UNIVERSIDAD NACIONAL AUTÓNOMA DE MEXICO

PROGRAMA DE MAESTRÍA Y DOCTORADO EN CIENCIAS QUÍMICAS

ESTUDIO DE MODELADO MOLECULAR Y
QUIMIOINFORMÁTICA DE MODULADORES
DE BLANCOS TERAPÉUTICOS
EPIGENÉTICOS

PROYECTO DE INVESTIGACIÓN
PARA OPTAR POR EL GRADO DE

MAESTRO EN CIENCIAS

PRESENTA

Químico Farmacéutico Industrial FERNANDO DANIEL PRIETO MARTÍNEZ

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

Ciudad Universitaria, CD. MX. Diciembre 2016

El presente trabajo se realizó en el DIFACQUIM perteneciente al Departamento de Farmacia de la Facultad de Química, UNAM. El proyecto fue posible gracias a los siguientes apoyos:

- Beca CONACyT (No. de becario 576637/ CVU 660465)
- Programa de Apoyo a la Investigación y al Posgrado (PAIP 5000-9163), Facultad de Química, UNAM

Los productos académicos de este proyecto son:

- Cartel de avances, presentado en las Jornadas de Investigación de la Facultad de Química, UNAM
- Cartel de avances, presentado en el 1° Congreso estudiantil de Ciencia sin fronteras en la Escuela Superior de Medicina, IPN.
- Publicación en una revista de impacto (*RSC Advances*) bajo el título “A chemical space odyssey of inhibitors of histone deacetylases and bromodomains”. RSC Adv.,2016,6, 56225. DOI: 10.1039/c6ra07224k
- Presentación de resultados en la UGM de MOE, en Montreal Canadá

El jurado asignado a este trabajo está comprendido por:

Dr. Luis Demetrio Miranda Gutiérrez	Instituto de Química, UNAM
Dr. Fernando Cortés Guzmán	Instituto de Química, UNAM
Dr. José Correa Basurto	Escuela Superior de Medicina, IPN
Dr. Alfonso Sebastián Lira Rocha	Facultad de Química, UNAM
Dra. Laura Domínguez Dueñas	Facultad de Química, UNAM

AGRADECIMIENTOS

A la Universidad Nacional Autónoma de México por permitirme realizar mi posgrado, en esta máxima casa de estudios.

Al Programa de Maestría y Doctorado en Ciencias Químicas, por el apoyo otorgado para asistir al UGM de MOE en Montreal, donde se presentaron avances del presente proyecto.

Al Consejo Nacional de Ciencia y Tecnología, por la beca otorgada durante el desarrollo de mis estudios de maestría.

Al Instituto de Investigaciones Biomédicas, a través del Programa de Nuevas Alternativas para el Tratamiento de Enfermedades Infecciosas (NUATEI), por el apoyo para la obtención de licencias de software usado en el presente trabajo.

Al Dr. José Luis Medina Franco, por todo su apoyo, consejo y sobre todo por el empeño que infundió en mi durante la realización del proyecto.

A mis sinodales, por sus contribuciones, comentarios y sugerencias para el enriquecimiento y mejora del presente informe.

A mis padres Anita y Fernando, porque todo lo que soy hoy, es por su esfuerzo y dedicación.

A Tatiana, por todo su apoyo, confianza y comprensión durante ésta etapa y las que están por venir.

Al candidato a Dr. en C. Eli Fernández de Gortari, por su ayuda y consejos durante las primeras etapas del proyecto.

A mis compañeros de DIFACQUIM, por la convivencia y todos esos momentos e intercambios en el cubículo, nuestra segunda casa.

INDICE

Resumen……..2

Introducción……2

Antecedentes……..3

Hipótesis……..5

Objetivos……..5

Métodos……6

Resultados y discusión……..11

Conclusiones……23

Perspectivas……..23

Referencias……23

Anexo (Publicación original)……..26

INDICE DE TABLAS Y FIGURAS

Tabla 1. Número de ligandos co-cristalizados encontrados en PDB con actividad inhibitoria de BETs……..10

Figura 1. Quimiotipos con mayor frecuencia en las bases BRDi y HDACi……...16

Figura 2. Curvas de recobro de núcleos base, para las colecciones con información epigenéticamente relevante y la colección GRAS……...17

Figura 3. Comparativo de núcleos base asociados a la inhibición de BRDs y posibles quimiotipos similares por representación bidimensional……19

Figura 4. Ejemplos de alineamiento flexible……..20

Figura 5. Validación del acoplamiento molecular……...21

Figura 6. Estructuras químicas de los hits identificados por el protocolo de cribado virtual, como inhibidores potenciales de BRD4……...21

Figura 7. Histograma de interacciones representativas, recuperadas por PLIFs usando los diferentes algoritmos de acoplamiento……..22

Lista de Abreviaturas

ADN: Ácido desoxirribonucleico

BET: Proteína con bromodominio extraterminal

BRD: Bromodominio

DIFAC: Diseño de fármacos asistido por computadora

DNMT: DNA-5-metil-transferasa

ECCFGP: Huellas digitales moleculares por conectividad extendida

EQ: Espacio químico

ESQUER: Espacio químico epigenéticamente relevante

FGP: Huella digital molecular (*fingerprint*)

GRAS: Compuestos generalmente reconocidos como seguros

HDACs: Histona-desacetilasas

MACCS: *Molecular Access System*

MOE: *Molecular Operating Enviroment*

PCA: Análisis de componentes principales (por sus siglas en inglés)

RESUMEN[1]

La epigenética comprende una serie de mecanismos independientes de ADN que involucran a una gran variedad de enzimas. Estas enzimas, por su naturaleza "editora", son directamente responsables del ciclo expresión/supresión de los genes. Diversos trabajos han reportado hasta ahora a una gran cantidad de inhibidores de éstas dianas. No obstante, hasta ahora el perfil de dichos compuestos no se ha discutido en un contexto quimioinformático. En esta investigación se presenta una muestra del espacio químico epigenéticamente relevante (ESQUER) y su posible relación con fármacos aprobados por la FDA y aditivos alimentarios. Mediante técnicas de análisis por componentes principales se obtuvo una representación visual aproximada de dicho espacio que también fue analizado utilizando métodos quimionformáticos. Se encontraron relaciones fisicoquímicas y estructurales importantes entre inhibidores epigenéticos y fármacos aprobados/GRAS. Los resultados pemitieron sugerir estudios posteriores para el reposicionamiento o caracterización nutracéutica de dichos compuestos.

Una vez caracterizado el espacio químico de inhibidores epigenéticos conocidos se desarrolló un protocolo de cribado virtual para sugerir nuevos inhibidores potenciales. El cribado virtual se enfocó únicamente a bromodominios, específicamente a la isoforma BRD4. Se utilizaron búsquedas por similitud molecular y alineamientos flexibles para identificar inhibidores potenciales. Para determinar las posibles interacciones ligando-proteína a nivel molecular de los compuestos identificadas en el paso anterior, se empleó el acoplamiento molecular automatizado usando cuatro programas: MOE, Autodock 4, Autodock Vina e ICM. A pesar de las diferencias notables en evaluación y algoritmos de cada programa se identificaron "*hits* consenso". Los resultados obtenidos del cribado virtual son prometedores porque las estructuras de los compuestos identificados difieren significativamente de los inhibidores actuales. Como principales perspectivas de este trabajo de investigación destacan: identificar un modelo farmacofórico para inhibidores de bromodominios y la evaluación biológica de los compuestos propuestos.

1. Introducción

La epigenética involucra cambios químico-estructurales reversibles; catalizados por enzimas, que son previos a la edición del ADN [1]. Las histonas son octámeros proteicos que empaquetan el ADN, de manera que su almacenamiento y transporte no interfiera con los procesos celulares. Las cadenas laterales son los puntos de modificación por una serie de enzimas que adicionan grupos funcionales como metilo, acetato o fosfato sobre ciertos residuos de aminoácido, p.ej. lisina. Estas modificaciones actúan como una etiqueta de instrucciones que determina el destino de una histona incluyendo su carga genética. Aspectos importantes de estas modificaciones es que son heredables y que responden a estímulos del ambiente. A mediados del siglo pasado se observó cierta correlación entre los patrones de modificación y la expresión de oncogenes. Ahora se sabe que esta relación se extiende a otros padecimientos crónico-degenerativos como esquizofrenia, enfermedad de Alzheimer, síndrome metabólico, diabetes mellitus, padecimientos cardiovasculares, etc.[2,3]

1. Por cuestiones de espacio, se han omitido figuras y tablas. Por favor refiérase al artículo anexo a este escrito, las referencias se harán mediante A-X, siendo X el número correspondiente a la figura del artículo original y SX referencias al material suplementario de dicho artículo.

El presente trabajo comprendió el análisis quimioinformático de compuestos reportados en bases de datos públicas como inhibidores de bromodominios (BRDs), ADN-5-metiltransferasas (DNMT) e histonas desacetilasas (HDACs). Dicho análisis involucró la proyección del espacio químico y métodos por similitud molecular. El comparativo se realizó contra las siguientes colecciones de compuestos como referencia: fármacos en fase clínica (*Clinic*), inhibidores enzimáticos diversos (General), fármacos aprobados (FDA) y aditivos alimentarios (GRAS). Después de la caracterización del ESQUER se realizó una búsqueda de moléculas de similitud estructural, considerable a la de moléculas activas, con el fin de proponer nuevos inhibidores ya sea por reposicionamiento de fármacos, valor nutracéutico o diseño de moléculas sonda para bromodominios. Con tal fin se desarrolló un protocolo de cribado virtual, partiendo de similitud molecular, para el posterior refinamiento por alineamiento en tres dimensiones y acoplamiento molecular.

2. Antecedentes

2.1. Epigenética

Conrad Waddington describió por primera vez el concepto de epigenética en 1940. Él estudiaba la derivación de genotipo a fenotipo, de ahí los mecanismos “superiores” al ADN [4]. En los eucariotas el material genético se agrupa como 146 pares de bases en un cromosoma; el cual, a su vez, es rodeado por el nucleosoma que forma parte del octámero de histonas [5]. Las cadenas laterales de la histona son reconocidas por las enzimas epigéneticas de las cuales se pueden distinguir tres grupos: escribas, supresoras y lectoras.

Los escribas adicionan grupos químicos que son reconocidos por otras enzimas. La modificación se da sobre residuos de aminoácido (lisina o arginina) [6]. Las supresoras son antagonistas a los escribas ya que identifican a los grupos adicionados o presentes en la histona y los eliminan. Este proceso de edición forma patrones particulares que regulan el ciclo “expresión-supresión”. Las lectoras “interpretan” los patrones de modificación en las “colas” de la histona. Este paso es crucial pues el dictamen emitido por las lectoras determinará el destino del ADN contenido por la histona.

Aunado a lo anterior, se ha investigado el rol de la salud materna durante la gestación y la epigenética asociada a este proceso. Por ejemplo, se ha visto que la nutrición y exposición a radiación UV son patrones determinantes en la salud y memoria epigenética [7].

2.2 Enzimas epigenéticas más relevantes

El interés por la investigación epigenética ha crecido exponencialmente. Una de las primeras evidencias que relacionó enzimas epigenéticas y neoplasia fue la conservación de patrones hipermetilados de ADN en líneas celulares de cáncer. En la misma forma se observó que una hiperexpresión de histona-acetil transferasas conduce a la mutación de genes y cambio en la expresión de oncogenes. Ante estos hallazgos la investigación se enfocó en las enzimas DNMT e histona acetiltransferasas (HAT). El desarrollo de nuevo conocimiento permitió elucidar el papel de otras enzimas como HDACs y BRDs, cruciales en el proceso de

1. Por cuestiones de espacio, se han omitido figuras y tablas. Por favor refiérase al artículo anexo a este escrito, las referencias se harán mediante A-X, siendo X el número correspondiente a la figura del artículo original y SX referencias al material suplementario de dicho artículo.

modulación epigenética. A continuación, se describen los aspectos más relevantes sobre estas familias de enzimas.

2.2.1. ADN-5-metil-transferasas (DNMTs)

El ADN está constituido por pares de bases púricas y pirimídicas que interactúan por medio de puentes de hidrógeno. Debido a la naturaleza química de esta interacción, el par base-base es característico. El par citosina-guanina que forman islas CpG inducen una modificación epigenética [8]. Específicamente, la metilación, es considerada como una de las modificaciones más estables [9], ocurre en la citosina (sobre el átomo C5). La ADN-metil-transferasa (DNMT) es la enzima encargada de esta función. Actualmente se distinguen dos familias: DNMT1 encargada principalmente de metilación de "mantenimiento" (asegurando la subsistencia del patrón de metilación) y DNMT3(A y B) encargadas de la metilación *de novo* [10].

2.2.2. Histona-descetilasas (HDACs)

Mientras que las histonas poseen en general una carga parcial positiva, el ADN posee carga negativa. Así, la interacción electrostática entre ambas permite la compactación del nucleosoma. La adición de un grupo electronegativo neutraliza esta interacción de cargas, por ello el ADN se encontrará más disponible para su transcripción. Se infiere entonces, que la remoción de este grupo se relaciona con la supresión génica.

Las enzimas encargadas de ésta labor remueven grupos acilo de lisinas. Esta familia de enzimas se pueden clasificar en cinco grupos (I, IIa, IIb, III y IV) cuyo mecanismo de acción se puede dividir a su vez, en dos variantes: catálisis por iones zinc 2+ y catálisis NAD dependiente [5]. De las casi 18 isoformas que componen esta familia, más del 50% está relacionada con alguna forma de cáncer. Entre ellas destacan cáncer prostático, gástrico-entérico, mama, pulmón y colon [11]. A pesar de lo anterior, actualmente sólo existen dos inhibidores aprobados para uso clínico en el tratamiento de linfoma por células T [12].

Entre los inhibidores de HDACs descritos destacan los derivados de ácido hidroxámico. Tras diferentes estudios de estructura actividad cualitativa y cuantitativa (SAR y QSAR, respectivamente y por sus siglas en inglés) se sabe que son esenciales tres características para explicar la inhibición: un grupo electronegativo que se coordina con el zinc, una cadena "puente" de entre 6 y 8 átomos de carbono que une a un grupo que favorezca la inhibición por impedimento estérico [13].

2.2.3. Bromodominios (BRDs)

Los bromodominios son enzimas lectoras. Su función característica es el reconocimiento de grupos acetilo; específicamente, aquellos ubicados sobre residuos de lisina [14]. Si bien su función puede parecer sencilla, al día de hoy se han caracterizado cerca de 30 bromodominios. Éstos a su vez pueden denominarse por su

1. Por cuestiones de espacio, se han omitido figuras y tablas. Por favor refiérase al artículo anexo a este escrito, las referencias se harán mediante A-X, siendo X el número correspondiente a la figura del artículo original y SX referencias al material suplementario de dicho artículo.

ubicación o proteína asociada. Este estudio se enfocó a la familia BET (por sus siglas en inglés) que se refiere a los bromodominios de dominio extraterminal [2]. El interés en estas dianas se debe al reciente desarrollo farmacológico y conocimiento parcial de su estructura [15], así como su relación a padecimientos crónico-degenerativos [14].

