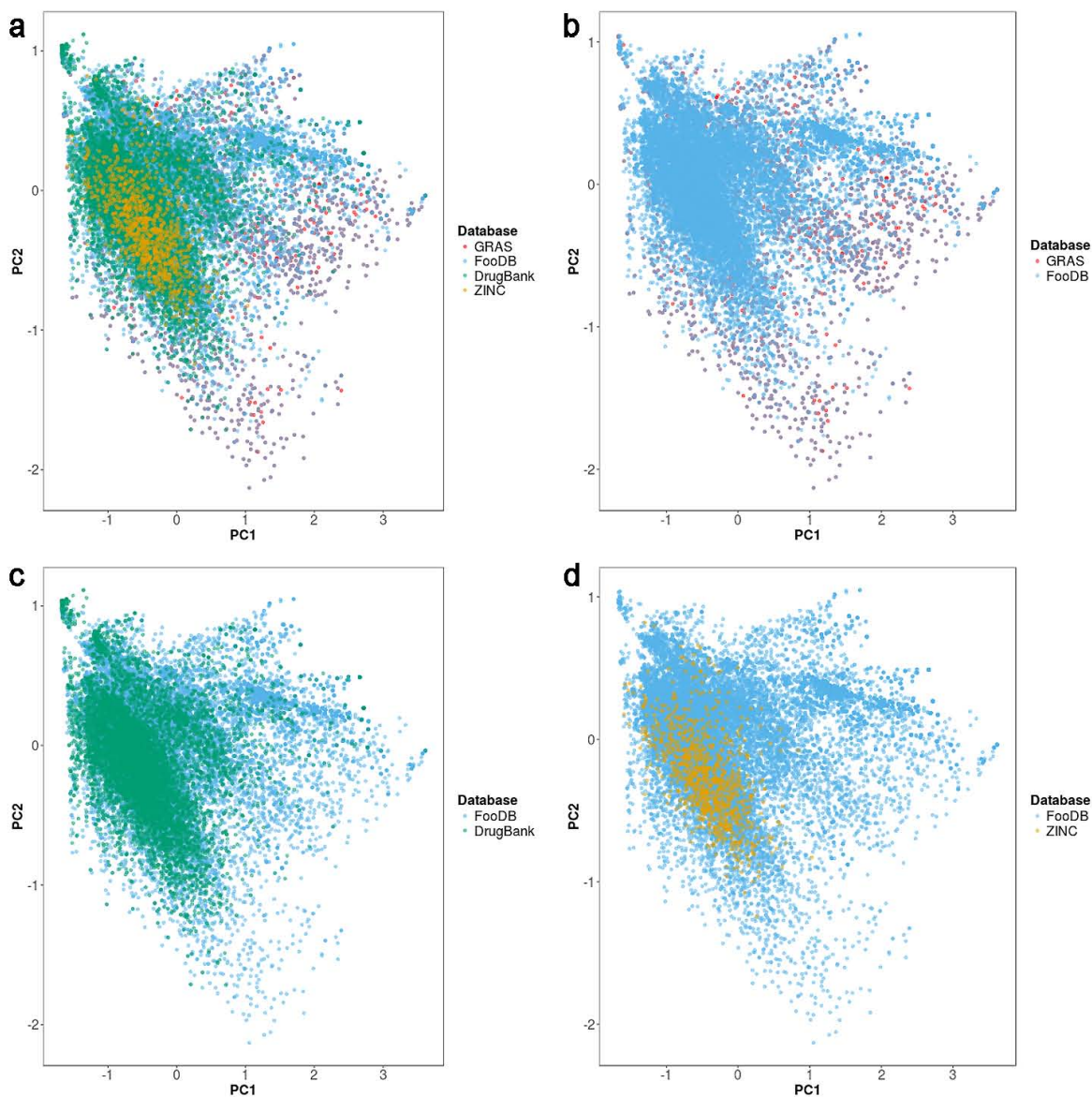
For FooDB, the fraction of  $sp^3$  carbons (mean: 0.62; standard deviation: 0.28) and the number of stereocenters (mean: 4.7; standard deviation: 7.1) indicated a high structural complexity. For comparison, it has reported that the mean of the fraction of  $sp^3$  carbons for approved drugs, compounds in the clinic and a general screening collections of organic compounds is 0.47, 0.41 and 0.32, respectively (González-Medina *et al.*, 2016; Lovering *et al.*, 2009). Moreover, the reported mean of the fraction of  $sp^3$  carbons for natural products collections ranges between 0.41 and 0.58 (for natural products in ZINC and Traditional Chinese Medicine (López-Vallejo *et al.*, 2012). The complexity of compounds in FooDB is comparable to molecules in GRAS (mean: 0.63; standard deviation: 0.28) (González-Medina *et al.*, 2016).

### Scaffold content

Figure 3 shows the frequency of the most common scaffolds in FooDB. Many compounds are acyclic (32%), followed by monocyclic compounds with a benzene (6%), cyclohexene (2%) and tetrahydropyran (1%) as a core structure. The benzene ring is the most common core scaffold in chemical databases used in drug discovery (Bemis & Murcko, 1996; Singh *et al.*, 2009; Yongye *et al.*, 2012). Many of the most frequent scaffolds in FooDB are also common in other compound databases of natural products (González-Medina *et al.*, 2017). In a follow-up work, it will be interesting to explore the type of functional groups commonly present in the acyclic structures of FooDB.

Recently, Schneider *et al.* published an analysis on the selectivity of Bemis-Murcko scaffolds based on public bioactivity data available in ChEMBL (Schneider & Schneider, 2017). 78 of the



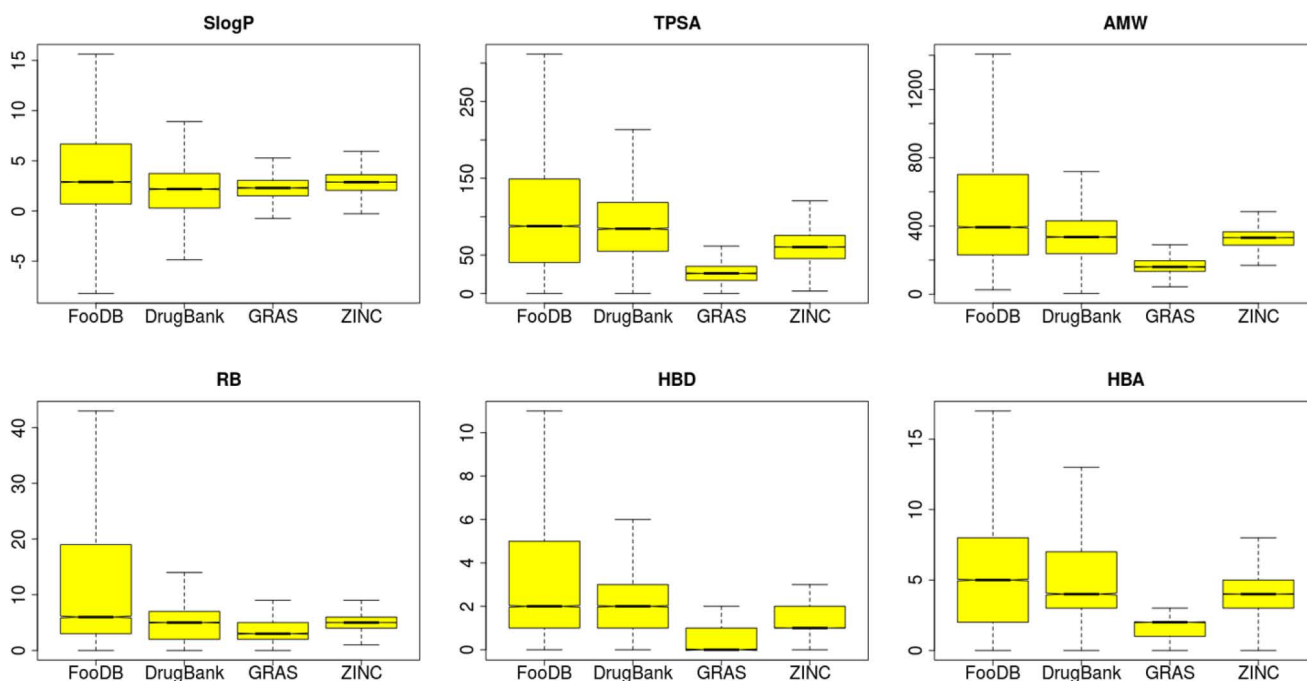


**Figure 1. Representation of the chemical space of FooDB.** The visual representation was generated with ChemMaps (Naveja & Medina-Franco, 2017). **a)** Comparison of FooDB with three reference collections. Panels **b-d)** show comparisons of FooDB with individual data sets.

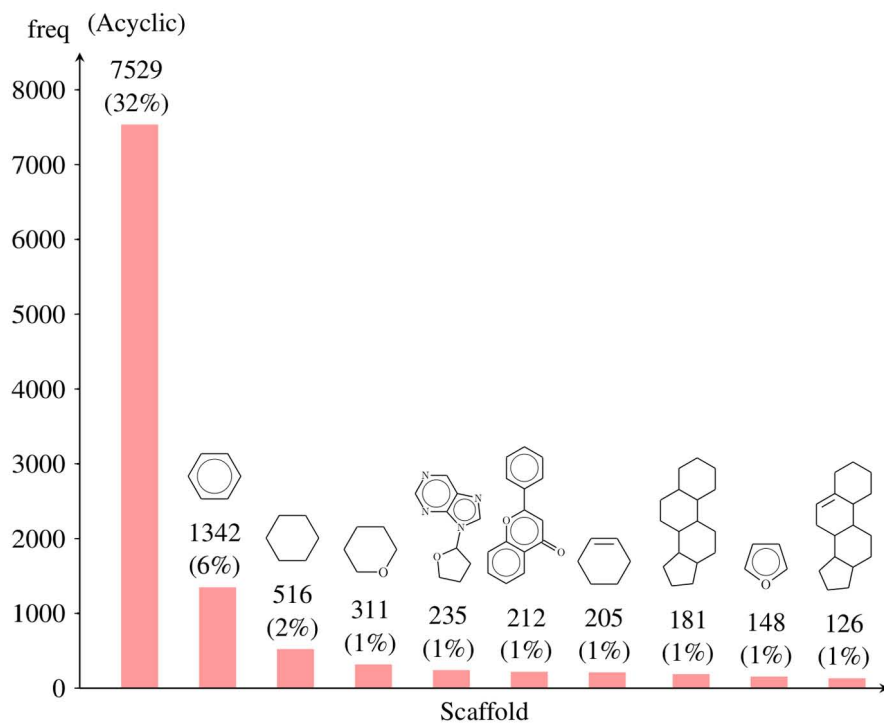
585 scaffolds reported therein were present in FooDB. The list of the 78 matching scaffolds, along with the original statistics calculated by Schneider *et al.*, is made available as [Dataset 1](#) (Naveja *et al.*, 2018a). Of note, the three most frequent scaffolds in FooDB (benzene, cyclohexane and tetrahydropyran, with more than 300 compounds - [Figure 3](#)) are matching scaffolds. Interestingly, the mean *Information content* (I) value of all 585 Schneider's scaffolds is 2.8 (sd= 0.6), while the subset of the 78 scaffolds also present in FooDB has a mean I value of only 2.1 (sd = 0.7). Lower I values point towards more promiscuous scaffolds (Schneider & Schneider, 2017), an expected

finding given the nature of the database. As example, [Table S2](#) ([Supplementary File 1](#)) shows and discusses briefly the statistics for the three most frequent matching scaffolds.

**Polyphenols.** Since polyphenols are an important class of compounds in food chemistry (Rasouli *et al.*, 2017), we investigated and quantified the amount of polyphenols in FooDB. Polyphenols are well-known antioxidants, which may play a role in the prevention of several diseases including type 2 diabetes, cardiovascular diseases, and some types of cancer (Neveu *et al.*, 2010). In this line, it is known that oxidative/nitrosative stress



**Figure 2. Distribution of physicochemical properties.** Box plots of the distribution of six physicochemical properties of FooDB and reference data sets. SlogP (partition coefficient), TPSA (topological polar surface area), AMW (atomic mass weight), RB (rotatable bonds), HBD (hydrogen bond donors) and HBA (hydrogen bond acceptors).



**Figure 3. Frequency of the ten most common scaffolds in FooDB.**

has a pivotal role in pathophysiology of neurodegenerative disorders and other kinds of disease (Ebrahimi & Schluesener, 2012). Polyphenols have been demonstrated to elicit several biological effects in *in vitro* and *ex vivo* tests (Del Rio *et al.*, 2010; Scalbert *et al.*, 2005).