El desarrollo de la investigación sobre estas dianas comenzó de forma fortuita en 2012, Filippakopoulos y colaboradores [16] analizaron de manera sistemática la interacción de los BETs y núcleos benzodiazepínicos. Gracias a estas investigaciones fue posible una caracterización completa de estos bromodominios con el desarrollo de potentes inhibidores como JQ1 e IBET (GSK525762A). En particular en la literatura la modulación de los BETs presenta la siguiente prioridad BRD4>BRD2>BRD3. Recientemente ha surgido el interés por otras familias de bromodominios como CREBP y BAZ2(A y B) de las cuales aún no se conoce claramente su función [17]. Es por esto que es importante identificar moléculas con alta afinidad y pocos efectos adversos que permitan un estudio similar al descrito para la familia BET [14].

De manera análoga a las enzimas epigenéticas discutidas anteriormente los bromodominios se asociaron con cáncer, particularmente con cáncer prostático [18]. Se sabe que los distintos bromodominios son cruciales en procesos celulares diversos. Además se han asociado con problemas de inflamación y diabetes [2].

3. Hipótesis

Si estudios anteriores identificaron fármacos aprobados como BRDi, entonces, existe una región de intersección en el espacio químico que puede extenderse a HDACi. Como las propiedades fisicoquímicas (PFQ) están ligadas a la estructura, dicha intersección permitirá identificar estructuras bioisosteras a nivel de estas dianas epigenéticas. Finalmente, ante las nuevas investigaciones sobre el efecto de la dieta y alimentación en el epigenoma, existe una relación entre aditivos alimentarios e inhibidores epigenéticos. En resumen; la presente quimiografía, pondrá en contexto al ESQUER ofreciendo una herramienta para el diseño de fármacos, polifármacos y nutracéuticos.

4. Objetivos

General

Identificar las posibles relaciones estructurales entre los inhibidores a nivel epigenético, con la posibilidad de describir de manera empírica la naturaleza farmacofórica epigenética relevante.

Particulares

1. Determinar la complejidad asociada a los inhibidores actuales con el fin de mejorar su selectividad.
2. Identificar estructuras afines a BRD e HDACs, a fin de proponer inhibidores duales.
3. Proponer inhibidores potenciales de la isoforma BRD4 con núcleos base distintos a los actuales.

1. Por cuestiones de espacio, se han omitido figuras y tablas. Por favor refiérase al artículo anexo a este escrito, las referencias se harán mediante A-X, siendo X el número correspondiente a la figura del artículo original y SX referencias al material suplementario de dicho artículo.

5. Métodos

Las colecciones de compuestos epigenéticos y de referencia se obtuvieron de bases de datos moleculares públicas. Las bases de datos curadas se analizaron y compararon usando representaciones moleculares complementarias: seis propiedades fisicoquímicas, tres huellas digitales moleculares y núcleos base. Los detalles del análisis cuantitativo y descriptivo se presentan en las siguientes secciones.

5.1. Bases de datos y curado de información

Las estructuras químicas y actividad biológica de inhibidores de HDACs (1, 2, 3, 8 y SIRT1) y BRDs (2, 3, 4) se descargaron de ChEMBL (www.ebi.ac.uk/chembl/) y Binding Database (BDB, www.bindingdb.org/). Los inhibidores analizados en esta investigación (BRDi y HDACi) se han asociado a distintas variedades de cáncer p.ej. colon, próstata, gástrico y cérvico-uterino [11].

Los inhibidores de BRDs se han asociado con algunas líneas celulares de cáncer [19]. Las estructuras recuperadas fueron obtenidas en formato molecular simplificado a una línea (SMILES, por sus siglas en inglés) y convertidas a estructuras tridimensionales usando el programa MOE. El curado de datos, paso fundamental en cualquier estudio quimionformático [20], se realizó mediante la función “Wash” implementada en MOE. Usando esta función las estructuras se prepararon mediante la desconexión de iones metálicos y neutralización de estados protonados. Posteriormente se realizó una optimización de la geometría con el fin de trabajar con el confórmero más estable mediante un algoritmo estocástico. La estereoquímica indicada en la cadena SMILES original fue conservada durante el proceso. Finalmente, las bases obtenidas se inspeccionaron en forma visual.

5.1.2. Bases de datos de referencia

Las bases en estudio se compararon con una base de inhibidores de DNMT1 analizada previamente utilizando un protocolo similar [21]. Otras colecciones de compuestos de referencia fueron: aditivos y saborizantes de la lista GRAS de FEMA (GRAS 25) previamente curada (‘GRAS’), fármacos aprobados por la FDA recuperados de DrugBank (‘FDA’), compuestos en fase clínica reportados en el ‘Therapeutic Target Database’ (TTD), (‘Clinic’); una colección comercial de inhibidores epigenéticos diversos, distribuida por Selleckchem (‘Epi-bloq’); y una colección de inhibidores enzimáticos en general, igualmente distribuida por Selleckchem (‘General’). Estas bases de referencia (con excepción de GRAS) fueron utilizadas previamente por Fernández-de Gortari y colaboradores [21].

El razonamiento tras el uso de estas colecciones es el siguiente: el posible reposicionamiento o uso alternativo de fármacos (FDA y *Clinic*), identificación de posibles moléculas sonda (General), validación de diversidad y características “epigenéticas” (Epi-bloq.) y la relación nutracéutica (GRAS).

1. Por cuestiones de espacio, se han omitido figuras y tablas. Por favor refiérase al artículo anexo a este escrito, las referencias se harán mediante A-X, siendo X el número correspondiente a la figura del artículo original y SX referencias al material suplementario de dicho artículo.

5.2. Representación molecular

Dado que el espacio químico (EQ) depende fuertemente de la representación de estructuras, se utilizaron diversos criterios (utilizados previamente por este y otros grupos de investigación [21–23]). Entre estos criterios se emplearon propiedades fisicoquímicas relevantes desde el punto de vista farmacéutico, huellas digitales moleculares y núcleos estructurales base. Además, se analizó la complejidad de las estructuras químicas usando métricas bien definidas y establecidas [24].

5.2.1. Propiedades fisicoquímicas

Se calcularon en MOE seis propiedades: coeficiente de partición octanol/agua (SLogP), número de enlaces rotables (RTB), aceptores de puente de hidrógeno (HBA), donadores de puente de hidrógeno (HBD), área topológica superficial (TPSA) y peso molecular (MW). Estas propiedades se utilizan con frecuencia en la búsqueda de nuevos fármacos y son la base de reglas empíricas de "carácter de fármaco" (*drug-likeness*) [21]. La distribución de las propiedades fue visualizada en gráficos "boxplot" mediante R Studio y se analizaron mediante el paquete PMCMR para estadística descriptiva. La visualización se realizó mediante un cálculo de componentes principales para expresar el EQ en función de los *eigenvectores* del espacio original. El análisis de componentes principales (PCA) se realizó con MOE y la gráfica fue generada en Datawarrior [25]. Para la comparación estadística se realizó una prueba de homosedacidad para las distribuciones. Esto se hizo mediante una prueba de Shapiro-Wilks seguida de una prueba ANOVA de Kruskal-Nemenyi en R Studio. El criterio de aceptación para la hipótesis nula fue un valor $p < 0.05$, estableciendo que no existe diferencia estadísticamente significativa entre los grupos de datos comparados.

5.2.2. Huellas digitales moleculares

La similitud intra-colección fue evaluada para BRDi y HDACi, así como de las colecciones de referencia para una comparación adecuada de la diversidad de las colecciones. Para ello se calcularon tres huellas digitales con MOE: MACCS keys, Gráfico farmacofórico triangular (GpiDAPH3) y Distancia Tipo (TGD). MACCS está basado en diccionario, esto quiere decir que compara la estructura contra una "lista de fragmentos predefinida". GpiDAPH3 identifica tres puntos farmacofóricos y asocia un triángulo que sirve como punto de comparación. TGD también se basa en puntos farmacofóricos, pero la comparación es la distancia comprendida entre dos puntos dentro de la molécula.

La similitud molecular entre huellas moleculares se calculó con el coeficiente de Tanimoto:

$$T(a,b) = \frac{c}{a + b - c}$$

Donde *a* y *b* indican el número de fragmentos del i-esimio y j-esimo compuestos y *c* representa el número de fragmentos compartidos entre i y j.

1. Por cuestiones de espacio, se han omitido figuras y tablas. Por favor refiérase al artículo anexo a este escrito, las referencias se harán mediante A-X, siendo X el número correspondiente a la figura del artículo original y SX referencias al material suplementario de dicho artículo.

Posteriormente, se construyó la matriz de similitud (NxN) comparando todos los pares de compuestos (N) y se extrajeron los N(N-1)/2 elementos de la diagonal. Con esto se construyó un gráfico de distribución acumulada contra similitud el cual se generó mediante el módulo matplotlib en Python [26].

5.2.3. Núcleos base

Los núcleos base o sistemas cíclicos son subestructuras que se generaron al eliminar los vértices o cadenas laterales. Esto se hizo con el programa Índices Moleculares Equivalentes (*Molecular Equivalence Indices*, MEQI). MEQI se ha utilizado intensivamente en el análisis de bases de datos muy diversas y grandes [27–29].
Esto núcleos base constituyen parte del quimiotipo, metodología desarrollada y discutida por Johnson y Xu. Para cada quimiotipo único se generó un código alfanumérico de cinco caracteres. Los sistemas cíclicos resultantes representan clases; es decir, no es posible que una molécula pertenezca a más de un quimiotipo. En este trabajo se analizaron los núcleos base de BRDi y HDACi, así como una comparación directa para los núcleos más frecuentes.

5.2.3.1. Diversidad de los núcleos base

Se evaluó mediante las curvas de recobro de quimiotipos (CRQ) [24]. El objetivo de estos gráficos fue comparar la fracción de núcleos base que contiene una fracción particular de compuestos de la base de datos. Para generar dicho gráfico la lista de quimiotipos se ordena por frecuencias. La curva fue caracterizada por el área bajo la curva (ABC) y por el valor de la fracción de núcleos base que comprende el 50% de los compuestos de la base estudiada (F_{50}). Colecciones de compuestos con alta diversidad de núcleos base tiene curvas que se acercan a la diagonal. La máxima diversidad se tiene cuando cada compuesto de la base de datos tiene un núcleo base diferente. En este caso la curva CRQ es una diagonal con ABC = 0.5.

5.2.3.2. Enriquecimiento de quimiotipos

Dado que es relativamente común que un quimiotipo frecuente sea común a compuestos activos e inactivos se realizó un gráfico de enriquecimiento. Primero se calculó la fracción de compuestos activos en toda la base de datos, Act(C), con la ecuación:

$$Act(C) = [C^*] / [C]$$

Donde [C] es el número total de compuestos, [C*] es el total de compuestos considerados como activos. Para el presente trabajo se definió como activo aquel compuesto con una CI_{50} menor a 1µM.

La fracción de compuestos activos contenidos en el quimiotipo λ, Act(C_λ), se obtuvo con la expresión:

$$Act(C_\lambda) = [C_\lambda^*] / [C_\lambda]$$

donde [C_λ] y [C_λ*] son el número total de compuestos y activos, respectivamente, para el quimiotipo λ.

El factor de enriquecimiento (EF) para un quimiotipo λ se calculó con la expresión:

1. Por cuestiones de espacio, se han omitido figuras y tablas. Por favor refiérase al artículo anexo a este escrito, las referencias se harán mediante A-X, siendo X el número correspondiente a la figura del artículo original y SX referencias al material suplementario de dicho artículo.

$$EF(C_\lambda) = Act(C_\lambda) / Act(C)$$

EF (C_λ) representa la proporción de compuestos activos relativos a un quimiotipo respecto al total de activos en la base de datos. De tal manera que cuanto mayor sea EF, el núcleo base correspondiente resulta más atractivo para estudio. Esto ayuda a hacer una mejor distinción de quimiotipos frecuentes pues suele ocurrir que núcleos frecuentes tienen poco enriquecimiento.

5.2.4. Complejidad molecular

La complejidad molecular suele ser la puerta a nuevas regiones del EQ. Previamente señalado por Lovering y colaboradores, se ha visto que los fármacos con mayor éxito son en general más estructuralmente complejos [30]. Adicionalmente, se ha observado que un incremento en la complejidad va acompañado por mayor selectividad, hablando específicamente del número de carbonos saturados. Otras métricas propuestas para evaluar la complejidad es el peso molecular [24]. Para este análisis se retomaron dos métricas bien definidas y reportadas, buscando caracterizar de manera apropiada la posible complejidad de las bases en estudio. El grado de saturación se define por la fracción de carbonos sp^3 (Fsp^3), que se obtiene del cociente (# de carbonos con hibridación sp^3/ # total de carbonos en la molécula). Por lo que a mayor valor de Fsp^3 se espera que una estructura sea “menos plana”. La complejidad estereoquímica se definió a partir de la proporción de carbonos quirales F-quiral = (# de átomos de carbono quirales/ # total de átomos de carbono). Las fracciones fueron calculadas en MOE y Maya Chem Tools (www.mayachemtools.org). La distribución de estos descriptores se analizó con *box plots* y estadística descriptiva obtenida con R Studio y el paquete PMCMR.

5.3. Cribado virtual

La búsqueda de inhibidores potenciales de BRD4 se realizó por similitud molecular usando *fingerprints* (representación en dos dimensiones -2D) y alineamientos flexibles (representación tridimiensional -3D).

5.3.1. Similitud en 2D

Se construyeron matrices de similitud para las bases de datos antes mencionadas, esto se hizo mediante los *scripts* de Mayachem Tools. Las representaciones utilizadas fueron *MACCS keys* (FGP por diccionario de fragmentos) y ECFP4 (FGP topológico). Las comparaciones se calcularon con el coeficiente de Tanimoto, usando como límite, similitudes mayores a 0.6 y 0.2, respectivamente. Los compuestos recuperados se sometieron a una prueba de similitud en 3D.

5.3.2. Alineamiento flexible (similitud tridimensional)

Para realizar un filtrado de compuestos “candidatos” se buscó que su estructura permitiera adoptar una conformación similar a inhibidores de BRD4 conocidos. Para ello se obtuvieron de *Protein Data Bank (*http://www.rcsb.org*)* los PDB asociados a BRDs con diversos inhibidores co-cristalizados (Tabla 1). Se extrajeron los ligandos del PBD y se utilizó el módulo FlexAlign en MOE, con el campo de fuerza MMFF94x.

1. Por cuestiones de espacio, se han omitido figuras y tablas. Por favor refiérase al artículo anexo a este escrito, las referencias se harán mediante A-X, siendo X el número correspondiente a la figura del artículo original y SX referencias al material suplementario de dicho artículo.

Dejando fijas las coordenadas del confórmero co-cristalizado se realizó una superposición buscando un alineamiento de grupos farmacofóricos representativos. El algoritmo iterativo busca el menor RMSD posible entre pares de átomos y asigna un puntaje total (S): cuanto menor sea el número expresado el alineamiento es mejor. MOE utiliza una aproximación por densidades de probabilidad: se calculan dos distribuciones Gausianas y se estima el traslape entre ambas densidades de probabilidad. Para refinar el cálculo se pueden considerar las propiedades ácido-base, aromaticidad, hidrofobicidad, carga y volumen [31].