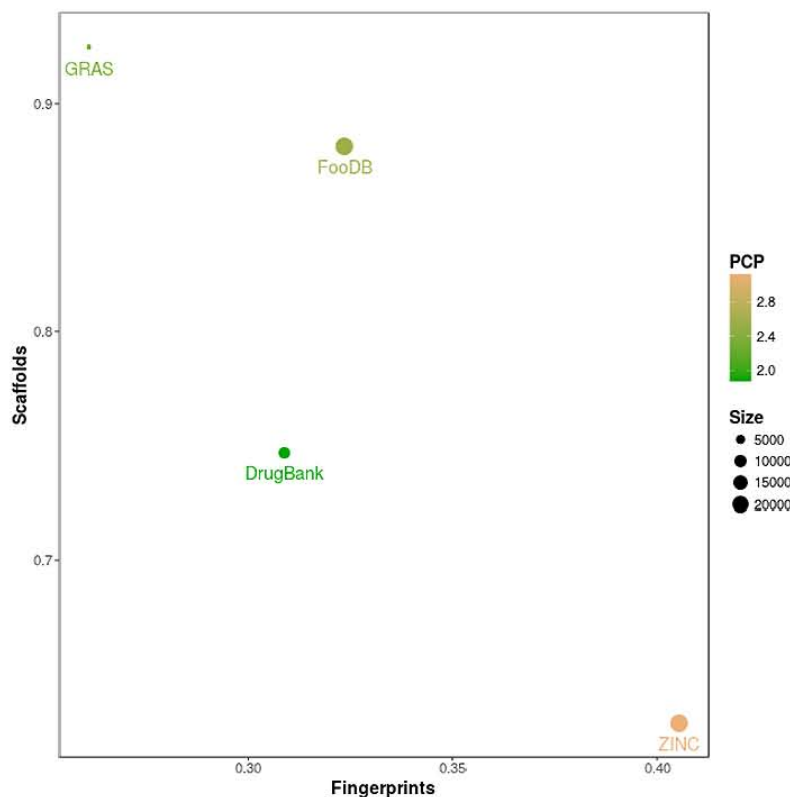
The molecular structure of polyphenols includes at least two phenolic groups, or one biphenol, and up to any additional number of OH substitutions in aryl rings. They may be classified by their structure in two major groups: flavonoids and non-flavonoids (phenolic acid derivatives) (Del Rio *et al.*, 2013). Some polyphenols, such as quercetin, are found in all plant products, whereas others are specific to particular foods. In many cases, food contain complex mixtures of polyphenols, which are often poorly characterized (Manach *et al.*, 2004).

Polyphenols are also a common chemical motif among natural products, and they are often associated to promiscuity (Tang, 2016). In this work it was found that 3,228 (13.5%) compounds in FooDB are polyphenolic. The list of all 3,228 polyphenolic compounds is made available as Dataset 2 (Naveja *et al.*, 2018b). This set of polyphenols is larger than the 502 polyphenols from food indexed in Phenol-Explorer (Neveu *et al.*, 2010).

For comparison, all the reference databases used in this work contained less polyphenols than FooDB. GRAS, ZINC and DrugBank contained 15 (0.6%), 24 (0.1%) and 325 (3.7%) polyphenols, respectively. The large list of polyphenols identified from FooDB is larger than the list of 1,395 polyphenols identified and used in the recent work of Lacroix *et al.* (Lacroix *et al.*, 2018) that was retrieved from Phenol-Explorer and the Dictionary of Natural Products. Indeed, the list of 3,228 polyphenolic compound made available in this work can be used to augment the already extensive polyphenol-protein interactome work of Lacroix *et al.* (Lacroix *et al.*, 2018).

### Global diversity

Since the diversity of compound data sets depend on the molecular representation (Sheridan & Kearsley, 2002), a global assessment of the diversity of FooDB was analyzed using different criteria: molecular fingerprints, scaffolds, physicochemical properties and number of compounds. The four criteria were analyzed in an integrated manner through a Consensus Diversity Plot generated as described in the Global diversity section of the Methods. The Consensus Diversity Plot in Figure 4 shows that FooDB has about average diversity both by fingerprints and relatively low diversity by scaffolds.



**Figure 4. Consensus Diversity Plot of FooDB and reference data sets.** The structural diversity of each data set is represented on the X-axis and was defined as the median Tanimoto coefficient of MACCS keys fingerprints. The scaffold diversity of each database is represented on the Y-axis and was defined as the area under the corresponding scaffold recovery curve. The diversity based on physicochemical properties (PCP) was defined as the Euclidean distance of six auto-scaled properties (SlogP, TPSA, AMW, RB, HBD, and HBA) and is shown as the filling of the data points using a continuous color scale. The relative number of compounds is represented with a different size of the data points (smaller data sets are represented with smaller data points).

Although PCP (represented with the color of the data points) are extremely diverse, structural motifs seem to reappear with slight variations. Figure 4 shows the overall large fingerprint and scaffold diversity of approved drugs (e.g., data points towards the lower left region of the plot). Similarly, the relative global diversity of GRAS i.e., high fingerprint diversity but low scaffold diversity (e.g., upper left region of the plot), is consistent with previous comparisons of these compounds with other reference data sets (González-Medina *et al.*, 2016; Medina-Franco *et al.*, 2012).

**Dataset 1. Schneidermatch.sdf. This file contains the list of the 78 matching scaffolds in SDF format, along with the original statistics calculated by Schneider *et al.***

<http://dx.doi.org/10.5256/f1000research.15440.d209071>

No special software is required to open the SDF files. Any commercial or free software capable of reading SDF files will open the data sets supplied

**Dataset 2. FooDBpolyphenols.sdf. This file contains 3,228 polyphenolic compounds available in FooDB, in SDF format**

<http://dx.doi.org/10.5256/f1000research.15440.d209072>

No special software is required to open the SDF files. Any commercial or free software capable of reading SDF files will open the data sets supplied

## Conclusions

FooDB is a novel, large and diverse library containing information of more than 23,000 compounds found in food. To date, it is the most informative public resource of food compounds. Visual representation of the chemical space revealed that FooDB largely contains and upgrades structural information from GRAS. Indeed, most of GRAS is contained in FooDB. Compounds in FooDB have a large diversity of physicochemical properties. The distributions of most physicochemical properties of FooDB compounds overlap with those of approved drugs and natural products in ZINC. GRAS mostly contains small-sized compounds. The global diversity indicates that FooDB has a large structural diversity as measured by molecular fingerprints, though it has relatively low scaffold diversity. One third of the compounds in FooDB are acyclic. The most frequent cyclic scaffolds are monocyclic. Of note, polyphenols represent a large fraction of FooDB. The list of 3,228 polyphenolic compounds identified in this work to enhance the on-going polyphenol-protein interactome studies. Analysis of the chemical complexity revealed that compounds in FooDB are more complex than approved drugs and natural products and have complexity comparable to GRAS compounds. A next

step of this work is to compare the chemical space of FooDB with that of natural products from different sources, e.g., plants, terrestrial, cyanobacteria. A second suggested future study is to perform the virtual screening of FooDB across a range of targets, for instance, the increasingly important epigenetic targets (Naveja & Medina-Franco, 2018). Virtual screening can be done using multiple methods, for instance, using similarity searching. In this case one needs to consider, however, the potential presence of activity cliffs i.e., compounds with similar structure but different activity (Stumpfe *et al.*, 2014). The goal of such study would be to identify systematically dietary components that may be participating in epigenetic regulatory processes (Martinez-Mayorga *et al.*, 2013). These efforts are ongoing in our group and will be reported in due course. Other perspective of this work is integrating the knowledge of FooDB with other large databases with the aim of identifying food-disease associations and food-drug interactions such as the works previously published by Jensen *et al.* (Jensen *et al.*, 2014; Jensen *et al.*, 2015).

## Data availability

**Dataset 1:** (Schneidermatch.sdf). **This file contains the list of the 78 matching scaffolds in SDF format**, along with the original statistics calculated by Schneider *et al.* No special software is required to open the SDF files. Any commercial or free software capable of reading SDF files will open the data sets supplied. [10.5256/f1000research.15440.d209071](http://dx.doi.org/10.5256/f1000research.15440.d209071) (Naveja, *et al.*, 2018a)

**Dataset 2:** (FooDBpolyphenols.sdf). **This file contains 3,228 polyphenolic compounds available in FooDB, in SDF format**. No special software is required to open the SDF files. Any commercial or free software capable of reading SDF files will open the data sets supplied. [10.5256/f1000research.15440.d209072](http://dx.doi.org/10.5256/f1000research.15440.d209072) (Naveja *et al.*, 2018b)

## Competing interests

No competing interests were disclosed.

## Grant information

This work was supported by a Consejo Nacional de Tecnología (CONACyT) scholarship [622969] (JJN). Programa de Apoyo a Proyectos de Investigación e Innovación Tecnológica (PAPIIT) Grant [IA203018] from the Universidad Nacional Autónoma de México (JLMF).

*The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.*

## Acknowledgements

The authors thank Karina Martínez-Mayorga, Andrea Peña-Castillo and Nicole Trujillo for rich discussions and valuable insights.

## Supplementary material

**Supplementary File 1: File with supporting tables.** Table S1: Summary statistics of the distribution of six PCP of FooDB and other reference collections. Table S2: Selected scaffold statistics as reported by (Schneider & Schneider, 2017).

[Click here to access the data.](#)



## References

- Bemis GW, Murcko MA: **The properties of known drugs. 1. Molecular frameworks.** *J Med Chem.* 1996; **39**(15): 2887–93.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- Berthold MR, Cebon N, Dill F, *et al.*: **KNIME: The Konstanz Information Miner.** In: Preisach C, Burkhardt H, Schmidt-Thieme L, Decker R, (Eds.), *Data Analysis, Machine Learning and Applications.* Berlin, Heidelberg: Springer Berlin Heidelberg. 2008; 319–326.  
[Publisher Full Text](#)
- Brown N, Jacoby E: **On scaffolds and hopping in medicinal chemistry.** *Mini Rev Med Chem.* 2006; **6**(11): 1217–29.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- Burdock GA, Carabin IG: **Generally Recognized as Safe (GRAS): history and description.** *Toxicol Lett.* 2004; **150**(1): 3–18.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- Del Rio D, Costa LG, Lean ME, *et al.*: **Polyphenols and health: what compounds are involved?** *Nutr Metab Cardiovasc Dis.* 2010; **20**(1): 1–6.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- Del Rio D, Rodríguez-Mateos A, Spencer JP, *et al.*: **Dietary (poly)phenolics in human health: structures, bioavailability, and evidence of protective effects against chronic diseases.** *Antioxid Redox Signal.* 2013; **18**(14): 1818–92.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Ebrahimi A, Schluesener H: **Natural polyphenols against neurodegenerative disorders: potentials and pitfalls.** *Ageing Res Rev.* 2012; **11**(2): 329–45.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- González-Medina M, Owen JR, El-Elmat T, *et al.*: **Scaffold Diversity of Fungal Metabolites.** *Front Pharmacol.* 2017; **8**: 180.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- González-Medina M, Prieto-Martínez FD, Naveja JJ, *et al.*: **Chemoinformatic expedition of the chemical space of fungal products.** *Future Med Chem.* 2016; **8**(12): 1399–412.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- 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](#)
- Irwin JJ, Shoichet BK: **ZINC—a free database of commercially available compounds for virtual screening.** *J Chem Inf Model.* 2005; **45**(1): 177–82.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Jensen K, Panagiotou G, Kouskoumvekaki I: **Integrated text mining and chemoinformatics analysis associates diet to health benefit at molecular level.** *PLoS Comput Biol.* 2014; **10**(1): e1003432.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Jensen K, Ni Y, Panagiotou G, *et al.*: **Developing a molecular roadmap of drug-food interactions.** *PLoS Comput Biol.* 2015; **11**(2): e1004048.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Lacroix S, Klicic Badoux J, Scott-Boyer MP, *et al.*: **A computationally driven analysis of the polyphenol-protein interactome.** *Sci Rep.* 2018; **8**(1): 2232.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Law V, Knox C, Djoumbou Y, *et al.*: **DrugBank 4.0: shedding new light on drug metabolism.** *Nucleic Acids Res.* 2014; **42**(Database issue): D1091–7.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- López-Vallejo F, Giulianotti MA, Houghten RA, *et al.*: **Expanding the medically relevant chemical space with compound libraries.** *Drug Discov Today.* 2012; **17**(13–14): 718–26.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- Lovering F, Bikker J, Humblet C: **Escape from flatland: increasing saturation as an approach to improving clinical success.** *J Med Chem.* 2009; **52**(21): 6752–6.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- Manach C, Scalbert A, Morand C, *et al.*: **Polyphenols: food sources and bioavailability.** *Am J Clin Nutr.* 2004; **79**(5): 727–47.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- Martínez-Mayorga K, Medina-Franco JL: **Chemoinformatics-applications in food chemistry.** *Adv Food Nutr Res.* 2009; **58**: 33–56.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- Martínez-Mayorga K, Medina-Franco JL: **Foodinformatics: Applications of chemical information to food chemistry.** Springer. 2014;  
[Publisher Full Text](#)
- Martínez-Mayorga K, Peppard TL, López-Vallejo F, *et al.*: **Systematic mining of Generally Recognized as Safe (GRAS) flavor chemicals for bioactive compounds.** *J Agric Food Chem.* 2013; **61**(31): 7507–14.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- Medina-Franco JL, Martínez-Mayorga K, Bender A, *et al.*: **Scaffold diversity analysis of compound data sets using an entropy-based measure.** *QSAR Comb Sci.* 2009; **28**(11–12): 1551–1560.  
[Publisher Full Text](#)
- Medina-Franco JL, Martínez-Mayorga K, Peppard TL, *et al.*: **Chemoinformatic analysis of GRAS (Generally Recognized as Safe) flavor chemicals and natural products.** *PLoS One.* 2012; **7**(11): e50798.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Méndez-Lucio O, Medina-Franco JL: **The many roles of molecular complexity in drug discovery.** *Drug Discov Today.* 2017; **22**(1): 120–126.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- Minkiewicz P, Darewicz M, Iwaniak A, *et al.*: **Internet databases of the properties, enzymatic reactions, and metabolism of small molecules-search options and applications in food science.** *Int J Mol Sci.* 2016; **17**(12): pii: E2039.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Naveja JJ, Medina-Franco JL: **ChemMaps: Towards an approach for visualizing the chemical space based on adaptive satellite compounds [version 2; referees: 3 approved with reservations].** *F1000Res.* 2017; **6**: pii: Chem Inf Sci-1134.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Naveja JJ, Medina-Franco JL: **Insights from pharmacological similarity of epigenetic targets in epipolypharmacology.** *Drug Discov Today.* 2018; **23**(1): 141–150.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- Naveja JJ, Rico-Hidalgo MP, Medina-Franco JL: **Dataset 1 in: Analysis of a large food chemical database: chemical space, diversity, and complexity.** *F1000Research.* 2018a.  
<http://www.doi.org/10.5256/f1000research.15440.d209071>
- Naveja JJ, Rico-Hidalgo MP, Medina-Franco JL: **Dataset 2 in: Analysis of a large food chemical database: chemical space, diversity, and complexity.** *F1000Research.* 2018b.  
<http://www.doi.org/10.5256/f1000research.15440.d209072>
- Neveu V, Perez-Jiménez J, Vos F, *et al.*: **Phenol-Explorer: an online comprehensive database on polyphenol contents in foods.** *Database (Oxford).* 2010; **2010**: bap024.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Oprea TI, Gottfries J: **Chemography: the art of navigating in chemical space.** *J Comb Chem.* 2001; **3**(2): 157–166.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- Peña-Castillo A, Méndez-Lucio O, Owen JR, *et al.*: **Chemoinformatics in Food Science.** In J. Gasteiger & T. Engel (Eds.), *Chemoinformatics - Volume 2: From Methods to Applications.* Weinheim, Germany: Wiley-VCH. 2018.  
[Publisher Full Text](#)
- Rasouli H, Farzei MH, Khodarahmi R: **Polyphenols and their benefits: A review.** *Int J Food Prop.* 2017; **20**(sup2): 1700–1741.  
[Publisher Full Text](#)
- Ruddigkeit L, Reymond JL: **The chemical space of flavours.** In K. Martínez-Mayorga & J. L. Medina-Franco (Eds.), *Foodinformatics.* Cham: Springer International Publishing. 2014; 83–96.  
[Publisher Full Text](#)
- Scalbert A, Johnson IT, Saltmarsh M: **Polyphenols: antioxidants and beyond.** *Am J Clin Nutr.* 2005; **81**(1 Suppl): 215S–217S.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- Schneider P, Schneider G: **Privileged Structures Revisited.** *Angew Chem Int Ed Engl.* 2017; **56**(27): 7971–7974.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Sheridan RP, Kearsley SK: **Why do we need so many chemical similarity search methods?** *Drug Discov Today.* 2002; **7**(17): 903–911.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- Singh N, Guha R, Giulianotti MA, *et al.*: **Chemoinformatic analysis of combinatorial libraries, drugs, natural products, and molecular libraries small molecule repository.** *J Chem Inf Model.* 2009; **49**(4): 1010–1024.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Stumpfe D, de la Vega De León A, Dimova D, *et al.*: **Advancing the activity cliff concept, part II.** *F1000Res.* 2014; **3**: 75.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Tang GY: **Why Polyphenols have Promiscuous Actions? An Investigation by Chemical Bioinformatics.** *Nat Prod Commun.* 2016; **11**(5): 655–656.  
[PubMed Abstract](#)
- The Metabolomics Innovation Centre: **FoodDB (Version 1).** Computer software, Canada: The Metabolomics Innovation Centre. 2017.  
[Reference Source](#)
- Yongye AB, Waddell J, Medina-Franco JL: **Molecular scaffold analysis of natural products databases in the public domain.** *Chem Biol Drug Des.* 2012; **80**(5): 717–724.  
[PubMed Abstract](#) | [Publisher Full Text](#)





# 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 cheminformatics 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** Cheminformatics, ChemMaps, Chemical space, Data mining, Epigenetics, Epigenomics, 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](https://doi.org/10.1007/7653_2018_6)) contains supplementary material, which is available to authorized users.

- c. Analyzing and comparing database diversity, and
- d. Visualizing their chemical space.

Of note, KNIME is an open-access initiative intended for generating data mining pipelines or workflows, which are capable of integrating multiple tools [5].

Although sufficiently detailed, this review aims at being a quick practical guide. More comprehensive tutorials in chemoinformatics can be found elsewhere [6, 7]. Additionally, web applications for cheminformatics methods that have been developed by our research group are mentioned in the respective subsections. These applications are part of the D-Tools initiative for generating open cheminformatics resources (available at <https://www.difacquim.com/d-tools/>). The D-Tools usage is further described elsewhere [4, 8–11], and these are not the focus of this review.

---

## 2 Methods

### 2.1 Construction and Curation of a Compound Database

Due to the increase in the amount of chemical information, where it is common to the concept of big data [12], the efficient management of information represents a challenge today. This is of particular importance in polypharmacology where large compound datasets contain information of screening across several biological endpoints. In response to this need, the construction of compound and other databases can be a convenient way to sort information according to the data available and the specific objectives of the study.

In chemoinformatics, construction of databases is a fundamental practice to perform various computational studies like the design of chemical libraries, characterization and comparison of the chemical space, the study of the structure–activity relationships (SAR), and virtual screening studies, among others.

Currently, web pages of large public databases such as DrugBank [13], ChEMBL [14], ZINC [15], and BindingDB [16] allow the user to download their own databases (complete or partial downloads) with information on approved drugs, drugs in the experimental phase, commercially available compounds, molecular targets, etc. However, these databases are not always updated, so they can be enriched with new information published in books or in scientific articles.

Also, in research groups devoted to the synthesis, isolation from natural sources and/or evaluation of new chemical entities can be carried out for the construction of completely new compounds' databases. Such collections are usually referred to as in-house databases.

The process of building and annotating chemical databases is not trivial. Each organization has its own rules, conventions, and

procedures. However, the steps that are considered essential are listed below:

1. Identify compounds and resources that contain information required, e.g., journals and databases with chemical information [4, 17].
2. In a spreadsheet, it is recommended that the user has the following information for each compound:
  - a. Name of each compound. This can be searched in public databases.
  - b. A number that identifies this compound in the database that has been consulted, for example, ChemSpider ID, Substance or Compound ID (SID, CID in PubChem, the CAS registry number, or an internal and consistent code if building an in-house collection).
  - c. Structure input. An example of this is the use of Canonical SMILES notation used for encoding molecular structures that can be imported to other molecular editing systems. It is worth noting the relevance of creating a single computational representation. This can be achieved by using various algorithms in a process known as canonicalization.
3. Once this information is collected in the spreadsheet, save the database preferably in *.csv* format (comma delimited). Other database formats with chemical information and compatible with most computer programs as KNIME are *sdf* (structure data file), *mol* (molecular data file), and *mol2* (tripos mol2 file).

For the management and analysis of databases, the KNIME Example Server provides access to many explanatory workflows. The example server is accessible via the KNIME Explorer panel within the KNIME workbench and represents a great help when starting a new workflow.

Some of the nodes to start working with files with chemical information are: *Molecule Type Cast*, a node useful for reading chemical data from a *.csv* file or database, and this node casts any string as a chemical type (i.e., It tells KNIME “This is a smiles string”) and *Marvin MolConverter*, a node provided by Chemaxon/Infocom that translates seamlessly between types (smiles ↔ sdf ↔ mrv).

An important aspect to consider when analyzing molecular databases generated by other scientists is that these may contain wrong information or unnecessary information for the intended application or project. Therefore, cleaning or curating the information is highly relevant to enhance the quality of the data and to avoid erroneous results [18].

As in the construction of databases, there is no widely accepted standard protocol for the preparation of small molecules. However,

hereunder are described the essential points in the preparation and curation of databases:

1. Normalize the chemical structures. In this step, each chemical structure is checked for valid atom types, valence checks, and functional groups such as nitro groups are converted to a consistent representation. This is followed by a standardization step in which chemical structures are converted to a canonical tautomeric form, aromatic structures are kekulized, placement of stereo bonds is standardized, and all implicit hydrogens are converted to explicit hydrogens [19].
2. Remove duplicates. After the molecules have been properly standardized, it is appropriate to detect duplicates. InChiKeys is a useful method to identify several states of protonation and tautomers of a molecule.
3. Discard inorganic and organometallic atoms or molecules if these are not the object of study. It is worth mentioning that the majority of the chemoinformatics programs currently available are developed to process small organic molecules.
4. Wash the compound database by applying to each molecule a set of rules of “cleaning” such as the elimination of salts and the adjustment of the protonation states. The purpose of this step is to ensure that each chemical structure is in a form suitable for the subsequent modeling.
5. Enumerate tautomers and stereoisomers. This step is important in virtual screening studies, particularly when using search methods such as docking or pharmacophore.
6. Optimize the geometry and minimize the energy if the database will be used to evaluate the potential of each compound to bind to a receptor or enzyme, or to calculate descriptors which depend on the three-dimensional conformation of the molecule. The specific method to optimize the geometry will largely depend on the type, quantity of molecules to optimize, and, most importantly, on the specific application.

In addition, if the quantity of compounds is too large to be examined or tested with the resources available, different strategies can be employed to reduce the number of compounds in a rational and consistent manner. Such strategies include: filtering—essentially imposing secondary search criteria to eliminate compounds, clustering—taking a representative subset of a larger set, and human inspection of the compound structures (with or without extra data) [20].

In several articles, the impact of the use of duplicates and inconsistencies in the molecular structures in prediction models had already been discussed [21]. For this reason, the project CERAPP (Collaborative Estrogen Receptor Activity Prediction Project) has

developed a workflow to curate databases [22]. A similar workflow can be found at the link <https://github.com/zhu-lab/curation-workflow/blob/master/Structure%20Standardizer2.zip>.

Gally et al. also report a workflow designed to prepare molecular databases but focused on studies of virtual screening [23]. In addition to carrying out of the standardization of chemical structures, the workflow of Gally et al. has implemented filters (based on molecular property distribution) to characterize specific subsets of chemical libraries such as drug-like, lead-like, or fragment-like subsets of compounds.

See Workflow 1 in the Supplementary Information for an example in KNIME.

The following analyses use an epigenomics chemical database that has already been curated and published [24].

## **2.2 Diversity Analysis**

In drug discovery projects focused on one single target or multiple targets, it is of high relevance quantifying the structural diversity of compound datasets. For instance, if the goal of a high-throughput screening campaign is to identify hit compounds with a desirable polypharmacological profile, it is desirable to screen a compound collection with high diversity. This will increase the possibilities to find active molecules with a desirable profile. If the goal of the screening campaign is to further develop a focused library (e.g., increase the structure–activity information of a focused region in chemical space [25]), it is desirable to screen a compound dataset with high internal similarity (low diversity).

The diversity in a chemical library can be assessed in multiple ways, mainly depending on the data under scrutiny. In addition to the diversity metric, a key aspect of diversity analysis is molecular representation [26, 27]. The most common ways to represent molecules in chemoinformatic applications are molecular descriptors (including physicochemical properties and molecular fingerprints), and chemical scaffolds [28]. Depending on the type of descriptor and the level of accuracy desired (considering the time of computation and the number of compounds to analyze), the input structures can be in two or three dimensions (the latter requires conformational analysis). The choice of molecular representation depends on the goals of the study.

A more detailed description on how to use molecular descriptors and scaffolds as an input for diversity analysis follows in the next paragraphs. See Workflow 2 in the Supplementary Information for an exemplary diversity analysis in KNIME.

### **2.2.1 Molecular Descriptors**

Molecular descriptors capture information of the whole molecule and are usually straightforward to interpret. Also, whole molecular properties such as physicochemical properties of pharmaceutical interest are usually part of empirical rules for drug likeness that aids to guide drug discovery programs. KNIME includes RDKit,



CDK, and Indigo nodes, with which complexity descriptors (e.g., chiral carbons, and fraction of  $sp^3$  carbon atoms), and physicochemical properties of pharmaceutical interest (including molecular weight, number of hydrogen bond donors and acceptors, number of rotatable bonds, logarithm of octanol–water partition coefficient, and topological polar surface area) [28].

Starting with curated databases (discussed in Sect. 2.1), the steps for quantifying diversity with molecular descriptors are:

1. Select the features to be evaluated (usually the six commonest physicochemical properties of pharmaceutical relevance, *vide supra*).
2. Scale the data using a *Z*-transformation. This transforms the data to dimensional units. The purpose is to improve the comparability of the variables and give a similar weight to all of them independently of the units with which they were originally measured.
3. Compute pairwise euclidean distance. For a database with  $n$  compounds,  $n \times (n - 1)/2$  pairwise comparisons are to be computed. Euclidean distance is calculated with the formula:

$$D(A, B) = \sqrt{\sum_{i=1}^n (a_i - b_i)^2},$$

where  $D(A, B)$  is the euclidean distance between compound A and B,  $a_i$  and  $b_i$  are the  $i$ -th descriptor, and  $n$  the total number of descriptors [29].  $D(A, B)$  can take any positive real number as value.

4. Compute a central tendency statistic (e.g., mean or median) for all the pairwise comparisons. The larger the mean or median, the more diverse the dataset is [30].
5. Finally, for comparison, the statistic can be computed for other reference databases or looked up at the literature if already reported.

### 2.2.2 Molecular Fingerprints

Many structural features escape the very general information obtained with physicochemical and complexity descriptors. Molecular fingerprints are vectors that aim towards a more comprehensive set of features (usually more than a hundred) to compare molecules. Every feature is encoded as a Boolean variable, where “0” represents absence and “1” represents presence of the feature. Therefore, repeated motifs are not generally acknowledged. For every molecule, a Boolean vector of features is obtained, and these are susceptible of standard set operations [31–33]. However, molecular fingerprints do have limitations, for example, they could be more difficult to interpret intuitively, and therefore pose a greater difficulty for extracting insights relevant for medicinal chemistry.