5.3.3. Acoplamiento molecular (*docking*)

Como etapa final de la evaluación *in silico,* se implementaron protocolos de acoplamiento usando a BRD4 como proteína de referencia (PBD ID: 3P5O, usada previamente en otros estudios [14,18,32]). La estructura se inspeccionó en el editor de secuencias de MOE para verificar errores de secuencia y/o átomos.

Tabla 1. Número de ligandos co-cristalizados encontrados en PDB con actividad inhibitoria de BETs

PDB ID	BRD
4J1P	2
5BT5	
3S92	3
3P5O	4
4CFL	
4YH4	

Posteriormente, se preparó la macromolécula mediante el módulo LigX el cual realiza una minimización energética, relaja la cadena de la proteína, elimina moléculas de disolventes comunes y solo conserva aquellas moléculas de agua a 4 Å del sitio activo. En el caso de Autodock 4 y Autodock Vina la estructura se preparó con un paso adicional, usando PyRx v.0.8 para cribado virtual [33].

Se utilizaron los algoritmos implementados por MOE, ICM (Molsoft), Autodock 4 y Autodock Vina. Para cada programa se realizó la validación correspondiente. Dicho proceso consistió en simular la unión al ligando co-cristalizado, probando si el algoritmo puede reproducir el confórmero activo o uno muy cercano en RMSD. Una vez validado el protocolo, se procedió al acoplamiento de las moléculas identificadas en los pasos anteriores.

Todos los programas se usaron con el módulo para cribado virtual para una búsqueda poco exhaustiva, preferentemente para evaluar los compuestos en condiciones similares. Una vez terminadas las corridas se eligieron los compuestos de acuerdo a la función de evaluación en cada programa: como criterio heurístico el puntaje debía ser cercano al del ligando co-cristalizado. Así, aquellos recuperados en consenso por cuando menos tres programas se consideraron “hits”.

5.3.4. Análisis de los confórmeros

A fin de realizar una evaluación sistemática de los modos de unión, los confórmeros acoplados se almacenaron en una base de datos. Posteriormente, usando MOE se calculó un FGP de interacciones

1. Por cuestiones de espacio, se han omitido figuras y tablas. Por favor refiérase al artículo anexo a este escrito, las referencias se harán mediante A-X, siendo X el número correspondiente a la figura del artículo original y SX referencias al material suplementario de dicho artículo.

farmacofóricas proteína-ligando (PLIF, por sus siglas en inglés). El PLIF resume todas las interacciones presentes para los diferentes ligandos contra una proteína (3P5O). Para reconocer inhibidores potenciales se buscaron las interacciones más frecuentes con BRD4 [14].

6. Resultados y discusión

Se analizaron en total 2000 y 207 compuestos únicos (eliminando duplicados), conocidos como inhibidores de HDACs y BRDs, respectivamente. Es importante aclarar que los compuestos analizados no son selectivos a isoformas específicas de HDACs o BRDs. Esto se hizo con el fin de elucidar de manera general la distribución y cobertura del espacio químico.

Se estableció un límite heurístico para clasificar la actividad biológica (CI_{50}): En el proyecto se definió a un compuesto como "activo" si su CI_{50} era menor o igual a 1 μM. Este criterio recuperó un alto porcentaje de activos contra HDACs (79.5%), mientras que en BRDs solo el 46.85% de los compuestos puede considerarse como activo. Estos valores corroboran un desarrollo y optimización más intensivos en HDACi.

6.1. Propiedades fisicoquímicas

La Figura A-2 (ver anexo) muestra a los *boxplots* con la distribución de propiedades. La estadística descriptiva y otros análisis estadísticos pueden encontrarse en el Anexo. Con respecto a los átomos donadores de puente de hidrógeno tanto en BRDi como en HDACi se mostraron distribuciones similares, que además son comparables a la colección de inhibidores ('General') y Epi-bloqueadores de origen comercial ('Epi-bloq') (ya que la mediana es 4). En general, DNMT mostró un mayor número de HBA, mientras que los fármacos aprobados y en fase clínica mostraron valores bajos de HBA. Para los valores de HBD se observó un comportamiento similar: HDACi y DNMT tienen valores más altos (valor de 3 para la mediana). Comparados nuevamente con 'General' y 'Epi-bloq' (valor de la mediana en 2); mientras que BRDi, FDA y Clinic se ubican más abajo con un valor de 1. Dicha tendencia se observó, para los valores de TPSA, medida asociada con la polaridad de la molécula.

En cuanto al SLogP tomado como medida de hidrofobicidad, BRDi tuvo los valores más altos en este parámetro. Para HDACi los valores son comparables con 'General' y pueden considerarse como los segundos más altos. Para evaluar la flexibilidad molecular se utilizó RTB. Como se puede notar la distribución de esta propiedad fue muy semejante entre colecciones, en este caso HDACi muestra los valores más altos. Esto puede atribuirse a una mayor presencia de compuestos peptídicos reportados como HDACi. Para el peso molecular se observó un comportamiento muy similar (con la excepción de GRAS), confirmando que los grupos de datos son comparables en este sentido.

Una prueba estadística de Kruskal-Nemenyi (ver Anexo) reveló que, en general, BRDi es semejante a 'General' como 'Epi-bloq'. Esto sugiere que los compuestos probados contra BRD probablemente proceden de bibliotecas comerciales. HDACi mostró un comportamiento similar, teniendo un parecido significativo con

1. Por cuestiones de espacio, se han omitido figuras y tablas. Por favor refiérase al artículo anexo a este escrito, las referencias se harán mediante A-X, siendo X el número correspondiente a la figura del artículo original y SX referencias al material suplementario de dicho artículo.

'General' y 'Epi-bloq', además de contar con propiedades semejantes a las de DNMTi (particularmente en términos de HBD, TPSA y MW).

Mediante un análisis estadístico se buscó demostrar si existe diferencia entre compuestos 'activos', (definidos en este proyecto si CI_{50}< 1 μM) tanto para HDACi como BRDi. Los resultados mostraron que, en cuanto a PFQ, no existe una diferencia estadísticamente significativa entre compuestos activos o inactivos.

6.1.1. Representación visual del espacio de propiedades

La Figura A-3 muestra representaciones en 2D y 3D del ESQUER, compuesto por BRDi, HDACi y DNMTi. La posición relativa se usó para comparar estos grupos de datos con las colecciones de referencia. Como se mencionó en la sección de Métodos, dicho espacio está en función de seis PFQ. Para claridad visual refiérase al Anexo (Figura S1), donde se muestran por separado BRDi, HDACi y GRAS en las mismas coordenadas de la Figura A-3.

Además, BRDi y HDACi cubren regiones similares a DNMTi, FDA y otras colecciones de referencia. En particular, BRDi cubre un espacio menor seguida por 'Epi-bloq' (Figura S1) ambas son cercanas en el ESQUER. Esto puede relacionarse tanto al número de compuestos (207 y 113, respectivamente) como al tipo de estructuras en ambas bases. Las pruebas de Nemenyi (Tabla S1) corroboraron una similitud estadísticamente significativa entre BRDi y 'Epi-bloq'. Compuestos evaluados contra HDACs tuvieron una distribución espacial más amplia. Los puntos fuera de escala pueden relacionarse con péptidos, ya que en general estos son más flexibles y polares que las moléculas pequeñas usadas en el diseño y desarrollo de fármacos. Aunado a lo anterior, estos péptidos según su naturaleza, pueden o no ser de alto MW.

Los primeros dos componentes principales (CP) recuperaron 82.2% de la varianza, mientras que los primeros tres recobraron 90.3%. HBA y HBD tienen la mayor contribución a CP1 mientras que SLogP tiene la mayor contribución a CP2. Nótese que estas propiedades están asociadas a la polaridad.

El análisis de GRAS mostró un comportamiento notable. Esto se debe a que GRAS comparte regiones del ESQUER, aunque con mayor dispersión (Figuras A-3 y S1). Muchos de los puntos fuera de escala son cercanos a HDACi; esto puede atribuirse a que ciertos aditivos alimenticios derivan de azúcares o péptidos.

Otro hallazgo interesante fue que GRAS tuvo una distribución similar de SLogP respecto a 'Epi-bloq'. En otros trabajos se ha discutido la importancia del log P como propiedad determinante en compuestos bioactivos. Esto puede relacionarse a la posible asociación entre los químicos alimenticios y epigenética.

Por otro lado, si bien las PFQ y la representación del ESQUER son un punto de partida crucial, estas herramientas no permiten evaluar o discernir la naturaleza química de los compuestos. Los aspectos estructurales de los epi-compuestos se discuten en la siguiente sección.

1. Por cuestiones de espacio, se han omitido figuras y tablas. Por favor refiérase al artículo anexo a este escrito, las referencias se harán mediante A-X, siendo X el número correspondiente a la figura del artículo original y SX referencias al material suplementario de dicho artículo.

6.2. Huellas moleculares estructurales

Las huellas moleculares (*fingerprints*) permiten una comparación de conectividades interatómicas de estructuras químicas. Como se describió en la sección anterior se usaron tres representaciones diferentes: MACCS keys, GpiDAPH3 y TGD. Estos descriptores se han usado previamente en trabajos de caracterización quimioinformática, así como la diversidad molecular para el caso de DNMT1 [21]. De esta manera fue posible hacer una evaluación cuantitativa de la similitud inter- e intra-colecciones.

6.2.1. Similitud intra-bibliotecas

El valor de similitud proviene de la comparación entre pares, mediante el coeficiente de Tanimoto a partir de las huellas digitales moleculares (*fingerprints,* FGP) mencionados. La Figura A-4 presenta la distribución de los valores de similitud usando funciones de distribución acumulada (CDFs). La estadística descriptiva de estas curvas puede encontrarse en el Anexo (Tabla S5). En general, los valores de similitud obtenidos por GpiDAPH3 y TGD para las bases estudiadas, difirieron notablemente de los valores en MACCS. De tal suerte que el orden decreciente de similitud fue: TGD> MACCS > GpiDAPH3 (Tabla S5). El orden relativo de los valores de similitud fue consistente por lo encontrado por otros grupos de investigación [28] [29]. De acuerdo a MACCS keys, BRDi posee una similitud intra-colección bastante alta seguida de HDACi, p.ej. los valores de la mediana son 0.463 y 0.423 respectivamente. Esto indica que ambas colecciones poseen una diversidad estructural baja respecto a otras, colecciones. Las colecciones más diversas fueron GRAS y FDA (con valores de 0.261 y 0.308, respectivamente). La diversidad estructural para fármacos aprobados se ha comentado previamente por López-Vallejo et al. [34] y es de esperarse dada la gran cantidad de dianas y mecanismos farmacológicos. Por esta razón es interesante que '*Clinic'* sea menos diverso que FDA, pero igualmente diverso que 'General'. Esta tendencia se mantuvo para TGD y GpiDAPH3, es interesante que para este último FGP BRDi y FDA son de similitud cercana.

Como se puede observar, la forma de representar un compuesto repercute mucho en la evaluación de la diversidad estructural y biológica; p.ej., considere HDACi, para GpiDAPH3 es una colección poco diversa mientras que por MACCS es una de los más diversas con una similitud comparable a '*Clinic'*.

Nótese que no existe una relación general entre el tamaño de las bases de datos y la diversidad estructural. A partir de esto podría anticiparse que cuanto más pequeño es un grupo, mayor será la similitud intra-colección. Bajo este razonamiento puede asociarse la baja diversidad en BRDi al número de compuestos (207). Sin embargo, se observa que 'Epi-bloq', es más diverso y cuenta con menos compuestos (113). Esto puede deberse a que 'Epi-bloq' a pesar de ser una colección enfocada, recopila inhibidores de dianas muy diversas. Bajo esta premisa, considere la baja diversidad en HDACi contrastada contra FDA, 'Clinic' y 'Epi-bloq' siendo que HDACi es la colección más grande de ellas (2000, 1490, 837 y 113, respectivamente).

En resumen, se concluyó que BRDi es una colección "limitada" teniendo en cuenta una representación por huellas digitales. Esto se ve reflejado en la cantidad de derivados específicos a ciertos núcleos base, como benzodiazepinas [17]. Por ello, se enfatiza la necesidad de buscar nuevos inhibidores para estas dianas,

1. Por cuestiones de espacio, se han omitido figuras y tablas. Por favor refiérase al artículo anexo a este escrito, las referencias se harán mediante A-X, siendo X el número correspondiente a la figura del artículo original y SX referencias al material suplementario de dicho artículo.

especialmente al ver que sus inhibidores conocidos cubren un área muy limitada del espacio químico. Por otro lado, HDACi tiene mayor diversidad química. Esto es de esperarse ya que diferentes quimiotipos se han estudiado como HDACi. A pesar de ello los derivados de ácido hidroxámico destacan como inhibidores de HDAC. Esta familia de compuestos tienen variaciones estructurales pero mantienen sus propiedades farmacofóricas [36]. Esta observación fue capturada con la representación por GpiDAPH3 que indica poca diversidad en HDACi. En contraste, MACCS (una representación más "sensible" a cambios estructurales) mostró una colección con estructuras significativamente distintas. Para GRAS se observó una diversidad significativa, la cual se mantiene en las diferentes representaciones. En base a estos resultados se concluyó que resulta más conveniente usar más de una representación estructural para determinar de mejor manera la diversidad de una colección [32] [33].

6.2.2. Similitud inter-bibliotecas

Esta parte del estudio hace referencia a la posible intersección dentro del ESQUER. Aquí se comparó la similitud molecular de dos familias de compuestos (BRDi y HDACi) contra las bases de datos moleculares de referencia. Las representaciones se construyeron con base en la similitud máxima entre cada inhibidor y los compuestos en colecciones de referencia utilizando CDFs. Para el análisis sólo se consideraron MACCS y TGD. GpiDAPH3 no fue considerada pues una gran cantidad de compuestos se evalúan con "similitud cero".

La Figura A-5 muestra los gráficos de CDFs comparando con MACCS keys y TGD, respectivamente. En ella se muestran cinco colecciones de compuestos con BRDi, HDACis y DNMTi (empleados como referencia). Los comparativos para las demás colecciones de referencia pueden encontrarse en el Anexo (Figura S2).

Los valores bajos de CDF, tanto para MACCS como TGD indicaron que en general, las estructuras dentro de los grupos epigenéticos (BRDi, HDACis y DNMTi) son muy diferentes a las contenidas en las referencias. Un resultado notable fue la obtención de valores de similitud de la unidad, esto sería si el mismo compuesto fuera común a diferentes colecciones (de acuerdo a esa representación molecular específica). Tras una inspección visual se encontró que, si bien los compuestos no son idénticos, comparten características farmacofóricas. Un ejemplo notable son los derivados de bisfenilo con ácido hidroxámico. De hecho, recientemente se ha publicado un inhibidor dual BRD/HDAC con el grupo hidroximato [38]. Otro grupo común entre ambas bases es la fenilsulfonamida. Tanto las CDFs como estadísticos mostraron que FDA y 'Clinic' fueron más similares a BRDi, HDACis y DNMTi que dichos grupos entre sí. Este resultado dio soporte a la observación que los inhibidores de BRDs, HDACs y DNMT son, en general, significativamente diferentes.

De acuerdo a MACCS keys el orden relativo de similitud de GRAS respecto a los grupos epigenéticos es: DNMTi>HDACis>BRDi. Dicho de otra forma, si se hiciera una búsqueda por similitud en GRAS, las moléculas serían más semejantes a DNMTi. Sin embargo, si consideramos TGD el orden relativo es: BRDi>HDACis>DNMTi. Por lo que, de manera análoga a la sección anterior se destaca el hecho de usar más de una representación molecular. Una discusión más profunda sobre los comparativos de todas las colecciones va más allá del alcance del presente trabajo, pues el enfoque es sobre compuestos epigenéticos.

1. Por cuestiones de espacio, se han omitido figuras y tablas. Por favor refiérase al artículo anexo a este escrito, las referencias se harán mediante A-X, siendo X el número correspondiente a la figura del artículo original y SX referencias al material suplementario de dicho artículo.

6.3. Contenido de núcleos base

La Figura 1 muestra los quimiotipos más recurrentes (como se definió en la sección de Métodos) para las colecciones BRDi y HDACis. Aquellos con una frecuencia superior a 10 (BRDi) y 20 (HDACis) son los que se muestran (la diferencia es debido al tamaño de la colección). El anillo de benceno (código RYLFV) no se incluye en la figura (*vide infra*). Los núcleos base más frecuentes para DNMTi pueden encontrarse en el trabajo de Fernández-de Gortari et al [21]. El análisis de núcleos base está organizado en tres partes: contenido, diversidad y enriquecimiento (distribución de actividad).

6.3.1 Quimiotipos (núcleos base)

Para BRDi el quimiotipo más frecuente (código 1AWRP) se presentó 14 veces (6.8%) seguido por el quimiotipo 49ZJ3 con 12 (5.8%) compuestos. Como se observa en la Figura 1, muchos de los núcleos frecuentes tienen de 3 a 4 anillos. En contraste, el quimiotipo más frecuente para HDACis tiene sólo un anillo aromático (41 compuestos, 2%). Lo interesante es que este núcleo es el bencimidazol, presente como subestructura de 1AWRP. La elevada prevalencia del anillo de bencimidazol en compuestos con actividad biológica está bien documentada [29] [35]. Estos hallazgos dan apoyo a la propuesta de diseñar “llaves maestras” usando al bencimidazol como punto de partida.

No fue de sorprender que el anillo de benceno (RYLFV) tuviera una prevalencia alta dentro de HDACis con una frecuencia de 127 (6.35%). Este anillo está presente con alta frecuencia en otras colecciones [35]. Lo que fue inesperado fue la ausencia de este anillo como núcleo base principal en BRDi (aunque está presente como subestructura). Una tendencia similar se obtuvo para el quimiotipo “00000’’ (compuestos acíclicos) cuya prevalencia fue similar a la del benceno [35]. Pero igualmente ausente en BRDi como núcleo base.

Otro resultado notable es que tanto para HDACis como BRDi los quimiotipos recuperados poseen una relación estructura-actividad continua. Esto se refiere a núcleos base que no son propensos a tener acantilados de actividad (*activity cliffs*). Sin embargo el sistema DM3VV puede resultar controversial debido a que posee un grupo tipo PAINs (del inglés Pan Assay Interference compounds), los cuales se caracterizan por causar falsos positivos en pruebas biológicas [40].

6.3.2. Diversidad de núcleos base

Las curvas de recobro de quimiotipos (CRQ) para las diferentes colecciones se muestran en la Figura 2. En forma similar a los gráficos anteriores la estadística descriptiva se muestra en el anexo (Tabla S8). Para generar las gráficas CRQ se comparó la fracción de cada quimiotipo respecto a la fracción de compuestos con dicho núcleo en la base de datos.

Considere entonces una colección "íntegramente diversa"; es decir, cada compuesto posee quimiotipo distinto. Entonces la CRQ será una línea recta y diagonal, dado que *x* e *y* son equivalentes el área bajo la curva (ABC) viene de la integral:

1. Por cuestiones de espacio, se han omitido figuras y tablas. Por favor refiérase al artículo anexo a este escrito, las referencias se harán mediante A-X, siendo X el número correspondiente a la figura del artículo original y SX referencias al material suplementario de dicho artículo.

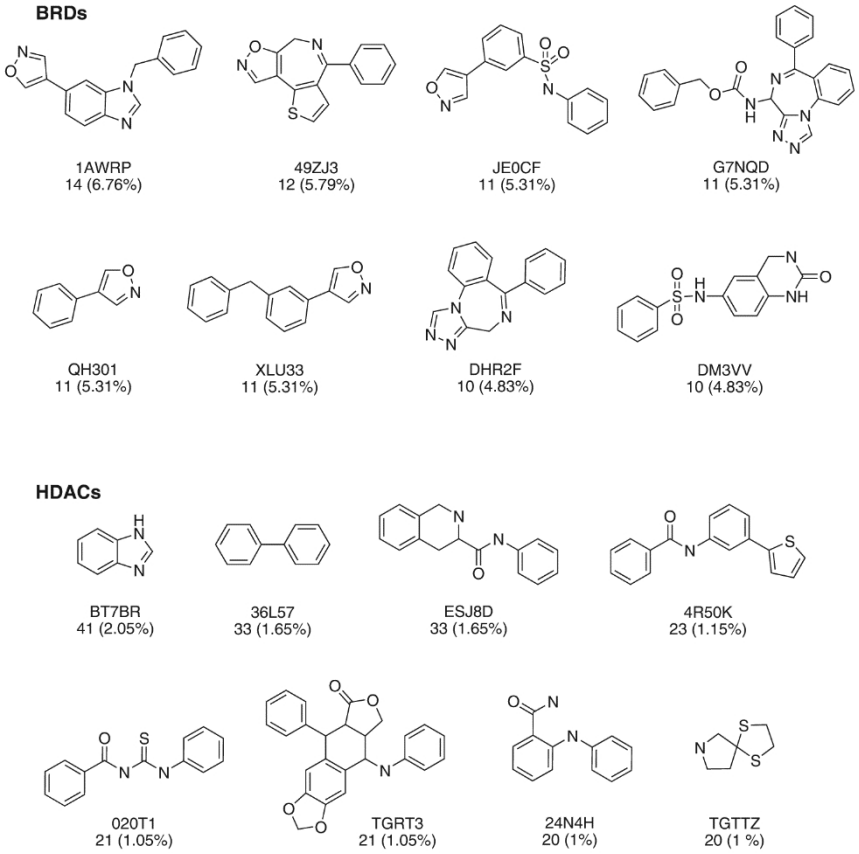


Figura 1. Quimiotipos con mayor frecuencia en las bases BRDi y HDACi. Se considera como alta frecuencia 10 y 15 estructuras presentes, respectivamente. La inclusión del anillo de benceno fue omitida. Se muestran los identificadores alfanuméricos de MEQI, así como el porcentaje de estructuras con dicho núcleo base.

$$\int_0^1 x\,dx = \frac{x^2}{2} \therefore A = 0.5$$

De la expresión anterior se desprende que la diversidad puede ser como máximo 0.5. Por lo tanto: $A \in [0.5, 1]$, siendo uno el valor máximo de la similitud (la base posee un quimiotipo único). Dicho esto, al analizar la Figura 2 se estableció que el orden relativo de diversidad de núcleos base fue DNMTi > BRDi > HDACis > GRAS. Resulta notable que, de acuerdo al ABC, BRDi y HDACis (ABC = 0.74 y 0.76, respectivamente) poseen una diversidad comparable a inhibidores de los receptores andrógino, de estrógeno, glucocorticoide y angiotensina [41].

La diversidad de quimiotipos para DNMTi es similar a la de inhibidores de la proteína acarreadora acetil-enolil reductasa [41]. El mismo orden relativo en diversidad se mantuvo para la F_{50}. Para GRAS el 50% de los compuestos se encuentra en 0.4% de los quimiotipos: En contraste para DNMTi la mitad de los compuestos se distribuyen en el 22% de los núcleos base. HDACis tuvo la menor diversidad de núcleos base lo cual se asocia con la alta presencia de inhibidores relacionados (hidroximatos) disminuyen el número de quimiotipos diferentes.

1. Por cuestiones de espacio, se han omitido figuras y tablas. Por favor refiérase al artículo anexo a este escrito, las referencias se harán mediante A-X, siendo X el número correspondiente a la figura del artículo original y SX referencias al material suplementario de dicho artículo.

Un punto que podría resultar contradictorio es que el orden relativo de diversidad cambia notablemente si se consideran FGP o núcleos base. Es decir, la diversidad molecular depende del criterio para representar a las estructuras. No obstante, hay que considerar que un FGP considera la naturaleza misma del compuesto incluyendo cadenas laterales y/o complejidad.

Un claro ejemplo de esto es la colección GRAS: de acuerdo a MACCS keys su diversidad es alta, mientras que la CRQ nos habla de una diversidad de quimiotipos escasa, si lo comparamos a otros grupos. Esto es de resaltar pues hasta ahora no se había reportado dicha diversidad para GRAS.

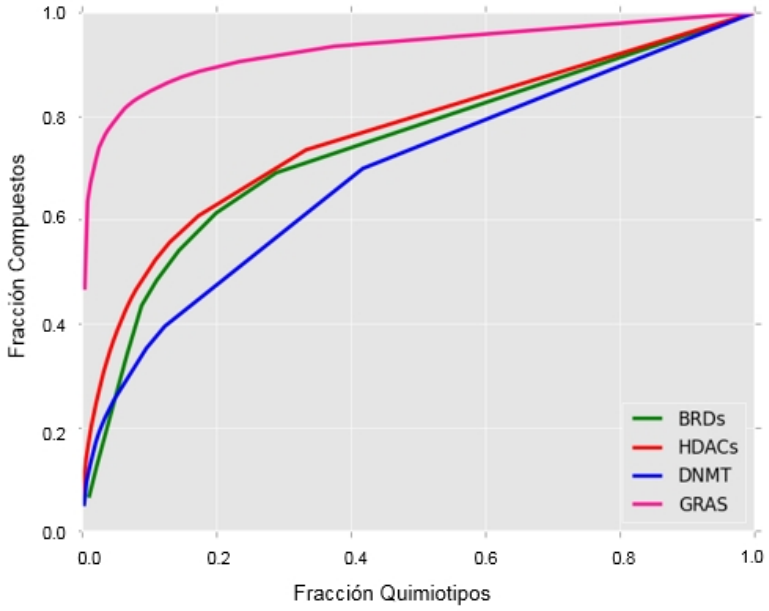


Figura 2. Curvas de recobro de núcleos base para las colecciones con información epigenéticamente relevante y la colección GRAS.

6.3.3. Distribución de actividad (enriquecimiento de quimiotipos)

Una vez caracterizado el contenido y diversidad de núcleos base dentro de BRDi y HDACi, se exploró si algún núcleo base posee una mayor proporción de compuestos activos. Este parámetro se denomina factor de enriquecimiento (EF). Los gráficos de enriquecimiento, así como la estadística involucrada se presentan en el anexo (Figura S3 y Tabla S8). Tanto para BRDi y HDACis el enriquecimiento resultó muy por debajo de la unidad. Esto quiere decir que en esta colección de compuestos no hay núcleos base particularmente 'enriquecidos' por compuestos activos. La Figura 1 (*vide supra*) muestra los quimiotipos con mayor enriquecimiento y frecuencia para BRDi (1AWRP, DM3VV, 49ZJ3 y G7NQD) y HDACis (BT7BR).

6.4. Complejidad molecular

Se ha comentado que una colección diversa permite explorar "regiones oscuras" dentro del espacio químico. Ante un concepto abstracto podría pensarse que encontrar regiones "vírgenes" en el universo de moléculas es sencillo. La dificultad surge al sondear dicho espacio en busca de áreas con relevancia terapéutica.

1. Por cuestiones de espacio, se han omitido figuras y tablas. Por favor refiérase al artículo anexo a este escrito, las referencias se harán mediante A-X, siendo X el número correspondiente a la figura del artículo original y SX referencias al material suplementario de dicho artículo.

Recientemente se han desarrollado metodologías que permiten aventurarse exitosamente en el océano químico. Una estrategia importante está enfocada en la complejidad molecular. Hasta donde sabemos este es el primer estudio sobre inhibidores epigenéticos donde se describe su complejidad. De manera análoga a análisis anteriores los resultados de complejidad se presentan como una distribución usando *boxplots* (Figura A-8). Representando la fracción de carbonos sp^3 (A-8A) y fracción de carbonos quirales (A-8B). En el primer caso, la fracción de C-sp^3 puede relacionarse con la "linearidad" de una molécula (cadenas laterales). Por tanto, no es de sorprender que GRAS destacara en este aspecto al ser principalmente moléculas de cadena larga. Este resultado estuvo de acuerdo con la alta prevalencia del quimiotipo 00000 (*vide supra*). En cuanto a las demás colecciones se observó una tendencia decreciente en el orden: FDA > Clinic > General > Epi-bloq. Este resultado sugiere la misma tendencia señalada por Lovering y cols. [30] (*vide supra*).

Precisamente, se observó que Epi-bloq es menos tridimensional y al considerar los grupos de datos contra dianas epigenéticas específicas (BRDi, HDACis, DNMTi) se observó una disminución notable en F-sp^3. Esto puede atribuirse a la naturaleza nucleósidica (DNMTi), cíclica (BRDi) y aromática (HDACis) de los compuestos. En contraste, la fracción de carbonos quirales mostró tendencias interesantes. Si bien, la tendencia en general es la similar, para este caso es FDA el punto de partida seguido por GRAS y Clinic, nuevamente se observó una relación subyacente entre la complejidad y la selectividad de un compuesto.

Este último punto se ha comentado previamente, en un estudio realizado por Clemons et al., [42] donde se estableció que un arreglo tridimensional es más adecuado para explorar el reconocimiento ligando-proteína. Esto podría parecer obvio, considerando la naturaleza disimétrica de los receptores y algunos ligandos; "un medio quiral distingue lo quiral" [43]. No obstante, es preciso recordar que muchos ligandos endógenos poseen una simplicidad notable y en muchos casos son reconocidos por una plétora de receptores. Aunado a esto estudios CoMFA/CoMSIA y modelos farmacofóricos se han empleado satisfactoriamente al describir moléculas en términos simples [36]. Por esta razón es prematuro considerar que la selectividad y complejidad están fuertemente relacionadas, si bien; es cierto que la complejidad ha resuelto problemas biofarmacéuticos y de formulación, en un contexto de diseño de fármacos hay muchas preguntas sin respuesta [44].

6.5. Cribado virtual

A partir de las colecciones antes mencionadas se realizó un protocolo de cribado virtual para identificar inhibidores potenciales de BRD4. Como primer paso se eligieron compuestos por similitud bidimensional. Debido a que este tipo de representación recupera moléculas que tienen grupos o conectividades similares, se utilizó un segundo criterio de similitud. En este caso se obtuvo el parecido 3D, es decir, se compararon los compuestos contra una serie de ligandos co-cristalizados. Esto se hizo para determinar si una molécula que no es propiamente similar en conectividad molecular puede acoplarse al sitio activo de BRD4.