The steps for computing diversity based on fingerprints are:

1. Select a molecular fingerprint. Although the selection of the “best” fingerprint could be different from case to case, it has been consistently found that MACCS keys 166-bits [34] are useful for quantifying database diversity. In turn, extended connectivity fingerprints of diameter 4 (ECFP4) [32] as well as other circular fingerprints are, overall, better suited for virtual screening, activity landscape modeling, and SAR studies in general.
2. Compute pairwise Tanimoto similarity [27, 35]. For a database with  $n$  compounds,  $n \times (n - 1)/2$  pairwise comparisons are to be computed. Tanimoto similarity is calculated with the expression:

$$T(A, B) = \frac{c}{a + b - c'}$$

where  $T(A, B)$  is Tanimoto similarity with possible values being any real number between 0 and 1,  $c$  is the number of features for which both molecules A and B have a “1” value,  $a$  is the number of features for which molecule A has a “1” value, and  $b$  is the number of features for which molecule B has a “1” value. Dissimilarity matrices implemented in KNIME are quite efficient at these calculations. However, by default they compute values as dissimilarities, the complement of similarities, or distance matrices. Conversion from Tanimoto dissimilarity to similarity is accomplished by just subtracting the value from 1 ( $T_s = 1 - T_d$ , where  $T_s$  is Tanimoto similarity and  $T_d$  is Tanimoto dissimilarity).

3. Compute a central tendency statistic (e.g., mean or median) for all the pairwise comparisons. Conversely to Euclidean distance (and any distance metric in general), the smaller the mean or median, the more diverse the dataset is [30].
4. Finally, for comparison, the statistic can be computed for other reference databases or looked up at the literature if already reported.

### 2.2.3 Molecular Scaffolds

KNIME has nodes for finding Murcko scaffolds [36, 37]. By definition, Murcko scaffolds contain all the cyclic systems in a molecule as well as the linkers between them. All other decorations and ramifications are omitted. The greatest benefit of working with scaffolds data is that, unlike molecular fingerprints, they are readily interpreted by medicinal chemists. Nonetheless, the representation is rougher and loses information from the side chains. Also, more advanced methods must be applied to account for the structural relations among the scaffolds.

It is logical and generally accepted that a dataset is more diverse when it has a large number of different scaffolds, and the proportions of compounds with each scaffold are evenly distributed. The procedure for measuring scaffold diversity is as follows:

1. Find Murcko scaffolds for every molecule in the dataset.
2. Compute a frequency table of the scaffolds.
3. From here, there are a number of different methods for assessing the diversity [38]:
  - a. Order the scaffolds by their frequency of occurrence and compute the median (i.e., the minimum number of scaffolds in the database that contain at least 50 % of the total entries). Lower values in this statistic mean higher diversity.
  - b. Order the scaffolds by their frequency of occurrence. This order would be an index from 1 to  $n$ , where  $n$  is the total number of different scaffolds in the dataset. Divide all indexes by  $n$ , such that the highest index value is 1. Using scaffold indexes in the  $x$ -axis and their respective cumulative proportions in the  $y$ -axis, compute the area under the curve as a diversity statistic. This statistic admits as value any real number in the domain [0.5, 1.0]. Lower values in this statistic mean higher diversity.
  - c. Compute scaled Shannon entropy (SSE) with the formula:

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

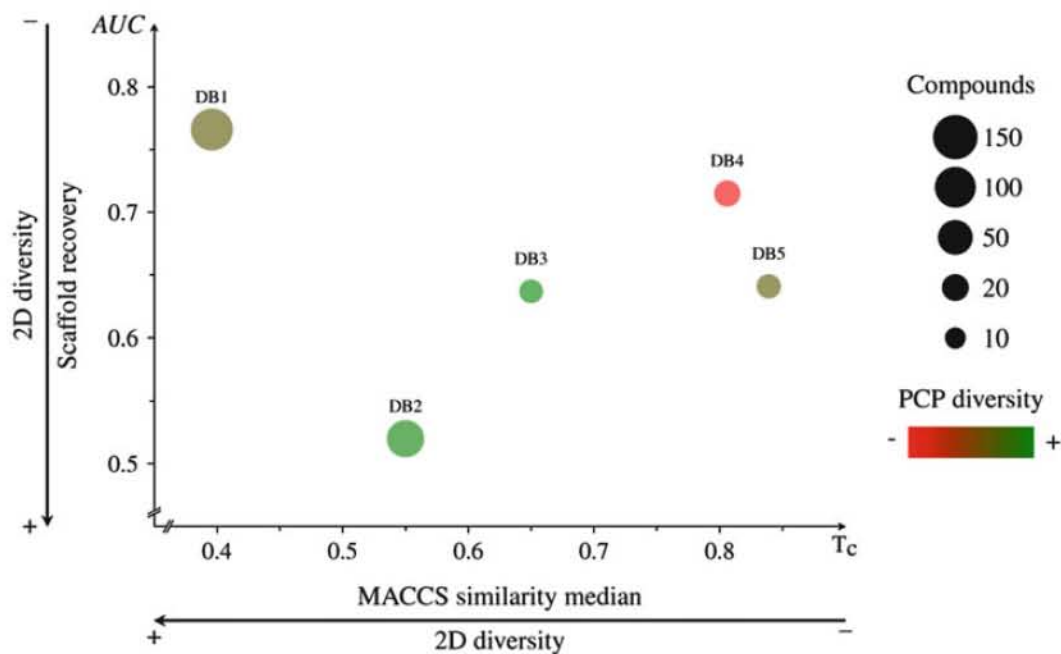
where  $\text{SE} = \sum_{i=1}^n -p_i \log_2 p_i$ ,

where  $p_i$  is the proportion in the dataset of the  $i$ -th scaffold (calculated by dividing the occurrence of this  $i$ -th scaffold by the total number of entries/molecules), SE is the Shannon entropy, and  $n$  is the total number of scaffolds in the dataset. SSE takes as value a real number in the range [0,1]. For this statistic, higher values mean higher scaffold diversity.

4. Finally, the statistic can be computed for other reference databases for comparison.

#### 2.2.4 Consensus Diversity Plots

In the light of numerous variables that can be used to quantify diversity, visual representations have been built in order to summarize multiple of them simultaneously. These are the consensus diversity plots (CDPs). A CDP, as defined by González-Medina et al. [10], renders 2D diversity measured by scaffolds, fingerprints, physicochemical properties, and the number of compounds in the databases. It is also possible to integrate 3D data [24]; however, we will not emphasize on 3D data usage here. The steps required for plotting a CDP from data are:



**Fig. 1** An exemplary consensus diversity plot (CDP). Each data point represents a compound database. Molecular fingerprints diversity is plotted in the  $x$ -axis, the scaffold diversity in the  $y$ -axis, the physicochemical properties diversity in a color continuous scale, and the relative number of compounds in the database as the data point size. *AUC* area under the curve, *PCP* physicochemical properties

1. Curate databases; calculate diversity with physicochemical properties, molecular fingerprints, and scaffolds (see above for details).
2. Plot the molecular fingerprints diversity in the  $x$ -axis, the scaffold diversity in the  $y$ -axis, the physicochemical properties in a color continuous scale, and the number of compounds in the database as the data point size. Every data point represents a database. (See Fig. 1 and Supplementary KNIME Workflow 3 for a few examples.)

As an alternative, an online server was developed for generating CDPs and is also available in D-Tools (see Sect. 1). A video tutorial is available at <https://youtu.be/lruo1ypKGBE>, and detailed written instructions about how to use it can be found at <http://132.248.103.152:3838/CDPlots/>.

### 2.3 Structure–Activity Relationship Analysis

A common assumption in virtual screening is that similar molecules are expected to have similar properties, e.g., comparable biological activity. This assumption is called the similarity principle. Although virtual screening is often useful for detecting active compounds, it is reassuring to verify whether the similarity principle is valid for the molecules under scrutiny. Such insights can be obtained through a subtype of SAR analysis, activity landscape modeling. SAR analysis of chemical libraries, for which activity against a biological target is

known, can also reveal substructures that are relevant for inhibiting the target in question. The next paragraphs give details onto some useful methods for assessing SAR of single and multiple libraries simultaneously. Workflow 4 in the Supplementary Information illustrates a KNIME implementation of the methods described below.

### 2.3.1 Structure–Activity Similarity Maps

Structure–activity similarity (SAS) maps are bidimensional activity landscape representations that contrast structural similarity (e.g., measured with Tanimoto coefficient of molecular fingerprints) and activity similarity (for example, as pIC<sub>50</sub> or p*K*<sub>i</sub>). Systematic pairwise compound comparisons are included in the plot [39]. Each point in a SAS map represents a pair of compounds and is colored according to the most active compound of the pair. The sequence of steps for generating and ultimately interpreting a SAS map is as follows:

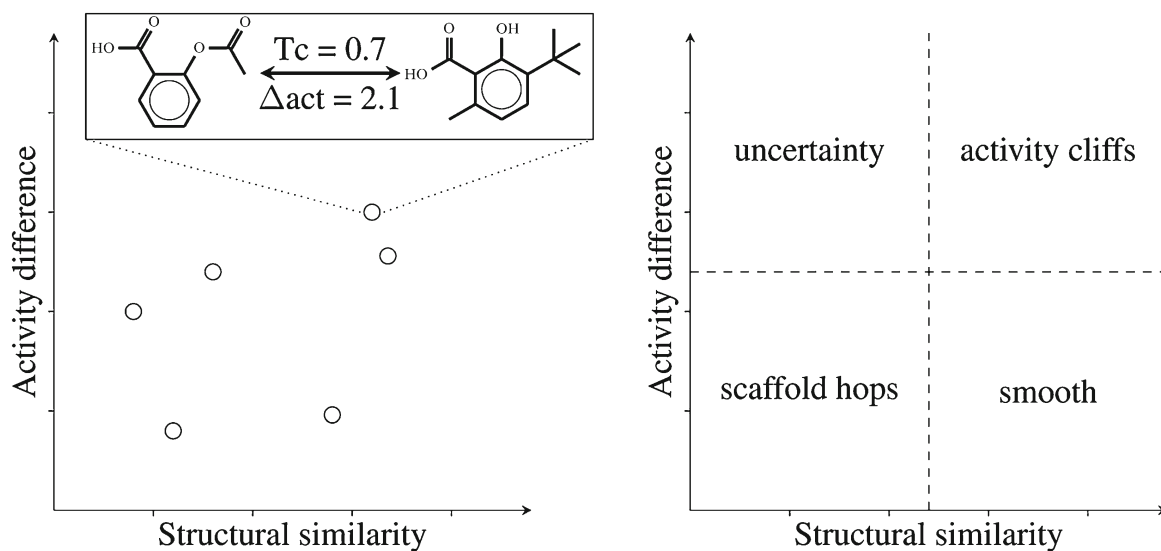
1. Given  $n$  compounds in a library, compute the  $n \times (n - 1)/2$  paired chemical similarity as described in Sect. 2.2.2.
2. Similarly, for the same paired comparisons calculate the absolute difference in potency. All compounds should have potency in pIC<sub>50</sub> units. It is calculated from IC<sub>50</sub> measurements in nanomolar concentration units with the formula (ideally, all compounds should have IC<sub>50</sub> values measured under the same protocol and assay conditions):

$$\text{pIC}_{50} = -\log_{10}(\text{IC}_{50} [\text{nM}]).$$

3. Plot the structural similarity in the  $x$ -axis and the potency difference in the  $y$ -axis. The color of the data points can also be set to render more information, for example, the maximum potency value in the pair.
4. The resultant plot, illustrated in Fig. 2, can be divided into four quadrants with thresholds defined a priori: (a) smooth (high structural similarity and low activity difference), (b) activity cliffs (high structural similarity but high activity difference), (c) scaffold hops (low structural similarity but low activity difference), and (d) uncertainty (low structural similarity and high activity difference) [40–42]. Typical potency thresholds are 2 for deep activity cliffs and 1 for shallow activity cliffs. In the case of structural similarity, 1 or 2 standard deviations above the mean could be used.

Alternatively, a web application for plotting SAS maps can be found at D-Tool under <https://unam-shiny-difacquim.shinyapps.io/ActLSmaps/>. A video tutorial is available at <https://youtu.be/52jHCcg5mXU>.





**Fig. 2** Structure–activity similarity (SAS) maps. Each data point represents a pair of compounds. The *x*-axis plots the structural similarity, while the *y*-axis plots the activity difference. Four quadrants are formed as described in Sect. 2.3.1. A color scale might be added to represent density of points or the maximum activity value in the pair. *Tc* Tanimoto coefficient

### 2.3.2 Scaffold Enrichment Factor

SAR can also be explored based on chemical scaffolds. For every dataset with activity annotations against a particular biological target, every scaffold could be considered as a cluster of molecules. At this point, it is interesting to find which clusters have a higher or lesser proportion of active molecules, pointing towards clusters of highly related molecules that tend to be more or less active than the average. This is the basis of the calculation of enrichment factors (EF) for scaffolds, which are obtained as follows:

1. If activity is represented quantitatively in the dataset, a threshold of activity should be set a priori. Often, a  $pIC_{50}$  of 5–6 or more is useful for defining a compound as active.
2. Essentially, the EF is an odds ratio, i.e., a ratio of proportions. Specifically, the proportion of active compounds with a given scaffold is divided by the proportion of active compounds in the general dataset. A more formal definition would be that, for every scaffold  $\lambda$ , an EF is calculated using the equation [43]:

$$EF(C_\lambda) = \frac{Act(C_\lambda)}{Act(C)}$$

$$\text{where } Act(C_\lambda) = \frac{|C_\lambda^+|}{|C_\lambda|}$$

$$\text{and } Act(C) = \frac{|C^+|}{|C|},$$

where, in turn,  $C$  is the total number of compounds tested,  $C^+$  the number of compounds active,  $C_\lambda$  the number of total compounds with a scaffold  $\lambda$  tested, and  $C_\lambda^+$  the number of

compounds with a scaffold  $\lambda$  active against the target. Values above 1 imply a positively enriched scaffold (i.e., a scaffold that has a higher proportion of active compounds than the general dataset), while values below 1 have the opposite meaning.