1. Por cuestiones de espacio, se han omitido figuras y tablas. Por favor refiérase al artículo anexo a este escrito, las referencias se harán mediante A-X, siendo X el número correspondiente a la figura del artículo original y SX referencias al material suplementario de dicho artículo.

6.6. Similitud por FGP (2D)

Como se observa en la Figura A-5, se estableció la existencia de compuestos similares a BRDi en las colecciones de referencia. A partir de la construcción de matrices de similitud fue posible determinar que compuestos son más similares en forma representativa. Como se detalló en la sección de Métodos se usaron dos tipos de representaciones; MACCS keys la cual recupera fragmentos, y la huella digital molecular *Extended Connectivity Fingerprint* a 4 enlaces (ECFP4). Este FGP hace referencia a la conectividad extendida de una molécula, por ésta razón pertenece a la categoría de representación topológica. En forma breve, este *fingerprint* construye las moléculas mediante vectores o "caminos": éstos parten de un átomo central determinado por un algoritmo (de Morgan) hasta un vértice de conectividad (normalmente 3 o 4 enlaces desde el átomo central). Se puede intuir entonces, que la representación topológica es más sensible a los cambios estructurales. Por tanto, es de esperar que los valores de similitud sean bajos (respecto a los valores por MACCS keys).

De ahí que los valores mínimos de selección sean distintos para cada *fingerprint*. Además, ECFP4 es una representación común en protocolos de cribado virtual por su alto desempeño [45]. Por otra parte, al usar la conectividad se garantiza que las estructuras difieran significativamente de los núcleos base, pero manteniendo patrones farmacofóricos semejantes. La Figura 3 presenta algunos núcleos base, estudiados como BRDi y aquellos recobrados por similitud.

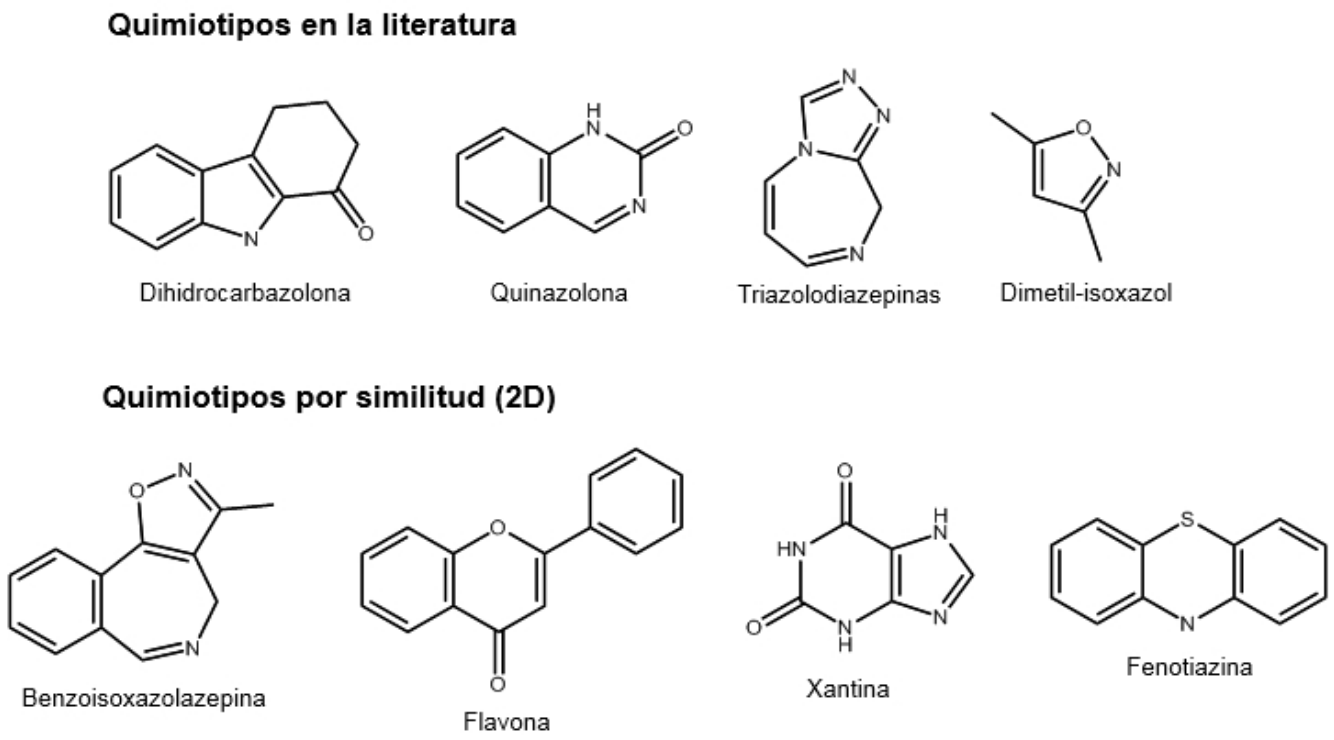


Figura 3. Comparativo de núcleos base asociados a la inhibición de BRDs y posibles quimiotipos similares por representación bidimensional.

1. Por cuestiones de espacio, se han omitido figuras y tablas. Por favor refiérase al artículo anexo a este escrito, las referencias se harán mediante A-X, siendo X el número correspondiente a la figura del artículo original y SX referencias al material suplementario de dicho artículo.

6.7. Alineamiento flexible (similitud 3D)

La similitud no sólo se limita a que un par de moléculas contengan fragmentos o grupos en común. A nivel farmacodinámico la similitud puede entenderse en términos de mecanismo de acción (bioisoterismo). Por esta razón se realizó la búsqueda por alineamiento flexible la cual compara estructuras en términos de forma y volumen (Figura 4). Se evaluaron cerca de 5000 estructuras en busca de similitud. De ellas más de 300 presentaron similitud bidimensional (MACCS > 0.6/ ECFP > 0.2). Para el alineamiento flexible, el número de compuestos con similitud significativa (tomando de línea base el valor promedio de S, -100 aprox.) fue: 70 estructuras de FDA, 20 productos naturales, 44 compuestos GRAS y 5 derivados de bencimidazoles.

De las estructuras encontradas destacan los núcleos flavonoides y los esteroides. Los flavonoides son productos naturales presentes en casi todas las especies vegetales [46]. Se han utilizado de manera satisfactoria como agentes antiinflamatorios [47], antidepresivos [46], sedantes [48] e incluso se han usado para tratar la demencia senil [49,50]. Por otra parte, los esteroides son compuestos que fungen como hormonas modulando una plétora de procesos fisiológicos [51]. La función más conocida es la regulación de caracteres y conductas sexuales [52]; no obstante, también se han asociado al insomnio [53] y a varios tipos de cáncer [54]. Es importante destacar este último punto, pues la posibilidad de que los esteroides actúen sobre los bromodominios sugiere una relación con la existencia de la isoforma BRDT (bromodomain específico testicular por sus siglas en inglés).

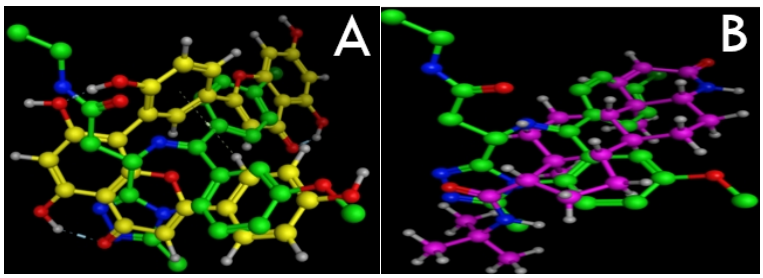


Figura 4. Ejemplos de alineamiento flexible. En verde se muestra la referencia. A) Núcleo flavonoide. B) Núcleo esteroide.

6.8. Acoplamiento molecular (*docking*)

Después de identificar núcleos base o estructuras similares, estas se sometieron a acoplamiento molecular. Para ello se utilizó como referencia la estructura BRD4 (PDB ID: 3P5O), usando el ligando co-cristalizado para comparaciones posteriores (*vide infra*).

El *docking* permite hacer estimaciones para determinar si una molécula puede interactuar con una proteína dada. Para ello utiliza algoritmos que tienen distintas maneras de evaluar el modo de unión y determinar un mejor confórmero. Por esta razón es recomendable usar más de un programa de *docking* a fin de hacer una mejor predicción. A esto se le llama *consensus scoring* [55].

Como se mencionó en la sección de Métodos se usaron los programas Autodock, Autodock Vina, MOE e ICM para elegir inhibidores potenciales. El primer paso para determinar si un programa es apto para realizar predicciones, es hacer la validación. Esto se realiza al acoplar nuevamente el ligando co-cristalizado,

1. Por cuestiones de espacio, se han omitido figuras y tablas. Por favor refiérase al artículo anexo a este escrito, las referencias se harán mediante A-X, siendo X el número correspondiente a la figura del artículo original y SX referencias al material suplementario de dicho artículo.

determinando mediante RMSD la diferencia entre confórmeros [56]. Los resultados de la validación se muestran en la Figura 5.

Una vez determinada una “línea base”, se usa el valor de energía del confórmero co-cristalizado como criterio de selección de compuestos. Para efectuar el consenso, la energía de unión debía ser cercana a la línea base en 3 de 4 algoritmos. La Figura 6 presenta los *hits* computacionales.

6.8.1. Análisis de confórmeros

Actualmente, se sabe poco sobre la función de los bromodominios. Recientemente se han caracterizado puntos críticos en el modo de unión, entre ellas la presencia de aguas estructurales e interacciones por puente de hidrógeno [18]. Como parte de la validación se observó si se conservan dichas interacciones.

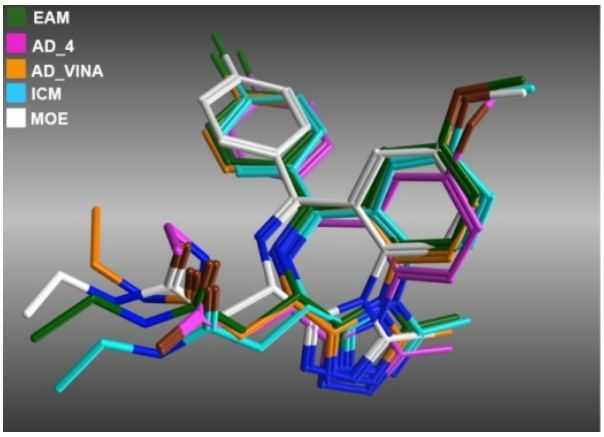


Figura 5. Validación del acoplamiento molecular. Se muestran los modos de unión (*poses*) con menor RMSD respecto a la referencia (verde), para los cuatro algoritmos utilizados.

Los *hits* consenso se evaluaron en forma semejante buscando estos modos de unión. Esto se conoce como evaluación de PLIF. El histograma de las interacciones más recurrentes se muestra en la Figura 7.

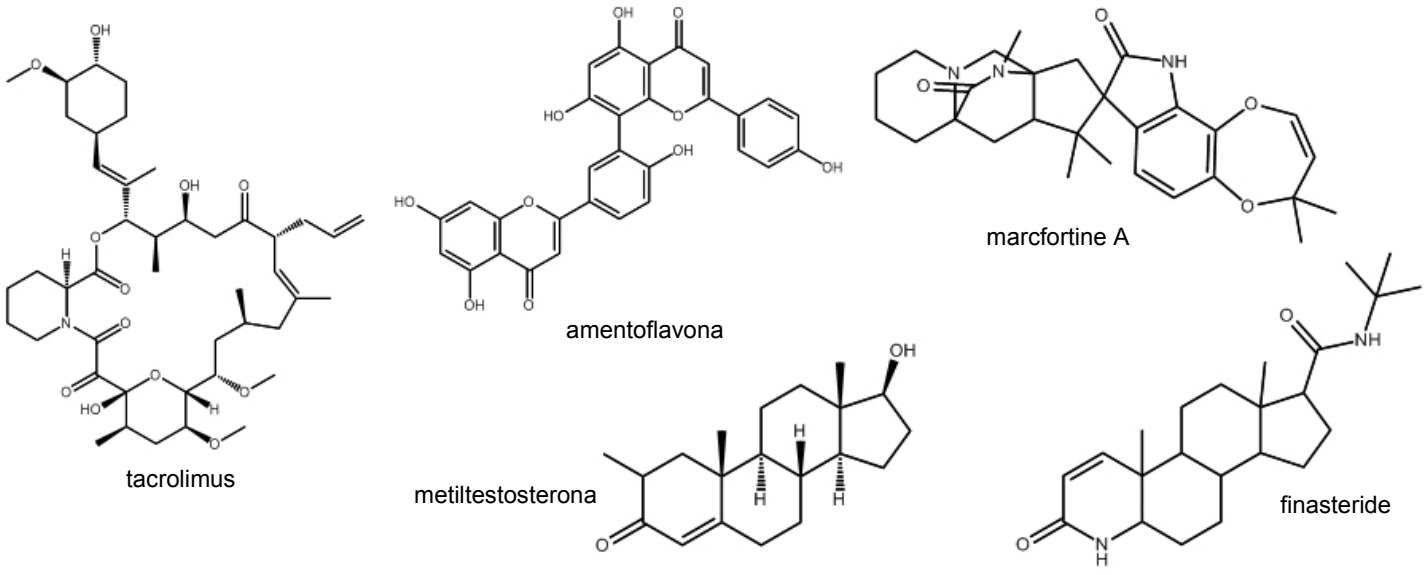


Figura 6. Estructuras químicas de los *hits* identificados por el protocolo de cribado virtual como inhibidores potenciales de BRD4.

1. Por cuestiones de espacio, se han omitido figuras y tablas. Por favor refiérase al artículo anexo a este escrito, las referencias se harán mediante A-X, siendo X el número correspondiente a la figura del artículo original y SX referencias al material suplementario de dicho artículo.

El mecanismo de reconocimiento y unión es el siguiente: el grupo acilo es reconocido mediante un puente de hidrógeno con el residuo Asp140 fungiendo como “ancla” del ligando. Posteriormente se establece un segundo puente de hidrógeno mediado por las aguas estructurales del sitio que interactúan con Trp81. También se reconoce la interacción con Leu92 y Leu94 que orientan redes aromáticas principalmente [14]. Se observó que para los distintos acoplamientos interactúan en forma distinta con el BRD4, no obstante, las interacciones recuperadas por acoplamiento incluyen a Trp81 y Leu92 con frecuencia significativa (Figura 6).

Las diferencias observadas pueden relacionarse a las distintas aproximaciones y algoritmo de búsqueda en los acoplamientos. En cuanto a Asp140, la naturaleza polar de la interacción dificulta su recuperación. En general los algoritmos de *docking* identificaron de manera diferente a los hidrógenos polares asignando torsiones y cargas variables.