3. EFs are susceptible of hypothesis testing. For finding statistically significant enriched scaffolds, chi-squared tests can be run using a  $2 \times 2$  contingency table for the compounds considering as variables whether they have a given scaffold and whether they are active. Since sometimes values in the cells might be lesser than 5, and this interferes with the analytic calculation of the chi-squared statistic, simulated values can be obtained.
4. After running all  $p$ -values for every scaffold, the false discovery rate correction (or other method for correcting for multiple hypothesis testing) should be applied.

### 2.3.3 Degree of Polypharmacology

The methods for SAR analysis mentioned above are useful for single target studies. However, sometimes inhibition data of multiple targets are available for single compounds. These data could lead to polypharmacology studies. Maggiora and Gokhale recently formalized the notion of polypharmacology and polyspecificity [44]. In practical terms, the degree of polypharmacology of a molecule equals the number of different targets against which the molecule is active, while the analogous degree of polyspecificity of a target equals the number of different molecules that are active against the target.

### 2.3.4 Multiple Structure–Activity Relationship Analysis

A review addressing SmARt analysis in epigenetics was recently published [3]. Two of the most useful SmARt tools are methodologically explained in the following paragraphs: dual-activity difference (DAD) maps and structure–promiscuity index difference (SPID). Similarly as for other SAR analyses, Workflow 4 in the Supplementary Information contains practical tools for computing them.

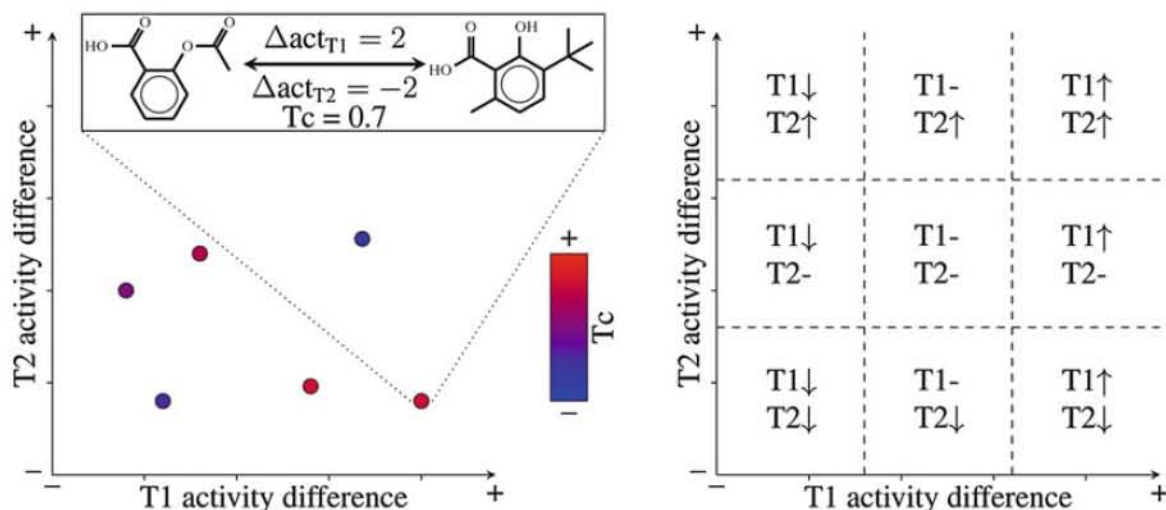
#### Dual-Activity Difference Maps

DAD maps are designed to compare at once the activity of compounds against two biological endpoints, in contrast to SAS maps [45]. However, DAD maps lose structural information, which is accounted for with SAS maps. The procedure for generating a DAD map is straightforward:

1. Select a library of compounds with the activity of each independently measured against two different endpoints.
2. Plot in the  $x$ -axis one of the measurements and on the  $y$ -axis the other. A general form of a DAD map is presented in Fig. 3.

#### Structure–Promiscuity Index Difference

Aiming towards a statistic for quantifying the relationship between structural similarity and polypharmacology (or promiscuity), the SPID was created [46]. It is computed with the formula:



**Fig. 3** Dual-activity difference (DAD) maps. Each data point represents a pair of compounds. The x-axis plots the activity difference of target 1, while the y-axis the activity difference of target 2. A color continuous scale might be added to the plot to represent chemical similarity of each pair of compounds. Up to nine regions can be distinguished depending on whether activity is conserved, increased, or decreased for any of the two targets.  $T_c$  Tanimoto coefficient,  $T1$  target 1,  $T2$  target 2

$$SPID(A, B) = \frac{|P_A - P_B|}{1 - T(A, B)}$$

where A and B are chemical compounds,  $P_A$  and  $P_B$  are the potencies of compounds A and B, respectively, and  $T(A, B)$  is the Tanimoto similarity of compounds A and B computed as in Sect. 2.2.2 using molecular fingerprints.

### 3 Chemical Space

Visual representations of the relationships of the compounds in a database are often useful for assessing libraries' diversity and SAR. Furthermore, the recent development of database fingerprints (DFPs) (described below) has made easier to chart multiple target-focused libraries in the chemical space, thereby providing polypharmacology insights [24]. Workflow 5 in the Supplementary Information illustrates a KNIME implementation of the methods described in this section.

#### 3.1 Principal Components Analysis for Charting Compounds

There are no universal methods for chemical space representations [47, 48]. A commonly used approach involves calculating similarity matrices, which capture all the pairwise comparisons. These matrices are squared and have  $n$  columns and rows, with  $n$  equal to the number of compounds in the dataset. Finally, principal components analysis (PCA) as well as other dimensionality reduction methods is useful to compress most of the relevant information in a few

variables. This makes possible to obtain visualizations of the chemical space. The concrete steps for creating visualizations of the chemical space using the approach presented above are as follows:

1. Select the set of descriptors with which the similarity or distance will be calculated. Common sets are: physicochemical properties (see Sect. 2.2.1) and molecular fingerprints (see Sect. 2.2.2). Compute the similarity matrix accordingly.
2. Apply PCA to the matrix. Select two or three principal components for plotting. It is useful to consider the percentage of variance captured with each principal component.

This method may become impractical for large datasets (>1000 compounds). See Sect. 3.3 for a chemical space visualization method that is less computationally expensive.

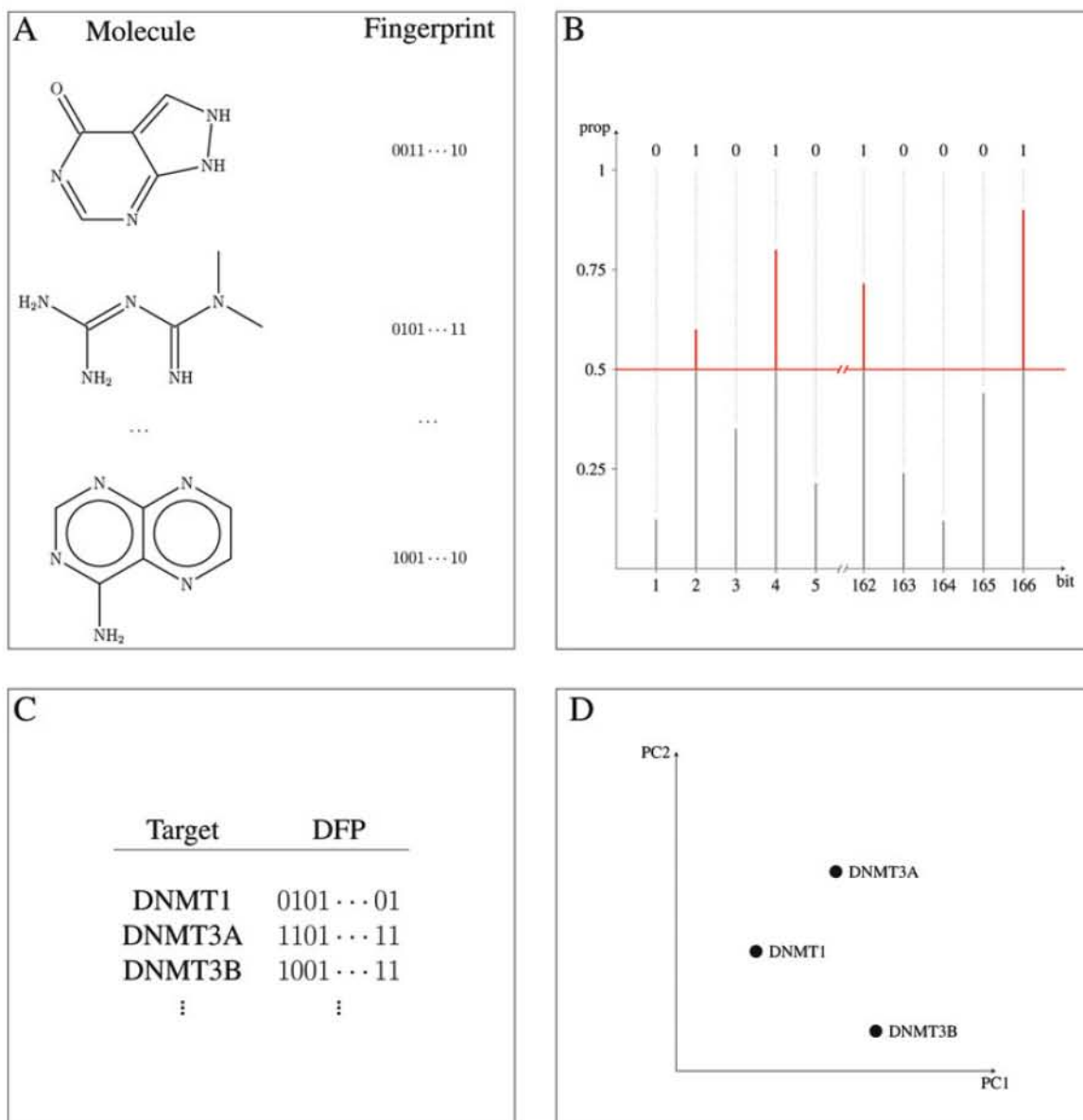
### **3.2 Comparing Multiple Libraries in the Chemical Space**

DFPs are a recently introduced approach to simplify the representation of all compounds in a dataset using a single bit-vector for each database, thereby summarizing every individual fingerprints it contains. DFPs retain the predominant information captured in the molecular fingerprints of the molecules in a given chemical dataset. Briefly, if a given bit had a “1” value in at least 50 % of the compounds in the dataset, it is set to “1” in the DFP, or as “0” otherwise. Further details of the DFPs standardization are described elsewhere [49]. This adds only one step prior to chemical space visualization as commented in Sect. 3.1. If it is intended to include SAR in these plots, libraries could be filtered to include only active compounds. Figure 4 shows schematically the concept of DFPs.

### **3.3 ChemMaps**

Several chemical space visualizations are based upon pairwise similarity measurements. Remarkably, computation of similarity matrices has exponential complexity. Thus, sometimes calculation times make impractical to chart the chemical space of more than 1000 compounds. ChemMaps aim at simplifying the computational task, by adaptively selecting some molecules in the database as comparison references or “satellites.” This method reduces up to 30 % of the time needed for generating a visualization of the chemical space, depending on the size and diversity of the database [50]. The method is as follows:

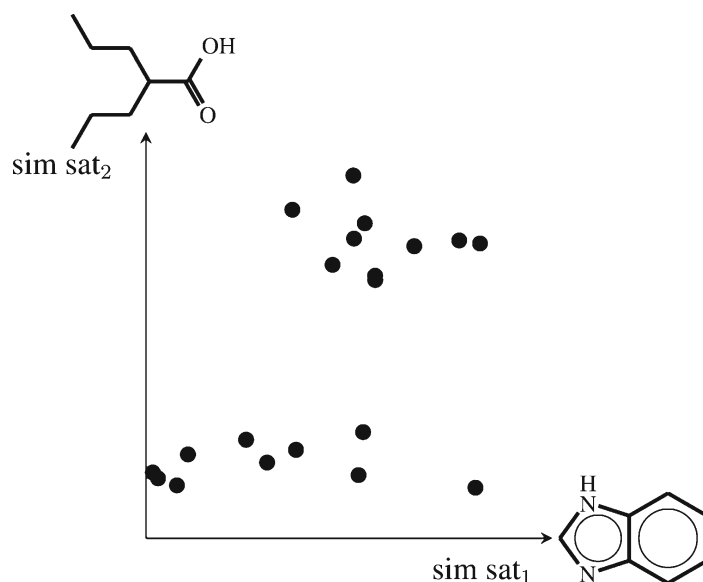
1. Select at random 25 % of the compounds in a library to use as satellites.
2. Compute the pairwise similarity matrix of all the compounds against the satellites.
3. Perform PCA on the matrix and select the first two or three principal components.
4. Using the principal components as descriptors, compute the distance matrix for all the charted compounds or a subset.



**Fig. 4** Database fingerprint (DFP). (a) For every compound in a chemical database, different kind of fingerprints might be obtained. (b) Usually, fingerprints store data in bits. If 50 % or more of the compounds in the database have a value of “1” for a given bit, then it is set as “1” in the DFP, otherwise it is set as “0.” (c) This procedure could be applied to many target-focused libraries. (d) DFPs of multiple libraries can be visualized to represent the chemical space of such libraries. DFPs can also be used for other applications, such as virtual screening. *DFP* database fingerprint

5. Add another 5 % of the database compounds to be used as satellites and repeat steps 2–4.
6. Calculate the correlation between the distances obtained with the PCA as descriptors and repeat step 5 until a correlation of 0.9 or higher is achieved.
7. Plot the chemical space. See Fig. 5.





**Fig. 5** ChemMaps concept. Chemical space is charted relative to adaptive chemical satellites. Two satellites are used in the example

### 3.4 Activity Landscape Sweeping

It is common that some structural clusters tend to form when analyzing the chemical space of libraries. Moreover, these clusters may also have different SAR morphologies, with a smoother or rougher application of the similarity principle [11, 51]. The SAR studies and their use for selecting clusters of molecules from a given library are named “activity landscape sweeping.” Such approach is useful to characterize discrete regions in the chemical space where predictive methods that heavily rely upon the similarity principle could be applied. The method is quite straightforward:

1. As a baseline, compute the general SAS map for the whole library as described in Sect. 2.3.1.
2. Plot the chemical space as described in either Sect. 3.1 or 3.3.
3. For defining clusters in the chemical space, apply some method for unsupervised clustering, such as  $k$ -means.  $K$ -means method could use many principal components for defining the clusters. For selecting a number of principal components to use, a rule of thumb is to plot the contribution of variances of the principal components and select the “elbow” of the curve (i.e., the inflexion point whereupon adding more principal components do not significantly add information). Given that  $k$ -means also requires to a priori define the number of clusters, a similar procedure as that for selecting the number of principal components could be applied. However, instead of plotting the variances contribution, the within groups sum of squares is used. However, the number of clusters can also be defined visually by manually adjusting it.

4. Once that clusters of compounds are defined, individual SAS maps per cluster are plotted as described in Sect. 2.3.1.
5. The SAS maps and the proportions of activity cliffs are compared, in order to identify regions with smoother SAR.

---

## 4 Target Fishing

In polypharmacology, the identification of all the likely targets for a given chemical compound is of utmost importance and has been an active area of research in recent years [52]. This problem is known as reverse virtual screening or “target fishing” [53]. There is a plethora of computational approaches applied in this field. Cheminformatics methods are mostly based on the principle of SAR [54] which suggests that similar compounds are likely to overlap between the sets of targets that they show activity against [55].

This identification of targets for a given compound can be carried out based on the similarity it presents with other compounds that are known to be active or inactive against some targets. If quantitative and comparable activity values are available, it is possible to build quantitative structure–activity relationships (QSAR) models [21, 56] for every target of interest. If the activity values are not completely reliable, a better alternative is the use of the categorical form of them to build machine-learning models for clustering and classification [57]. Although the general objective of most of these methodologies is the identification of targets for a given compound, the amount and type of biological information available can lead to various applications. This section describes the methodologies implicated in them.

### 4.1 Target Identification

The most general application of target fishing strategies consists of predicting all the possible targets for a given compound, or at least all of them for which bioactivity data is known. Most of these strategies treat the target fishing problem as a multi-label classification problem, in which every target is a label that a given compound belongs to and for which a predictive model is constructed [52, 58]. The main differences between different approaches are the molecular representation employed and the predictive models used. This work is not intended to provide a detailed description on the construction of these models, which can be found in several other works [59, 60], but of the general strategy for their application.

#### 4.1.1 Multi-label Classifiers

One of the most used alternatives to face the target fishing problem is by building a multi-label classifier. The general steps to build such model are described below:

1. Given a set of targets of interest, a set of compounds, and a defined bipartite activity relation between them, construct and

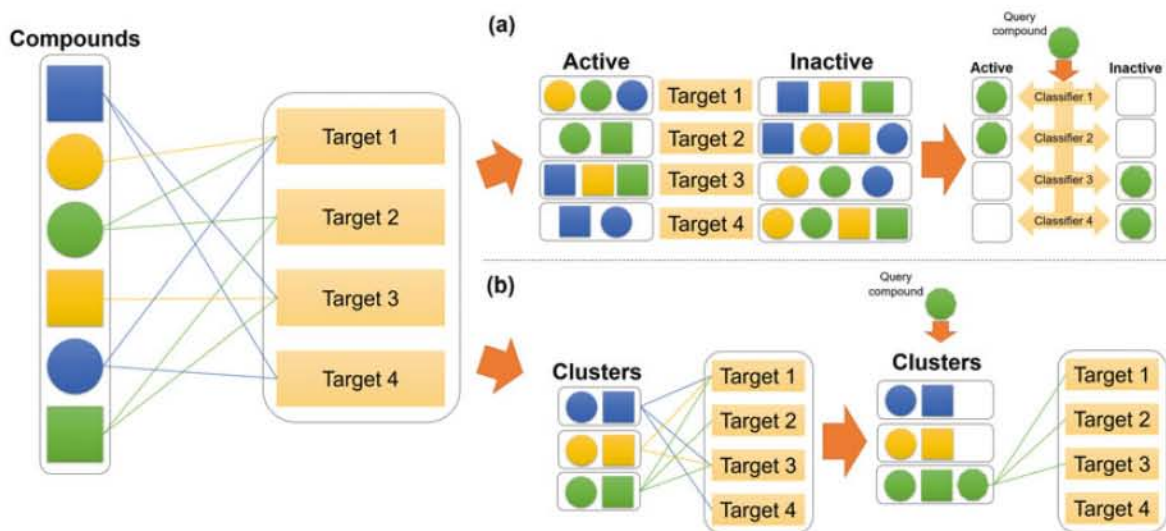
curate compound databases for each target according to the methods discussed in Sect. 2.1.

2. Build and validate a binary classifier for each database, which allows to distinguish between active and inactive compounds. At this point lies the main difference between distinct models, because the pertinence of a compound to one class or another can be defined according to a priori defined thresholds for a given score. For instance, a similarity coefficient when dealing with similarity searches (discussed in Sect. 2.2), an activity value in the case of QSAR models, or the probability coming from a machine-learning model.
3. Finally, evaluate a compound of interest with all binary models. The targets associated to that compound will be those for which the binary classifiers assign a score higher than the established threshold.

The general scheme of a multi-label classifier is presented in Fig. 6a. The application of these types of strategies in drug design projects is discussed in other works [21, 61, 62] and currently there are several web implementations of these methods [58, 63].

#### 4.1.2 Cluster Analysis

Another methodology to address the multi-label classification problem of target fishing is clustering, which is the task of grouping objects (compounds) such a way that objects belonging to the same group are more similar to each other in comparison to those belonging to other groups. This kind of methodologies only take into account the structure and properties of compounds known to



**Fig. 6** (a) Representation of a multi-label classifier. The targets associated to the query compound are those for which the corresponding classifiers identify them in the active class. (b) Representation of a clustering analysis. The targets associated to the query compound are those associated to the cluster in which such compound is grouped

be active against each target of interest. The general strategy is as follows:

1. Given a set of targets of interest, a set of compounds, and a defined bipartite activity relation between them, construct and curate a database considering only the compounds known to be active against at least one target.
2. Split the compound database into multiple groups by using a clustering algorithm. This grouping task can be performed according to different criteria, for example, by scaffolds or by molecular similarity, discussed in Sect. 2.2, or employing a machine-learning algorithm like  $k$ -means, discussed in Sect. 3.4.
3. For each cluster, identify all the targets against which at least one compound in the cluster is active, those will be the targets associated to that cluster.
4. Finally, assign a compound of interest to one cluster by using the same criteria involved in step 2, the targets associated to that cluster will be the predicted targets for the query compound.

Figure 6b presents the general scheme of a cluster analysis. Recent applications of this type of approaches in different research areas and web implementations are discussed in other publications [64, 65].

#### 4.2 Target Deconvolution

Although the knowledge of compounds with activity against one or several targets is fundamental for the development of the strategies presented in Sect. 4.1, these are not the only bioactivity data available. In addition to data from target-based methodologies, the amount of data from cell-based phenotypic screenings has increased considerably in recent years [66]. One of the major advantages of this kind of information is that it provides a more direct view of the responses taking place in the context of a complex biological system, such as a cell [67].

Identifying the molecular targets of active hits from phenotypic screens is a required process to understand the mechanisms of action involved and thus direct the optimization of such compounds. This task is referred as target deconvolution and the cheminformatic approaches to address the problem are essentially the same as those presented in the previous section, with the major difference being that the set of targets to analyze is reduced to those relevant for the phenotype under study [64, 68].

---

## 5 Future Prospects

The increasing awareness of polypharmacology in drug discovery and developments will continue demanding the application of cheminformatics approaches to accelerate the process. Computational

methods initially developed for drug discovery focused on a single target are being adapted to develop compounds for multiple targets. Typical examples are SMART and inverse virtual screening or target fishing. In this regard, it is expected that such approaches are further refined to improve accuracy. It is also expected that new computational approaches will emerge to boost the development of polypharmacological drugs.

---

## Acknowledgements

This work was supported by the Programa de Apoyo a Proyectos de Investigación e Innovación Tecnológica (PAPIIT) grant IA203718 and National Council of Science and Technology (CONACyT), Mexico grant number 282785. JJN, FIS-G, and NS-C are thankful to CONACyT for the granted scholarships number 622969, 629458, and 335997, respectively.

## References

1. Rosini M (2014) Polypharmacology: the rise of multitarget drugs over combination therapies. *Future Med Chem* 6:485–487. <https://doi.org/10.4155/fmc.14.25>
2. Méndez-Lucio O, Naveja JJ, Vite-Caritino H et al (2016) Review. One drug for multiple targets: a computational perspective. *J Mex Chem Soc* 60:168–181
3. Saldívar-González FI, Naveja JJ, Palomino-Hernández O, Medina-Franco JL (2017) Getting SMART in drug discovery: chemoinformatics approaches for mining structure–multiple activity relationships. *RSC Adv* 7:632–641. <https://doi.org/10.1039/C6RA26230A>
4. González-Medina M, Naveja JJ, Sánchez-Cruz N, Medina-Franco JL (2017) Open chemoinformatic resources to explore the structure, properties and chemical space of molecules. *RSC Adv* 7:54153–54163. <https://doi.org/10.1039/C7RA11831G>
5. Berthold MR, Cebron N, Dill F et al (2009) KNIME – the Konstanz information miner. *SIGKDD Explor Newsl* 11:26. <https://doi.org/10.1145/1656274.1656280>
6. Varnek A (2017) Tutorials in chemoinformatics. <https://doi.org/10.1002/9781119161110>
7. Saldívar-González FI, Hernández-Luis F, Lira-Rocha A, Medina-Franco JL (2017) Manual de Quimioinformática, 1st edn. Universidad Nacional Autónoma de México, Mexico City
8. González-Medina M, Medina-Franco JL (2017) Platform for unified molecular analysis: PUMA. *J Chem Inf Model* 57:1735–1740. <https://doi.org/10.1021/acs.jcim.7b00253>
9. González-Medina M, Méndez-Lucio O, Medina-Franco JL (2017) Activity landscape plotter: a web-based application for the analysis of structure-activity relationships. *J Chem Inf Model* 57:397–402. <https://doi.org/10.1021/acs.jcim.6b00776>
10. González-Medina M, Prieto-Martínez FD, Owen JR, Medina-Franco JL (2016) Consensus diversity plots: a global diversity analysis of chemical libraries. *J Cheminform* 8:63. <https://doi.org/10.1186/s13321-016-0176-9>
11. Naveja JJ, Oviedo-Osornio CI, Trujillo-Minero NN, Medina-Franco JL (2017) Chemoinformatics: a perspective from an academic setting in Latin America. *Mol Divers*. <https://doi.org/10.1007/s11030-017-9802-3>
12. Richter L, Ecker GF (2015) Medicinal chemistry in the era of big data. *Drug Discov Today Technol* 14:37–41. <https://doi.org/10.1016/j.ddtec.2015.06.001>
13. Law V, Knox C, Djoumbou Y et al (2014) Drug-Bank 4.0: shedding new light on drug metabolism. *Nucleic Acids Res* 42:D1091–D1097. <https://doi.org/10.1093/nar/gkt1068>
14. Bento AP, Gaulton A, Hersey A et al (2014) The ChEMBL bioactivity database: an update. *Nucleic Acids Res* 42:D1083–D1090. <https://doi.org/10.1093/nar/gkt1031>
15. Irwin JJ, Shoichet BK (2005) ZINC – a free database of commercially available