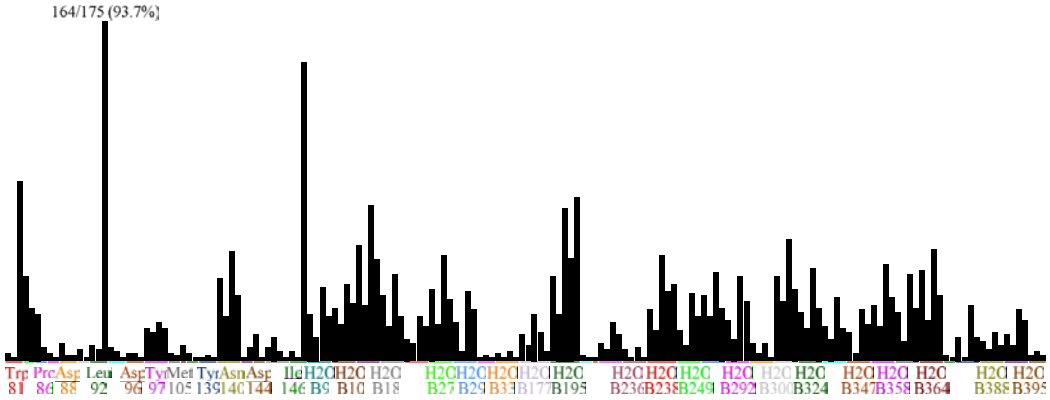


Figura 7. Histograma de interacciones representativas, representadas por PLIFs usando los diferentes algoritmos de acoplamiento. Para más información refiérase al texto principal.

Además, al estar vinculada por puente de hidrogeno a moléculas de agua es difícil reproducir la interacción pues muchos programas de *docking* no están optimizados para manejar estos sistemas [57]. Por tanto, la validación encontró adecuadamente la conformación probable del ligando. Sin embargo, hasta el momento con los algoritmos clásicos de *docking* resultará complicado reproducir el modo de unión por la presencia de aguas estructurales dentro del sitio activo. No obstante, fue posible recuperar muchas de las interacciones en un *hit*, particularmente usando MOE.

Los compuestos que resultaron más prometedores tanto por su conformación como modo de unión fueron el amentoflavona (flavonoide) y la finasterida (esteroide sintético) (Figura 6). Los flavonoides son compuestos con varios grupos polares, por lo cual no es extraño que interactúen por puentes de hidrógeno. No obstante, existen varios estudios sobre flavonoides como agentes terapéuticos, entre ellos sobre el receptor GABA [58]. Se ha observado que los flavonoides semisintéticos son excelentes moduladores del receptor. Más aún, de acuerdo a los trabajos de Fernández y cols [59], los grupos que son cruciales a esta acción no son precisamente polares si no electronegativos (Br, Cl, NO_2).

En cuanto a los esteroides, su mecanismo fisiológico está íntimamente relacionado con su capacidad de traslocar receptores. Es mediante este proceso nuclear que da inicio la síntesis de proteínas y otras señales

1. Por cuestiones de espacio, se han omitido figuras y tablas. Por favor refiérase al artículo anexo a este escrito, las referencias se harán mediante A-X, siendo X el número correspondiente a la figura del artículo original y SX referencias al material suplementario de dicho artículo.

mediadas por hormonas. Finasterida es un fármaco derivado de estereoides cuya indicación terapéutica es inhibidor de la 5-α-reductasa [60]. No obstante, en un inicio este fármaco se utilizó como terapia de cáncer prostático [61]. Esto sugiere la posibilidad de interacción con bromodominios (específicamente BRDT).

7. Conclusiones

- **Se desarrolló una quimiografía del ESQUER la cual mostró relaciones significativas a nivel fisicoquímico y estructural entre fármacos e inhibidores a BRDs y HDACs. Esta herramienta permitirá el mejor entendimiento de las relaciones farmacofóricas para estas dianas, facilitando el diseño de nuevos inhibidores y/o polifármacos.**
- **Las similitudes subyacentes entre compuestos epigenéticos con GRAS confirmaron que la dieta impacta de manera significativa en el epigenoma. Se espera que la caracterización de compuestos GRAS podrá facilitar la aplicación nutracéutica con fines de prevención.**
- **Se identificaron compuestos con potencial acción inhibitoria a BRD4, cuyos quimiotipos difieren significativamente de los reportados actualmente en la literatura.**
- **La presencia de núcleos esteroides; como BRDi potenciales, sugiere una posible modulación de la isoforma BRDT. Esta información podrá contribuir en el diseño del primer inhibidor específico a esta diana.**

8. Perspectivas

La epigenética continúa desarrollándose a un paso notable. No obstante, aún hay muchas preguntas que requieren respuesta. La clave para comprender y modular estos sistemas enzimáticos es dar el contexto apropiado al conocimiento actual. Este primer paso ha mostrado relaciones alentadoras sobre el ESQUER. El siguiente paso será profundizar en el estudio sobre las estructuras base aquí identificadas con el fin de elucidar el SAR asociado a BRDs y proponer nuevas estructuras privilegiadas.

Finalmente, los bromodominios son dianas epigenéticas diversas y relevantes, pero aún falta mucho por saber. Por tanto es importante realizar esfuerzos en la búsqueda e identificación de moléculas sonda. Con esto sería posible realizar estudios más precisos que permitan elucidar la fisiología de los bromodominios.

9. Referencias

1. D.C. Dolinoy. Epigenetics and Carcinogenesis. *Compr. Toxicol.* **14,** 293–309 (2010).
2. Wang, C. & Filippakopoulos, P. Beating the odds : BETs in disease. *Trends Biochem. Sci.* (2015).
3. Coletta, D. K. *Genetic and Epigenetics of Type 2 Diabetes*. *Pathobiology of Human Disease* (2014).
4. Casadesús, J. & Noyer-Weidner, M. Epigenetics. *Brenner's Encycl. Genet.* **2,** 500–503 (2013).
5. Lu, D. Epigenetic modification enzymes: catalytic mechanisms and inhibitors. *Acta Pharm. Sin. B* **3,** 141–149 (2013).
6. Wilting, R. H. & Dannenberg, J.-H. Epigenetic mechanisms in tumorigenesis, tumor cell heterogeneity and drug resistance. *Drug Resist. Updat.* **15,** 21–38 (2012).
7. Laker, R. C., Wlodek, M. E., Connelly, J. J. & Yan, Z. Epigenetic origins of metabolic disease: The impact of the maternal condition to the

1. Por cuestiones de espacio, se han omitido figuras y tablas. Por favor refiérase al artículo anexo a este escrito, las referencias se harán mediante A-X, siendo X el número correspondiente a la figura del artículo original y SX referencias al material suplementario de dicho artículo.

offspring epigenome and later health consequences. *Food Sci. Hum. Wellness* **2,** 1–11 (2013).

8. Majumder, S. & Advani, A. The epigenetic regulation of podocyte function in diabetes. *J. Diabetes Complications* (2015).
9. Liu, Y., Liu, K., Qin, S., Xu, C. & Min, J. Epigenetic targets and drug discovery: Part 1: Histone methylation. *Pharmacol. Ther.* (2014).
10. Yoo, J., Kim, J. H., Robertson, K. D. & Medina-Franco, J. L. *Molecular modeling of inhibitors of human DNa methyltransferase with a crystal structure: Discovery of a novel dnmt1 inhibitor*. *Advances in Protein Chemistry and Structural Biology* **87,** (2012).
11. Robert, C. & Rassool, F. V. in *Histone Deacetylase Inhibitors as Cancer Therapeutics* (2012).
12. Cherblanc, F. L., Davidson, R. W. M., Di Fruscia, P., Srimongkolpithak, N. & Fuchter, M. J. Perspectives on natural product epigenetic modulators in chemical biology and medicine. *Nat. Prod. Rep.* (2013).
13. Chen, Y. D., Jiang, Y. J., Zhou, J. W., Yu, Q. Sen & You, Q. D. Identification of ligand features essential for HDACs inhibitors by pharmacophore modeling. *J. Mol. Graph. Model.* (2008).
14. Ferri, E., Petosa, C. & Mckenna, C. E. Bromodomains : Structure , function and pharmacology of inhibition. *Biochem. Pharmacol.* (2015).
15. Duffy, B. C. *et al.* Bioorganic & Medicinal Chemistry Letters Discovery of a new chemical series of BRD4 (1) inhibitors using protein-ligand docking and structure-guided design. *Bioorg. Med. Chem. Lett.* (2015).
16. Filippakopoulos, P. *et al.* Bioorganic & Medicinal Chemistry Benzodiazepines and benzotriazepines as protein interaction inhibitors targeting bromodomains of the BET family. *Bioorg. Med. Chem.* (2012).
17. Galdeano, C. & Ciulli, A. Selectivity on-target of bromodomain chemical probes by structure-guided medicinal chemistry and chemical biology. *Future Med. Chem.* (2016).
18. Zhang, G., Smith, S. G. & Zhou, M.-M. Discovery of Chemical Inhibitors of Human Bromodomains. *Chem. Rev.* (2015).
19. Crawford, T. D. *et al.* Diving into the Water: Inducible Binding Conformations for BRD4, TAF1(2), BRD9, and CECR2 Bromodomains. *J. Med. Chem.* (2016).
20. Fourches, D., Muratov, E. & Tropsha, A. Curation of chemogenomics data. *Nat. Chem. Biol.* (2015).
21. Gortari, E. F. & Medina-Franco, J. L. Epigenetic relevant chemical space: a chemoinformatic characterization of inhibitors of DNA methyltransferases. *RSC Adv.* (2015).
22. Prieto-Martinez, F. D., Gortari, E. F., Mendez-Lucio, O. & Medina-Franco, J. L. A chemical space odyssey of inhibitors of histone deacetylases and bromodomains. *RSC Adv.* (2016).
23. Lovering, F. Escape from Flatland 2: complexity and promiscuity. *Medchemcomm* (2013).
24. Medina-Franco, J. L., Martínez-Mayorga, K., Peppard, T. L. & Del Rio, A. Chemoinformatic analysis of GRAS (Generally Recognized as Safe) flavor chemicals and natural products. *PLoS One* (2012).
25. Sander, T., Freyss, J., Kor, M. Von & Rufener, C. DataWarrior: An Open-Source Program For Chemistry Aware Data Visualization And Analysis. (2015).
26. Droettboom, M. *et al.* matplotlib: matplotlib v1.5.1. doi:10.5281/zenodo.44579
27. Xu, Y. J. & Johnson, M. Algorithm for Naming Molecular Equivalence Classes Represented by Labeled Pseudographs. *J. Chem. Inf. Comput. Sci.* (2001).
28. Medina-Franco, J. L., Petit, J. & Maggiora, G. M. Hierarchical strategy for identifying active chemotype classes in compound databases. *Chem. Biol. Drug Des.* (2006).
29. Yongye, A. B. & Medina-Franco, J. L. Toward an Efficient Approach to Identify Molecular Scaffolds Possessing Selective or Promiscuous Compounds. *Chem. Biol. Drug Des.* (2013).
30. Lovering, F., Bikker, J. & Humblet, C. Escape from Flatland: Increasing Saturation as an Approach to Improving Clinical Success. *J. Med. Chem.* (2009).
31. Molecular Operating Environment (MOE). (2016).
32. Smith, S. G., Sanchez, R. & Zhou, M. Review Privileged Diazepine Compounds and Their Emergence as Bromodomain Inhibitors. *Chem. Biol.* (2014).
33. Cha, S. Comprehensive Survey on Distance / Similarity Measures between Probability Density Functions. *Int. J. Math. Model. Methods Appl. Sci.* (2007).
34. López-Vallejo, F., Giulianotti, M. A., Houghten, R. A. & Medina-Franco, J. L. Expanding the medicinally relevant chemical space with compound libraries. *Drug Discov. Today.* (2012).
35. Singh, N. *et al.* Chemoinformatic analysis of combinatorial libraries, drugs, natural products, and molecular libraries Small Molecule Repository. *J. Chem. Inf. Model.* (2009).
36. Chen, Y. *et al.* 3D-QSAR studies of HDACs inhibitors using pharmacophore-based alignment. *Eur. J. Med. Chem.* (2009).
37. Saldívar-González, F., Prieto-Martínez, F. D. & Medina-Franco, J. L. Descubrimiento y desarrollo de fármacos: un enfoque computacional. *Educ. Química.* (2016).
38. Atkinson, S. J. *et al.* The structure based design

1. Por cuestiones de espacio, se han omitido figuras y tablas. Por favor refiérase al artículo anexo a este escrito, las referencias se harán mediante A-X, siendo X el número correspondiente a la figura del artículo original y SX referencias al material suplementario de dicho artículo.

of dual HDAC/BET inhibitors as novel epigenetic probes. *Med. Chem. Commun.* (2014).

39. Langdon, S. R., Blagg, J. & Brown, N. Scaffold Diversity in Medicinal Chemistry Space. *J. Chem. Inf. Model.* (2014).
40. Baell, J. & Walters, M. a. Chemical con artists foil drug discovery. *Nature* (2014).
41. Medina-Franco, J., Martínez-Mayorga, K., Bender, A. & Scior, T. Scaffold Diversity Analysis of Compound Data Sets Using an Entropy-Based Measure. *QSAR Comb. Sci.* (2009).
42. Clemons, P. A. *et al.* Small molecules of different origins have distinct distributions of structural complexity that correlate with protein-binding profiles. *Proc. Natl. Acad. Sci.* (2010).
43. Anslyn, E. V. & Dougherty, D. A. *Modern physical organic chemistry*. (2006).
44. Méndez-Lucio, O. & Medina-Franco, J. L. The many roles of molecular complexity in drug discovery. *Drug Discov. Today* (2016).
45. Da, C. & Kireev, D. Structural Protein–Ligand Interaction Fingerprints (SPLIF) for Structure-Based Virtual Screening: Method and Benchmark Study. *J. Chem. Inf. Model.* (2014).
46. Guan, L.-P. & Liu, B.-Y. Antidepressant-like effects and mechanisms of flavonoids and related analogues. *Eur. J. Med. Chem.* (2016).
47. Catarino, M. D., Talhi, O., Rabahi, A., Silva, A. M. S. & Cardoso, S. M. *The Antiinflammatory Potential of Flavonoids: Mechanistic Aspects*. *Studies in Natural Products Chemistry.* (2016).
48. Marder, M. & Paladini, A. C. GABA(A)-receptor ligands of flavonoid structure. *Curr. Top. Med. Chem.* (2002).
49. Perry, N. S. L., Bollen, C., Perry, E. K. & Ballard, C. Salvia for dementia therapy: review of pharmacological activity and pilot tolerability clinical trial. *Pharmacol. Biochem. Behav.* (2003).
50. Hwang, S.-L., Shih, P.-H. & Yen, G.-C. in *Diet and Nutrition in Dementia and Cognitive Decline* (2015).
51. Auchus, R. J. in *Knobil and Neill's Physiology of Reproduction.* (2015).
52. Ogino, Y., Sato, T. & Iguchi, T. in *Handbook of Hormones*. (2016).
53. Llanas, A. C., Hachul, H., Bittencourt, L. R. a & Tufik, S. Physical therapy reduces insomnia symptoms in postmenopausal women. *Maturitas* (2008).
54. Handa, R. J., Larco, D. O. & Wu, T. J. in *Reference Module in Biomedical Sciences* (2014).
55. Houston, D. R. & Walkinshaw, M. D. Consensus docking: Improving the reliability of docking in a virtual screening context. *J. Chem. Inf. Model.* (2013).
56. Tuccinardi, T., Poli, G., Romboli, V., Giordano, A. & Martinelli, A. Extensive Consensus Docking Evaluation for Ligand Pose Prediction and Virtual Screening Studies. *J. Chem. Inf. Model.* (2014).
57. Forli, S. & Olson, A. J. A Force Field with Discrete Displaceable Waters and Desolvation Entropy for Hydrated Ligand Docking. *J. Med. Chem.* (2012).
58. Ghamali, M. *et al.* Combining DFT and QSAR computation to predict the interaction of flavonoids with the GABA (A) receptor using electronic and topological descriptors. *J. Taibah Univ. Sci.* (2016).
59. Fernández, S. P., Wasowski, C., Paladini, A. C. & Marder, M. Synergistic interaction between hesperidin, a natural flavonoid, and diazepam. *Eur. J. Pharmacol.* (2005).
60. Thareja, S. Steroidal 5α-reductase inhibitors: A comparative 3D-QSAR study review. *Chem. Rev.* (2015).
61. Thompson, I. M. *et al.* Long-Term Survival of Participants in the Prostate Cancer Prevention Trial. *N. Engl. J. Med.* (2013).