- compounds for virtual screening. *J Chem Inf Model* 45:177–182. <https://doi.org/10.1021/ci049714+>
16. Liu T, Lin Y, Wen X et al (2007) BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. *Nucleic Acids Res* 35:D198–D201. <https://doi.org/10.1093/nar/gkl999>
  17. Lavecchia A, Cerchia C (2016) In silico methods to address polypharmacology: current status, applications and future perspectives. *Drug Discov Today* 21:288–298. <https://doi.org/10.1016/j.drudis.2015.12.007>
  18. Fourches D, Muratov E, Tropsha A (2016) Trust, but verify II: a practical guide to chemogenomics data curation. *J Chem Inf Model* 56:1243–1252. <https://doi.org/10.1021/acs.jcim.6b00129>
  19. Hersey A, Chambers J, Bellis L et al (2015) Chemical databases: curation or integration by user-defined equivalence? *Drug Discov Today Technol* 14:17–24. <https://doi.org/10.1016/j.ddtec.2015.01.005>
  20. Miller MA (2002) Chemical database techniques in drug discovery. *Nat Rev Drug Discov* 1:220–227. <https://doi.org/10.1038/nrd745>
  21. Cherkasov A, Muratov EN, Fourches D et al (2014) QSAR modeling: where have you been? Where are you going to? *J Med Chem* 57:4977–5010. <https://doi.org/10.1021/jm4004285>
  22. Mansouri K, Abdelaziz A, Rybacka A et al (2016) CERAPP: collaborative estrogen receptor activity prediction project. *Environ Health Perspect* 124:1023–1033. <https://doi.org/10.1289/ehp.1510267>
  23. Gally J-M, Bourg S, Do Q-T et al (2017) Vsprep: a general KNIME workflow for the preparation of molecules for virtual screening. *Mol Inform*. <https://doi.org/10.1002/minf.201700023>
  24. Naveja JJ, Medina-Franco JL (2017) Insights from pharmacological similarity of epigenetic targets in epipolypharmacology. *Drug Discov Today*. <https://doi.org/10.1016/j.drudis.2017.10.006>
  25. Medina-Franco JL, Martinez-Mayorga K, Meurice N (2014) Balancing novelty with confined chemical space in modern drug discovery. *Expert Opin Drug Discov* 9:151–165. <https://doi.org/10.1517/17460441.2014.872624>
  26. Sheridan RP, Kearsley SK (2002) Why do we need so many chemical similarity search methods? *Drug Discov Today* 7:903–911. [https://doi.org/10.1016/S1359-6446\(02\)02411-X](https://doi.org/10.1016/S1359-6446(02)02411-X)
  27. Medina-Franco JL, Maggiora GM (2013) Molecular similarity analysis. In: Bajorath J (ed) *Cheminformatics for drug discovery*. Wiley, Hoboken, NJ, pp 343–399. <https://doi.org/10.1002/9781118742785.ch15>
  28. Singh N, Guha R, Giulianotti MA et al (2009) Cheminformatic analysis of combinatorial libraries, drugs, natural products, and molecular libraries small molecule repository. *J Chem Inf Model* 49:1010–1024. <https://doi.org/10.1021/ci800426u>
  29. Xu J, Hagler A (2002) Chemoinformatics and drug discovery. *Molecules* 7:566–600. <https://doi.org/10.3390/70800566>
  30. Gortari EF, Medina-Franco JL (2015) Epigenetic relevant chemical space: a cheminformatic characterization of inhibitors of DNA methyltransferases. *RSC Adv* 5:87465–87476. <https://doi.org/10.1039/C5RA19611F>
  31. Eckert H, Bajorath J (2007) Molecular similarity analysis in virtual screening: foundations, limitations and novel approaches. *Drug Discov Today* 12:225–233. <https://doi.org/10.1016/j.drudis.2007.01.011>
  32. Rogers D, Hahn M (2010) Extended-connectivity fingerprints. *J Chem Inf Model* 50:742–754. <https://doi.org/10.1021/ci100050t>
  33. Ewing T, Baber JC, Feher M (2006) Novel 2D fingerprints for ligand-based virtual screening. *J Chem Inf Model* 46:2423–2431. <https://doi.org/10.1021/ci060155b>
  34. 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>
  35. Jaccard P (1901) Etude Comparative de la Distribution Florale dans une Portion des Alpes et des Jura. *Bull Soc Vaudoise Sci Nat* 37:547–579
  36. Bemis GW, Murcko MA (1996) The properties of known drugs. I. Molecular frameworks. *J Med Chem* 39:2887–2893. <https://doi.org/10.1021/jm9602928>
  37. Xu Y-J, Johnson M (2002) Using molecular equivalence numbers to visually explore structural features that distinguish chemical libraries. *J Chem Inf Comput Sci* 42:912–926
  38. Medina-Franco J, MartÃ-nez-Mayorga K, Bender A, Scior T (2009) Scaffold diversity analysis of compound data sets using an entropy-based measure. *QSAR Comb Sci* 28:1551–1560. <https://doi.org/10.1002/qsar.200960069>
  39. Shanmugasundaram V, Maggiora GM (2001) Characterizing property and activity landscapes using an information-theoretic approach. CINF-032. In 222nd ACS National Meeting, Chicago, IL, United States; August 26–30,

- 2001; American Chemical Society, Washington, DC, 2001.
40. Guha R (2012) Exploring structure-activity data using the landscape paradigm. Wiley Interdiscip Rev Comput Mol Sci. <https://doi.org/10.1002/wcms.1087>
  41. Bajorath J, Peltason L, Wawer M et al (2009) Navigating structure-activity landscapes. Drug Discov Today 14:698–705. <https://doi.org/10.1016/j.drudis.2009.04.003>
  42. Medina-Franco JL (2012) Scanning structure-activity relationships with structure-activity similarity and related maps: from consensus activity cliffs to selectivity switches. J Chem Inf Model 52:2485–2493. <https://doi.org/10.1021/ci300362x>
  43. Medina-Franco JL, Petit J, Maggiora GM (2006) Hierarchical strategy for identifying active chemotype classes in compound databases. Chem Biol Drug Des 67:395–408. <https://doi.org/10.1111/j.1747-0285.2006.00397.x>
  44. Maggiora G, Gokhale V (2017) A simple mathematical approach to the analysis of polypharmacology and polyspecificity data. [version 1; referees: 3 approved, 1 approved with reservations]. F1000Res. <https://doi.org/10.12688/f1000research.11517.1>
  45. Pérez-Villanueva J, Santos R, Hernández-Campos A et al (2011) Structure–activity relationships of benzimidazole derivatives as anti-parasitic agents: dual activity-difference (DAD) maps. Med Chem Commun 2:44–49. <https://doi.org/10.1039/C0MD00159G>
  46. Yongye AB, Medina-Franco JL (2012) Data mining of protein-binding profiling data identifies structural modifications that distinguish selective and promiscuous compounds. J Chem Inf Model 52:2454–2461. <https://doi.org/10.1021/ci3002606>
  47. Osolodkin DI, Radchenko EV, Orlov AA et al (2015) Progress in visual representations of chemical space. Expert Opin Drug Discov 10:959–973. <https://doi.org/10.1517/17460441.2015.1060216>
  48. Medina-Franco J, Martinez-Mayorga K, Giulianotti M et al (2008) Visualization of the chemical space in drug discovery. Curr Comput Aided Drug Des 4:322–333. <https://doi.org/10.2174/157340908786786010>
  49. Fernández-de Gortari E, García-Jacas CR, Martínez-Mayorga K, Medina-Franco JL (2017) Database fingerprint (DFP): an approach to represent molecular databases. J Cheminform 9:9. <https://doi.org/10.1186/s13321-017-0195-1>
  50. Naveja JJ, Medina-Franco JL (2017) Chem-Maps: towards an approach for visualizing the chemical space based on adaptive satellite compounds [version 1; referees: 1 approved, 2 approved with reservations]. F1000Res. <https://doi.org/10.12688/f1000research.12095.1>
  51. Naveja JJ, Medina-Franco JL (2015) Activity landscape sweeping: insights into the mechanism of inhibition and optimization of DNMT1 inhibitors. RSC Adv 5:63882–63895. <https://doi.org/10.1039/C5RA12339A>
  52. Wale N, Karypis G (2009) Target fishing for chemical compounds using target-ligand activity data and ranking based methods. J Chem Inf Model 49:2190–2201. <https://doi.org/10.1021/ci9000376>
  53. Jenkins JL, Bender A, Davies JW (2006) In silico target fishing: predicting biological targets from chemical structure. Drug Discov Today Technol 3:413–421. <https://doi.org/10.1016/j.ddtec.2006.12.008>
  54. Hansch C, Maloney PP, Fujita T, Muir RM (1962) Correlation of biological activity of phenoxyacetic acids with hammett substituent constants and partition coefficients. Nature 194:178–180. <https://doi.org/10.1038/194178b0>
  55. Nettles JH, Jenkins JL, Bender A et al (2006) Bridging chemical and biological space: “target fishing” using 2D and 3D molecular descriptors. J Med Chem 49:6802–6810. <https://doi.org/10.1021/jm060902w>
  56. Cramer RD (2012) The inevitable QSAR renaissance. J Comput Aided Mol Des 26:35–38. <https://doi.org/10.1007/s10822-011-9495-0>
  57. Lavecchia A (2015) Machine-learning approaches in drug discovery: methods and applications. Drug Discov Today 20:318–331. <https://doi.org/10.1016/j.drudis.2014.10.012>
  58. Yao Z-J, Dong J, Che Y-J et al (2016) Target-Net: a web service for predicting potential drug-target interaction profiling via multi-target SAR models. J Comput Aided Mol Des 30:413–424. <https://doi.org/10.1007/s10822-016-9915-2>
  59. Nidhi, Glick M, Davies JW, Jenkins JL (2006) Prediction of biological targets for compounds using multiple-category Bayesian models trained on chemogenomics databases. J Chem Inf Model 46:1124–1133. <https://doi.org/10.1021/ci060003g>

60. Kawai K, Fujishima S, Takahashi Y (2008) Predictive activity profiling of drugs by topological-fragment-spectra-based support vector machines. *J Chem Inf Model* 48:1152–1160. <https://doi.org/10.1021/ci7004753>
61. Nikolic K, Mavridis L, Djikic T et al (2016) Drug design for CNS diseases: polypharmacological profiling of compounds using cheminformatic, 3D-QSAR and virtual screening methodologies. *Front Neurosci* 10:265. <https://doi.org/10.3389/fnins.2016.00265>
62. Rognan D (2010) Structure-based approaches to target fishing and ligand profiling. *Mol Inform* 29:176–187. <https://doi.org/10.1002/minf.200900081>
63. Awale M, Reymond J-L (2017) The polypharmacology browser: a web-based multi-fingerprint target prediction tool using ChEMBL bioactivity data. *J Cheminform* 9:11. <https://doi.org/10.1186/s13321-017-0199-x>
64. Kunimoto R, Dimova D, Bajorath J (2017) Application of a new scaffold concept for computational target deconvolution of chemical cancer cell line screens. *ACS Omega* 2:1463–1468. <https://doi.org/10.1021/acsomega.7b00215>
65. Reker D, Rodrigues T, Schneider P, Schneider G (2014) Identifying the macromolecular targets of de novo-designed chemical entities through self-organizing map consensus. *Proc Natl Acad Sci U S A* 111:4067–4072. <https://doi.org/10.1073/pnas.1320001111>
66. Zheng W, Thorne N, McKew JC (2013) Phenotypic screens as a renewed approach for drug discovery. *Drug Discov Today* 18:1067–1073. <https://doi.org/10.1016/j.drudis.2013.07.001>
67. Lee J, Bogyo M (2013) Target deconvolution techniques in modern phenotypic profiling. *Curr Opin Chem Biol* 17:118–126. <https://doi.org/10.1016/j.cbpa.2012.12.022>
68. Mugumbate G, Mendes V, Blaszczyk M et al (2017) Target identification of mycobacterium tuberculosis phenotypic hits using a concerted chemogenomic, biophysical, and structural approach. *Front Pharmacol* 8:681. <https://doi.org/10.3389/fphar.2017.00681>

## 7. Discusión, conclusiones y perspectivas

Se realizó una búsqueda de pares de dianas relevantes en líneas celulares de cáncer (ver Sección 4). Estos pares de dianas son interesantes porque los compuestos polifarmacológicos que los inhiben son todos activos contra las células, mientras que existen compuestos inactivos que inhiben sólo a una de las dianas del par; esto resalta la importancia de la inhibición combinada. La relevancia de los pares de dianas identificados se confirmó al observar que los mismos se asocian con combinaciones sinérgicas de compuestos, es decir, las combinaciones de compuestos en los que el par de dianas en cuestión es inhibido tienen una mayor tendencia a mostrar sinergia. Para explicar los mecanismos de sinergia se utilizó un modelo de redes de interacción de proteínas, que proporcionó información biológica interpretable.

Además, se obtuvo información valiosa acerca de la similitud farmacológica de 52 dianas epigenéticas (Sección 5). Esto implicó el estudio de espacio químico más amplio hasta el momento realizado en dianas epigenéticas. De manera interesante, la información contenida en las bibliotecas de inhibidores epigenéticos puede ser utilizada como referencia para comparar a las dianas desde un punto de vista farmacológico. Por ejemplo, las bibliotecas de dianas que ejercen funciones semejantes, tienden a ser similares, y sus bibliotecas de inhibidores se agrupan en el espacio químico. Posteriormente, se identificó una asociación entre ciertas regiones del espacio químico y una relación estructura-actividad más directa. Esto permitió identificar dianas epigenéticas para las cuales los métodos predictivos pueden funcionar mejor.

Además, en el curso del proyecto se encontraron oportunidades para desarrollar, aplicar y difundir nuevas metodologías relacionadas con la exploración del espacio químico y el estudio de la polifarmacología (Sección 6). Específicamente, se desarrollaron métodos de identificación, análisis y visualización de series de análogos, así como otras técnicas de exploración del espacio químico, como los gráficos de constelaciones (*constellation plots*) ChemMaps y el barrido de panoramas de actividad (*activity landscape sweeping*). En relación con el análisis de series de análogos, se creó una nueva metodología para identificar, de forma eficiente y consistente, moléculas “análogas”, es decir, que comparten núcleos base. Del mismo modo, se desarrollaron aplicaciones de esta estrategia en el estudio de las relaciones estructura-actividad (análisis CSAR) y en la visualización de ensayos de alto rendimiento (gráficos de constelaciones). Por su parte, ChemMaps es una alternativa más rápida para representar el espacio químico, ya que se basa en definir compuestos “satélites” que sirven como referencia, lo cual simplifica los cálculos. ChemMaps fue uno de los análisis que se aplicaron en el análisis de una biblioteca de compuestos presentes en alimentos. En cuanto al barrido de panoramas de actividad, este combina el espacio químico con los estudios de relación estructura-actividad. Se desarrolló para separar las regiones del espacio químico donde la relación estructura-actividad parece cumplirse de las que no. Este método se aplicó en el análisis de una biblioteca química de xenoestrógenos, donde se identificaron a los derivados de esteroides y flavonas como las estructuras con la relación estructura-actividad más débil.