1. Por cuestiones de espacio, se han omitido figuras y tablas. Por favor refiérase al artículo anexo a este escrito, las referencias se harán mediante A-X, siendo X el número correspondiente a la figura del artículo original y SX referencias al material suplementario de dicho artículo.

RSC Advances

PAPER

Cite this: *RSC Adv.*, 2016, **6**, 56225

A chemical space odyssey of inhibitors of histone deacetylases and bromodomains†

Fernando D. Prieto-Martínez, Eli Fernández-de Gortari, Oscar Méndez-Lucio and José L. Medina-Franco*

Received 18th March 2016
Accepted 4th June 2016

DOI: 10.1039/c6ra07224k

www.rsc.org/advances

The interest in epigenetic drug and probe discovery is growing as reflected in the large amount of structure-epigenetic activity information available. Therefore, the significance of understanding the entire or fractions of the epigenetic relevant chemical space is increasing. Major epigenetic targets are histone lysine deacetylases (HDACs), bromodomains (BRDs), and DNA methyltransferases (DNMTs). However, with the exception of DNMTs, characterization of the chemical space of these epi-targets is limited. This work is the first chemoinformatic analysis of the physicochemical properties, structural diversity, and coverage of the chemical space of compounds screened as inhibitors of HDACs and BRDs. The chemical space was compared to DNMTis, approved drugs, commercial screening compounds, and generally recognized as safe (GRAS) molecules. The structural complexity of compounds directed towards epigenetic targets was also addressed. The outcome of this analysis indicated that it is required to increase the structural diversity and molecular complexity of screening libraries tested as modulators of DNMTs, HDACs and BRDs. Results also suggested that it is feasible to develop dual inhibitors targeting HDACS and BRDs. This work has implications in repurposing of food chemicals with potential epigenetic activity and design of poly-epigenetic compounds.

1. Introduction

Every living being has the ability to inherit its genetic material. However this process is not flawless. After a few decades, the study of DNA repair lead to the discovery of higher order mechanisms and the term 'epigenetics' was coined.[1] Initially, inhibition of epigenetic targets was considered a novel alternative for the treatment of cancer. While this approach may be true, current research showed that environmental factors such as radiation exposure, nutritional history, dietary intake, reproductive factors, among others, also play a key role on the expression of specific epigenetic modifications.[1–3] Nowadays it is accepted that epi-modulation can act as a link between genotype and environment stimuli[4] and may be used as a Rosetta Stone to better understand, prevent and/or cure diseases. While this is yet to be proved, many researchers consider epigenetics as the missing link on the biogenesis of chronic diseases like Alzheimer's, dementia, schizophrenia, diabetes, metabolic syndrome, to name few examples.[5,6]

Chemical modifications are key features in epigenetics. Although the number of reactions and enzymes involved are different and comprise more than one hundred, it is possible to distinguish three functions: writers, erasers and readers. Writers add chemical groups that can be labile or stable. Erasers remove the groups added by writing enzymes. Readers are 'effector proteins' that identify specific chemical groups associated with epigenetic modifications and produce large scale changes such as chromatin remodeling or recruitment of other enzymes involved in DNA replication or gene expression.[7]

The correlation between epigenetic changes and carcinogenesis attracted attention to histone deacetylases (HDACs). Acetylation on lysine residues is one of the most common processes on epigenetics.[1] Eighteen different HDACs have been identified, characterized and classified in three classes. Class I comprises HDACs 1, 2, 3 and 8, that are located on the nucleus with involvement on development of numerous cancer types.[8] Class III gathers seven HDACs that are NAD^+ dependent and sirtuin constituted and are known as SIRT 1–7. This class has been mainly involved with pancreas and breast cancer, nevertheless some of them (*e.g.* SIRT1) may be involved with type II diabetes.[9] Although the removal of acetate groups from histone tails may be conceived as the first step towards transcriptional repression, it has been shown that regulation by HDACs goes beyond histones acting in a plethora of cellular pathways.[10] Despite major efforts from industry and academia and the baffling amount of chemical and biochemical studies towards these enzymes, the Food and Drug Administration (FDA) of the United States has approved only four drugs so far for clinical use: vorinostat, romidepsin, belinostat, and panobinostat.

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

† Electronic supplementary information (ESI) available. See DOI: 10.1039/c6ra07224k

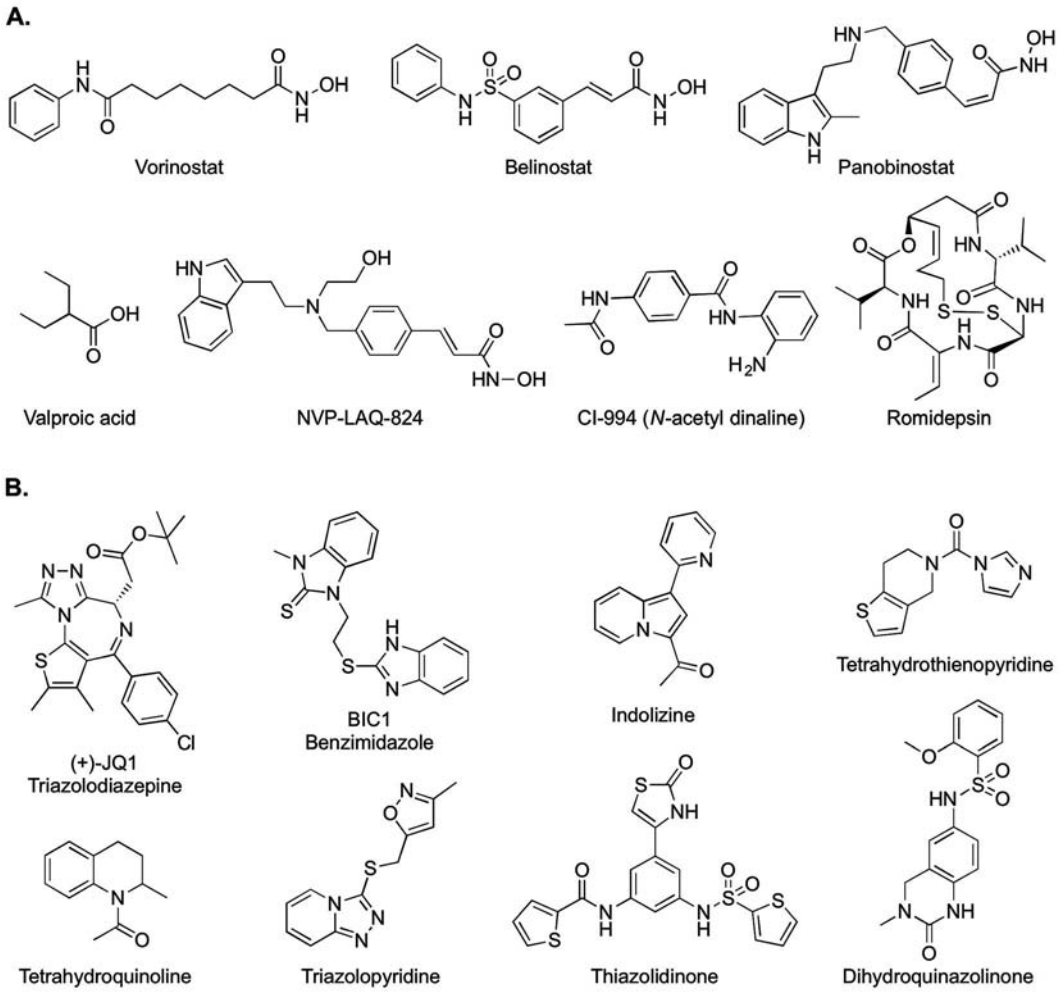


Fig. 1 Chemical structures of representative (A) HDAC inhibitors and (B) BRD inhibitors.

Fig. 1A shows the chemical structures of different HDAC inhibitors (HDACis) including those approved for clinical use.

More recently, epigenetic drug and probe discovery has turned to other targets such as the Bromodomain and Extra-Terminal domain (BET) protein family. From this family, the bromodomains (BRDs) BRD2, BRD3 and BRD4 have become of particular therapeutic interest. BRDs are structural motifs that recognize acetylated lysines, mainly those located in histone tails.[11] BRDs decide the ultimate fate of histones and their function has been correlated with cancer and inflammatory disease.[12] Also BRDs are responsible for crucial steps during cell cycles.[13] Fig. 1B depicts representative BRDs inhibitors (BRDis) including different scaffolds and acetyl-lysine mimicking groups.

Drug development is a daunting task and epigenetic drug discovery is not an exception. One of the main drawbacks is the high attrition rate,[14] which may be associated with the traditional trend in drug discovery to design specific drugs. Indeed, there is increasing evidence that drugs may act as 'master key compounds',[15] *i.e.* drugs exhibit biological activity through the interaction with a set of selected targets with reduced affinity for off-targets (at therapeutic doses). Thus, polypharmacology is largely influencing drug discovery strategies including the discovery of epigenetic drugs.[16] One of the strategies to explore the development of poly-epigenetic compounds is the assessment of the diversity and chemical space coverage of compounds with known epigenetic activity. The putative intersection in chemical space by approved drugs and compounds with epigenetic activity may lead to strategies to conduct drug repurposing.[17] Similarly, the intersection in chemical space of epigenetic compounds and diverse screening collections offers the possibilities to guide focused library design within novel regions in chemical space.[18] Furthermore, the comparison of chemical spaces of epigenetic compounds and food-related molecules may lead to the systematic elucidation of food chemical as bioactive epigenetic molecules. As a proof-of-concept, chemoinformatic mining of generally recognized as safe (GRAS) compounds, widely used in the food industry, lead to the identification of HDACis.[19]

There are several chemoinformatic tools that enable the analysis of the coverage and diversity of the chemical space of public data sets of epigenetic compounds. Despite the fact that public databases do not necessarily cover the entire current knowledge of epigenetic collections, they represent a reasonable

starting point to better understand the entire or fractions of the chemical space covered by epigenetic-related compounds *i.e.*, Epigenetic Relevant Chemical Space (ERCS). In fact, the wide variety of structures and molecules currently studied by epigenetics accounts for more than 5000 compounds available in the public domain with structure–activity data. As preliminary data, the chemical space of DNMT1 inhibitors (DNMTis) has been recently reported.[20]

The goal of the present study is to characterize the chemical space of HDACis and BRDis currently stored in two major public databases: ChEMBL and Binding Database (see below). Therefore, in light of the emerging research area of Epi-informatics,[21] this work contributes to the understanding of a fraction of ERCS. Of note, the compounds analyzed thought the study and deposited in ChEMBL and Binding Database have been developed not as part of drug discovery projects but also in the development of molecular probes. Chemoinformatic characterization of compound data sets is extensively documented by our and other groups to be an essential component of drug, lead and molecular probe discovery projects.[22,23] As discussed through the study and emphasized in the Conclusions section, the outcome of the analysis provided specific information that can be used to develop novel and improved epigenetic compounds. These collections were compared to DNMTis, approved drugs, compounds in clinical trials, a general screening collection typically used in high-throughput screening (HTS), and a commercial screening collection focused on epigenetic targets. Moreover, the epigenetic-related libraries were compared to GRAS chemicals. In order to conduct the comparisons, four complementary criteria were used including physicochemical properties (PCP) of pharmaceutical relevance, molecular fingerprints, molecular scaffolds, and established measures of molecular complexity. To the best of our knowledge, this is the first analysis of the structural complexity of epigenetic databases. As discussed throughout the study, the insights of these analyses provided sound basis to conduct computer-aided drug repurposing, identify bioactive compounds from food chemicals, and guide library design. The intersection of epigenetic target spaces may also indicate the feasibility of developing dual or poly-epigenetic modulators. Of note, the findings of this work directly impact not only the development of therapeutic agents but also on molecular probes.

2. Methods

The epigenetic and reference collections were retrieved from public databases. Curated data sets were analyzed and compared to each other using complementary molecular representations: six PCP, three structural fingerprints, and molecular scaffolds. The methods used to represent the molecular structures and perform the visual and quantitative comparisons are described in this section.

2.1 Datasets and data curation

The chemical structures and activity data of HDACis (against isoforms 1, 2, 3, 8 and SIRT1) and BRDis (against isoforms 2, 3, 4) were downloaded from ChEMBL[24] version 20, and Binding Database (BDB).[25] The HDACis analyzed in this work had been associated with the treatment of different types of cancer including colon, prostate, gastric and cervical,[8] and they may also play key roles on diabetes[26] and cardiovascular diseases.[27] Likewise, the BRDis had been associated with activity against cancer cell lines.[26] The structures were downloaded as simplified molecular input line entries (SMILES) and then converted to three-dimensional (3D) structures using Molecular Operating Environment (MOE).[28] To curate the data sets, which is a major step in chemoinformatic analysis of compound data sets[29] we employed the 'Wash' module implemented in MOE. Using this module, structures were prepared by disconnecting metal salts, remove simple components, and rebalancing protonation states. Molecules were energy minimized with the MMFF94x force field using the default stochastic sampling algorithm followed by LowModeMDsearch implemented in MOE. The stereochemistry annotated in the SMILES string was used. The resulting data sets were visually inspected and final details were corrected. Table 1 summarizes the compounds initially downloaded and the number of compounds after curation.

2.1.1 Reference databases. The data sets of BRDs and HDACis were compared to a database of DNMTis analyzed recently[20] and five additional reference collections: a data set of food additives and flavors obtained from FEMA GRAS list (GRAS 25) previously curated ('GRAS');[30] drugs approved for clinical use obtained from DrugBank ('Drugs'); compounds currently on clinical trials as reported by the Therapeutic Target Database (TTD), ('Clinic'); a commercial collection focused on epigenetic targets available at Selleckchem ('Epi-focused'); and a general screening collection obtained from Selleckchem ('General'). With the exception of GRAS, these data sets were used as reference in a recent study by Fernández-de Gortari *et al.*[20] Table 1 summarizes the reference data sets.