Dentro de las perspectivas del proyecto se incluye la exploración del efecto biológico y molecular de inhibir múltiples dianas epigenéticas de manera simultánea. Por otra parte, se buscarán colaboraciones experimentales para probar las predicciones. Finalmente, los métodos desarrollados se pueden aplicar en otros modelos biológicos diferentes al cáncer (p.ej., síndrome metabólico).

## Agradecimientos

Agradezco al Plan de Estudios Combinados en Medicina (PECEM) y a su coordinadora, la Dra. Ana Flisser, por la oportunidad, única en el mundo, que plantea a los estudiantes de medicina para que puedan realizar investigación a la par del aprendizaje clínico; estoy convencido que, de no ser por el PECEM, no habría tenido motivación suficiente para terminar la licenciatura en Medicina. Agradezco a la Universidad Nacional Autónoma de México (UNAM) por brindarme una educación realmente universal; por abrir, a través de las dos estancias en el extranjero que me permitió, mi panorama académico más allá de las fronteras físicas de nuestro país, así como por las incontables oportunidades que me proporcionó para interactuar con profesores y alumnos de nivel académico mucho mayor al mío: de esta forma me enseñó el valor de la humildad y a aprender de –y, cuando sea posible, enseñar a– los demás. Agradezco al Consejo Nacional de Ciencia y Tecnología (CONACyT) por la beca de doctorado número 622969, así como por la beca mixta que me permitió hacer una estancia en la Universidad de Bonn, Alemania.

Agradezco especialmente a mi tutor, José Luis Medina Franco, por las enseñanzas brindadas y su apoyo durante la realización de este proyecto. (¡Qué suerte que regresó a México justo a tiempo para que fuera mi tutor!) También muchas gracias a los estudiantes del grupo de investigación DIFACQUIM, por su amistad y por compartirme sus conocimientos de la Química; en particular gracias a Eli Fernández, Óscar Méndez, Norberto Sánchez, Fernando Prieto, Mariana González, Fernanda Saldívar, Oscar Palomino y Bárbara Díaz. Del mismo modo, gracias a mis compañeros del PECEM, porque pude transitar este camino compartiendo con ellos: en particular a Leonardo Zapata, Marco Tapia, Eduardo Cervantes, Omar Bello y Mauricio Ostrosky. Igualmente agradezco a mis compañeros de la Facultad de Medicina, con quienes he compartido muy gratos momentos y el aprendizaje de la clínica: Ricardo Espinosa, Víctor Marín, Diego Ángeles, Enrique López, Javier Villalpando, Javier Hernández. A Susy le agradezco por los momentos bonitos y por su compañía y cariño durante los momentos difíciles. También agradezco a los tutores con los que tuve oportunidad de realizar estancias durante mi formación en la investigación, en especial a Sergio Guerrero, Mario Calcagno, Flavio Contreras, Luis Herrera, Diego Oliva, Rodrigo Cáceres, Carlos Cirlos, Alfredo Rodríguez, Mónica Campillos y Jürgen Bajorath; haber aprendido de cada uno de ellos fue verdaderamente un privilegio. En general, gracias a todos y cada uno de los que hicieron posible que disfrutara tanto mi estancia en la UNAM y en la Ciudad de México. Gracias a mi comité tutor (Iwin Leenen, José Díaz, Maximino Aldana y José Correa) y al comité sinodal (Rafael Castillo, Karina Martínez, Luis Herrera, Félix Recillas) porque las valiosas aportaciones de cada uno de ellos permitieron mejorar el escrito final de la tesis. Finalmente, agradezco a mis padres, Jesús Naveja Macías y Gemma Romero, y a mis hermanos, Gemma y Miguel Naveja, por su apoyo incondicional.



## Referencias

- [1] Moffat JG, Rudolph J, Bailey D. Phenotypic screening in cancer drug discovery—past, present and future. *Nat Rev Drug Discov.* 2014;13(8):588.
- [2] Hopkins AL, Mason JS, Overington JP. Can we rationally design promiscuous drugs? *Curr Opin Struct Biol.* 2006;16(1):127–136.
- [3] Chen X, Ji ZL, Chen YZ. TTD: therapeutic target database. *Nucleic Acids Res.* 2002;30(1):412–415.
- [4] Zagidullin B, Aldahdooh J, Zheng S, Wang W, Wang Y, Saad J, et al. DrugComb: an integrative cancer drug combination data portal. *Nucleic Acids Res.* 2019;47(W1):W43–W51.
- [5] Anighoro A, Bajorath J, Rastelli G. Polypharmacology: Challenges and Opportunities in Drug Discovery. *J Med Chem.* 2014;57(19):7874–7887.
- [6] Guo J, Liu H, Zheng J. SynLethDB: synthetic lethality database toward discovery of selective and sensitive anticancer drug targets. *Nucleic Acids Res.* 2015;44(D1):D1011–D1017.
- [7] Szalay KZ, Csermely P. Perturbation centrality and turbine: a novel centrality measure obtained using a versatile network dynamics tool. *PLoS ONE.* 2013;8(10):e78059.
- [8] Al-Ali H. The evolution of drug discovery: from phenotypes to targets, and back. *Med Chem Commun.* 2016;7(5):788–798.
- [9] Shoemaker RH. The NCI60 human tumour cell line anticancer drug screen. *Nature Reviews Cancer.* 2006;6(10):813.
- [10] Chabner BA. NCI-60 cell line screening: a radical departure in its time. *JNCI: Journal of the National Cancer Institute.* 2016;108(5).
- [11] Holbeck SL, Camalier R, Crowell JA, Govindharajulu JP, Hollingshead M, Anderson LW, et al. The National Cancer Institute ALMANAC: A comprehensive screening resource for the detection of anticancer drug pairs with enhanced therapeutic activity. *Cancer research.* 2017;77(13):3564–3576.
- [12] Gladstone M, Su TT. Chemical genetics and drug screening in *Drosophila* cancer models. *J Genet Genomics.* 2011;38(10):497–504.
- [13] Willoughby LF, Schlosser T, Manning SA, Parisot JP, Street IP, Richardson HE, et al. An in vivo large-scale chemical screening platform using *Drosophila* for anti-cancer drug discovery. *Dis Model Mech.* 2013;6(2):521–529.
- [14] Gao H, Korn JM, Ferretti S, Monahan JE, Wang Y, Singh M, et al. High-throughput screening using patient-derived tumor xenografts to predict clinical trial drug response. *Nat Med.* 2015;21(11):1318.
- [15] Jung HJ, Kwon HJ. Target deconvolution of bioactive small molecules: the heart of chemical biology and drug discovery. *Arch Pharm Res.* 2015;38(9):1627–1641.
- [16] Kunimoto R, Dimova D, Bajorath J. Application of a new scaffold concept for computational target deconvolution of chemical cancer cell line screens. *ACS Omega.* 2017;2(4):1463–1468.

- [17] Liu X, Baarsma HA, Thiam CH, Montrone C, Brauner B, Fobo G, et al. Systematic Identification of Pharmacological Targets from Small-Molecule Phenotypic Screens. *Cell Chem Biol.* 2016;23(10):1302–1313.
- [18] Gayvert KM, Madhukar NS, Elemento O. A Data-Driven Approach to Predicting Successes and Failures of Clinical Trials. *Cell Chem Biol.* 2016;23(10):1294–1301.
- [19] Kuhn M, Al Banchaabouchi M, Campillos M, Jensen LJ, Gross C, Gavin AC, et al. Systematic identification of proteins that elicit drug side effects. *Mol Syst Biol.* 2013;9(1):663.
- [20] Al-Ali H, Lee DH, Danzi MC, Nassif H, Gautam P, Wennerberg K, et al. Rational polypharmacology: systematically identifying and engaging multiple drug targets to promote axon growth. *ACS Chem Biol.* 2015;10(8):1939–1951.
- [21] Campillos M, Kuhn M, Gavin AC, Jensen LJ, Bork P. Drug target identification using side-effect similarity. *Science.* 2008;321(5886):263–266.
- [22] Helal KY, Maciejewski M, Gregori-Puigjané E, Glick M, Wassermann AM. Public Domain HTS Fingerprints: Design and Evaluation of Compound Bioactivity Profiles from PubChem's Bioassay Repository. *J Chem Inf Model.* 2016;56(2):390–398.
- [23] Cowen L, Ideker T, Raphael BJ, Sharan R. Network propagation: a universal amplifier of genetic associations. *Nat Rev Genet.* 2017;18(9):551.
- [24] Allis CD, Jenuwein T. The molecular hallmarks of epigenetic control. *Nat Rev Genet.* 2016;17(8):487.
- [25] Flavahan WA, Gaskell E, Bernstein BE. Epigenetic plasticity and the hallmarks of cancer. *Science.* 2017;357(6348):eaal2380.
- [26] Meng S, Zhang L, Tang Y, Tu Q, Zheng L, Yu L, et al. BET inhibitor JQ1 blocks inflammation and bone destruction. *J Dent Res.* 2014;93(7):657–662.
- [27] Anand P, Brown JD, Lin CY, Qi J, Zhang R, Artero PC, et al. BET bromodomains mediate transcriptional pause release in heart failure. *Cell.* 2013;154(3):569–582.
- [28] Banerjee C, Archin N, Michaels D, Belkina AC, Denis GV, Bradner J, et al. BET bromodomain inhibition as a novel strategy for reactivation of HIV-1. *J Leukoc Biol.* 2012;92(6):1147–1154.
- [29] Heerboth S, Lapinska K, Snyder N, Leary M, Rollinson S, Sarkar S. Use of epigenetic drugs in disease: an overview. *Genet Epigenet.* 2014;6:9–19.
- [30] Kilgore M, Miller CA, Fass DM, Hennig KM, Haggarty SJ, Sweatt JD, et al. Inhibitors of class 1 histone deacetylases reverse contextual memory deficits in a mouse model of Alzheimer's disease. *Neuropsychopharmacology.* 2010;35(4):870–880.
- [31] Arguelles AO, Meruvu S, Bowman JD, Choudhury M. Are epigenetic drugs for diabetes and obesity at our door step? *Drug Discov Today.* 2016;21(3):499–509.
- [32] Whyne TF. Epigenetics in the development, modification, and prevention of cardiovascular disease. *Mol Biol Rep.* 2015;42(4):765–776.
- [33] Dueñas-González A, Naveja JJ, Medina-Franco JL. Introduction of Epigenetic Targets in Drug Discovery and Current Status of Epi-Drugs and Epi-Probes. In: *Epi-Informatics.* Elsevier; 2016. p. 1–20.

- [34] Jones PA, Issa JPJ, Baylin S. Targeting the cancer epigenome for therapy. *Nat Rev Genet.* 2016;17(10):630–641.
- [35] Naveja J, Dueñas-González A, Medina-Franco J. Chapter 12: Drug Repurposing for Epigenetic Targets Guided by Computational Methods. In: Medina-Franco J, editor. *Epi-informatics. Discovery and development of small molecule epigenetic drugs and probes.* 1st ed. Academic Press. Elsevier; 2016. p. 327–357.
- [36] Csoka AB, Szyf M. Epigenetic side-effects of common pharmaceuticals: a potential new field in medicine and pharmacology. *Med Hypotheses.* 2009;73(5):770–780.
- [37] Hunter P. The second coming of epigenetic drugs: a more strategic and broader research framework could boost the development of new drugs to modify epigenetic factors and gene expression. *EMBO Rep.* 2015;16(3):276–279.
- [38] Raynal NJM, Da Costa EM, Lee JT, Gharibyan V, Ahmed S, Zhang H, et al. Repositioning FDA-Approved Drugs in Combination with Epigenetic Drugs to Reprogram Colon Cancer Epigenome. *Mol Cancer Ther.* 2017;16(2):397–407.
- [39] Dekker FJ, van den Bosch T, Martin NI. Small molecule inhibitors of histone acetyltransferases and deacetylases are potential drugs for inflammatory diseases. *Drug Discov Today.* 2014;19(5):654–660.
- [40] Sato T, Cesaroni M, Chung W, Panjarian S, Tran A, Madzo J, et al. Transcriptional selectivity of epigenetic therapy in cancer. *Cancer Res.* 2017;77(2):470–481.
- [41] de Lera AR, Ganesan A. Epigenetic polypharmacology: from combination therapy to multitargeted drugs. *Clin Epigenetics.* 2016;8(1):105.
- [42] Fernández-de Gortari E, García-Jacas CR, Martínez-Mayorga K, Medina-Franco JL. Database fingerprint (DFP): an approach to represent molecular databases. *J Cheminform.* 2017;9(1):9.
- [43] Dobson CM. Chemical space and biology. *Nature.* 2004 dec;432(7019):824–828.
- [44] Oprea TI, Gottfries J. Chemography: the art of navigating in chemical space. *J Comb Chem.* 2001 apr;3(2):157–166.
- [45] Osolodkin DI, Radchenko EV, Orlov AA, Voronkov AE, Palyulin VA, Zefirov NS. Progress in visual representations of chemical space. *Expert Opin Drug Discov.* 2015 jun;10(9):959–973.
- [46] Peters JU. Polypharmacology – Foe or Friend? *J Med Chem.* 2013;56(22):8955–8971.
- [47] Poret A, Boissel JP. An in silico target identification using Boolean network attractors: Avoiding pathological phenotypes. *C R Biol.* 2014;337(12):661–678.