2.2 Molecular representation

Since the chemical space depends on the structure representation, multiple criteria were implemented, namely; PCP of pharmaceutical relevance, structural fingerprints of different design and molecular scaffolds.[20,30] Structural complexity using well-established methods was also measured.[31] Details of each method are provided hereunder.

2.2.1 Physicochemical properties. Six PCP were computed using MOE, namely, partition coefficient octanol/water (Slog *P*), rotatable bonds (RTB), hydrogen bond acceptors (HBA), hydrogen bond donors (HBD), topological surface area (TPSA), and molecular weight (MW). This set of properties is typically used to compare compound data sets in drug discovery[20] and was employed to make cross-comparisons with studies that report the characterization of other data sets. The distribution of the properties was assessed with box plots, generated with R Studio using the PMCMR package,[32] and summary statistics of the distributions. In order to generate a visual representation of the chemical space based on properties, principal component analysis (PCA) was conducted with MOE. Data visualization was performed with Data Warrior.[33] The statistical comparison of the properties was made by assessment of homoscedasticity

Table 1 Summary of the data sets analyzed in this work

	Size			
	Number of compounds	Unique molecules	Source	URL
Data set				
HDACs	>5000	2000	ChEMBL	https://www.ebi.ac.uk/chembl/
BRDs	~2000	207	ChEMBL, BDB	https://www.bindingdb.org/
Reference sets				
DNMTs	>5000	565	ChEMBL, BDB, HEMD	http://xlink.rsc.org/?DOI=C5RA19611F
Drugs	>5000	1490	DrugBank	http://www.drugbank.ca/drugs/
Clinic	1151	837	Therapeutic target database	http://bidd.nus.edu.sg/group/cjttd/
General	1224	1100	Selleck	http://www.selleckchem.com
Epi-focused	128	113	Selleck	http://www.selleckchem.com
GRAS	2200	2200	FEMA	http://dx.plos.org/10.1371/journal.pone.0050798

(data values follow a normal distribution). For this approach we made a Shapiro test, followed by a Kruskal–Nemenyi test. Allowing pairwise comparisons for which $p < 0.05$ for acceptance of null hypothesis *i.e.*, there is not a statistically significant difference between the distributions in the data sets.

2.2.2 Structural fingerprints. The intra-library similarity was measured for the BRDs and HDAC data sets which are the main focus of this work. In addition, the inter-library similarity was analyzed between different epigenetic and reference collections. In order to compute intra- and inter-set similarity, three molecular fingerprints of different design were calculated with MOE, namely; molecular access system (MACCS) keys (166 bits), pharmacophore graph triangle (GpiDAPH3), and Typed Graph Distance (TGD). MACCS is a dictionary-based representation that matches pre-defined fragments from a list with the structure of the molecule; GpiDAPH3 is a fingerprint based on pharmacophoric features of a molecule by three point representation. TGD is also a pharmacophore-based fingerprint, but in this case it searches for two points.

Structural similarity was computed with the Tanimoto coefficient:[34,35]

$$T(a,b) = \frac{c}{a+b-c}$$

where a and b are the number of fragment bits corresponding to the i-th and j-th compounds and c is the number of fragments shared between i and j.

The intra-library similarity of a given data set with N compounds was measured as the distribution of the $N(N-1)/2$ pairwise similarity values. The inter-library similarity was analyzed by means of nearest-neighbor curves. These curves represent the distribution of the maximum similarity values of molecules in a test set with respect to the molecules in the reference set.[36] In this study, six data sets (*i.e.*, BRDs, HDAC and DNMT1, GRAS, 'Drugs' and 'Clinic') were used as reference and test sets, *e.g.*, they were compared to each other. The distribution of the intra- and inter-library similarity values was analyzed by means of cumulative distribution functions (CDF) generated with matplotlib.pyplot Python scripts.[37]

2.2.3 Molecular scaffolds. The molecular scaffolds or cyclic systems were generated by deleting the side chains from the molecules *i.e.*, removal of the vertex with degree one, with the program molecular equivalent indices (MEQI).[38,39] MEQI has been extensively used for the scaffold analysis of a number of compound databases.[38,40–42] Cyclic systems are part of the chemotypes defined in the methodology developed by Johnson and Xu. For each cyclic system a unique chemotype identifier o chemotype alpha-numeric code with five characters was assigned.[38,39] The resulting cyclic systems represent equivalent classes, that is, all molecules classified in a cyclic system do not fall into other chemotype class. In this study, cyclic systems were computed for all the curated datasets and the overlap between the scaffolds of the BRDs and HDACs datasets was analyzed by direct comparison of the chemotype codes and cyclic systems.

2.2.3.1 Scaffold diversity. The scaffold diversity of the BRDs and HDACs data sets was evaluated using cyclic systems recovery (CSR) curves. Briefly, a CSR curve measures the fraction of cyclic systems contained in a given fraction of the database. To generate this curve, the list of cyclic systems for each data set was ordered by frequency. Then, the fraction of cyclic systems was plotted on the X axis and the fraction of compounds containing cyclic systems was plotted on the Y axis. The CSR curve was characterized by the area under the curve (AUC) and the fraction of cyclic systems that contain 50% of the corresponding data set (F_{50}).[43] The development and application of CSR curves is discussed elsewhere.[43,44]

2.2.3.2 Chemotype enrichment. For the most frequent scaffolds of BRDs and HDACs, the proportion of active compounds in a given scaffold relative to the fraction of active compounds in the entire data set was analyzed. The chemotype enrichment analysis was based on the definition of scaffolds given by Johnson and Xu (*vide supra*). To measure scaffold enrichment we employed an established methodology.[45,46] The background or baseline activity of the corresponding data set was calculated using the equation:

$$\mathrm{Act}(C) = [C^*]/[C]$$

where $[C]$ is the total number of compounds, and $[C^*]$ is the total number of active compounds. For this analysis, a compound was defined as 'active' if the IC_{50} was lower than 1 μM.

The fraction of active compounds in a specific chemotype Act(C_λ) was calculated with the expression:

$$\mathrm{Act}(C_\lambda) = \left[C_\lambda^*\right]/\left[C_\lambda\right]$$

where $[C_\lambda]$ and $[C_\lambda^*]$ are the total number of compounds and active compounds, respectively, in the chemotype class λ.

The enrichment factor (EF) for chemotype λ was calculated with the equation:

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

Thus, EF(C_λ) measured the proportion of active molecules of a particular chemotype relative to the proportion of active compounds in the dataset. In this manner, the molecular scaffolds with the highest EF were flagged as the most attractive. To further differentiate the most attractive cyclic systems *i.e.*, molecular scaffolds with the highest frequency, chemotype enrichment plots were generated plotting the EF on the *X*-axis and the cyclic systems frequency on the *Y*-axis.[45] Chemotype enrichment plots have been used in the scaffold analysis of compound databases, including the scaffold analysis of DNMTis.[20,45,47]

2.2.4 Molecular complexity. Molecular complexity is an attractive parameter for entering into uncharted regions of chemical space. Lovering *et al.*[31,48] underline the relationship of complexity markers towards the successful development of new drugs. Also this has shown to improve selectivity, as a measure of saturated carbon atoms. Several measures of complexity have been reported including MW.[49–51] In this work we focused on metrics that are frequently used to characterize the complexity of compound collections so that the results can be directly compared with other reports. Carbon bond saturation was defined by fraction sp^3 (F-sp^3), where F-sp^3 = (number of sp^3 hybridized carbons divided by total carbon count).[31] Overall, a larger F-sp^3 value indicates that the molecule is more likely to have a 3D structure *i.e.* the structure would be less flat.[31,52] Stereochemical complexity was defined as the fraction of chiral carbon atoms;[12] F-chirality = (number of chiral carbon atoms divided by total carbon count). The fraction of chiral atoms and the fraction of sp^3 carbon atoms were computed with MOE and MayaChem Tools (http://www.mayachemtools.org). The distribution of both metrics was analyzed using box plots and summary statistics. The statistical analysis was generated with the PMCMR package.

3. Results and discussion

Table 1 summarizes the number of compounds before and after data curation. In total, we analyzed 2000 unique compounds tested against HDACs and 207 unique molecules tested against BRDs. Of note, in this work we did not differentiate between different isoforms of HDACs and BRDs; first we aimed to obtain a first overall assessment of the chemical space coverage and distribution of the compounds that have been tested with these epigenetic targets. Despite the fact that most of the HDACs inhibitors studied in this work are from class I, a follow-up study will be conducted to distinguish the chemical space of different classes.

The degree of activity of each set was explored taking as reference an IC_{50} value of 1 μM to define an 'active' compound. Following this heuristic criterion, a high percentage (79.5%) of the HDAC data set was composed of active compounds, whereas only 46.85% of the molecules in the BRDs dataset had $IC_{50} < 1$ μM. This result is in line with the amount of activity data accumulated to optimize the activity of HDACis as compared to the current development of BRDis.

3.1 Physicochemical properties

Fig. 2 shows box plots of the distribution of the six PCPs calculated for the epigenetic and reference data sets. Table S1 in the ESI† summarizes the statistics of each distribution for all data sets. BRDs and HDACs sets had similar distributions of HBA that are comparable to the distribution for the general screening collection ('General') and the commercial collection focused on epigenetic targets ('Epi-focused') (*i.e.* median value of 4). Overall, DNMTs had a slightly higher number of HBA while approved drugs, compounds in the clinic ('Clinic') and, more markedly GRAS, have fewer HBA. In a similar way, HDACs and DNMTs had slightly higher number of HBD (*i.e.* median of 3) compared to 'General' and 'Epi-focused' datasets that presented a median of 2 HBD, or to BRDs, 'Drugs', and 'Clinic' dataset with a median of 1 HBD. Similar trends where obtained for TPSA that is other measure associated with polarity. HDACs and DNMTs had higher values of TPSA than other reference collections.

Regarding Slog *P* as a measure of hydrophobicity, BRDs had, on average, the highest values across all the data sets. The Slog *P* of HDACs was comparable to the 'General' set and had the second highest values. As per compound flexibility, measured by RTB, BRDs presented a similar mean value compared to the other reference collections, including GRAS. Noteworthy, HDACs had a higher mean number of RTB. This is due to the presence of peptide molecules that have been largely explored as HDACis. Overall, HDACs presented the highest mean MW, nevertheless all the other dataset (with exception of GRAS) presented comparable values of MW.

Statistical analysis using the Nemenyi test (Table S3 in the ESI†) revealed that, overall, BRDs is similar to 'General' and 'Epi-focused'. This result could suggest that compounds tested for BRD inhibition came from generally screening collections commercially available. Of note, the 'Epi-focused' set is also commercially available (Table 1) and it was assembled from a generally screening library. The HDACs set also showed to be similar to the 'General' and 'Epi-focused' sets, and it has some degree of similarity to the PCP of the DNMT set (for example in terms of HBD, TPSA, and MW).

Statistical tests were used to determine whether there is a significant difference between 'active' compounds ($IC_{50} < 1$ μM) in the HDACs and BRDs and the entire sets. It was found that the most active compounds were not statistically different from the inactive compounds based on PCP (data not shown).

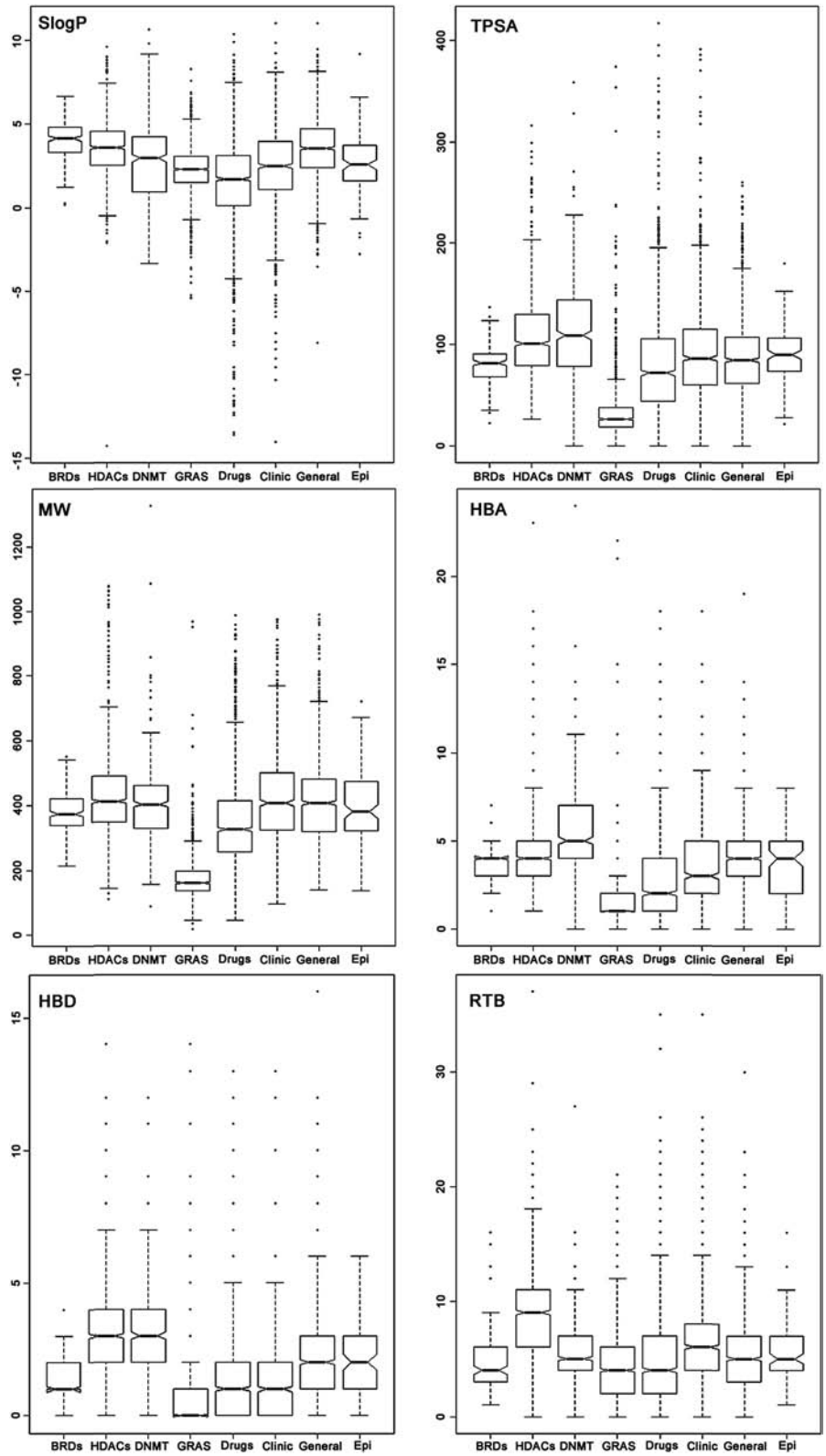


Fig. 2 Box plots of the distribution of six physicochemical properties of pharmaceutical relevance for the BRDs, HDACs, DNMTs and reference data sets. Summary statistics are in Table S1 of the ESI.†

3.1.1 Visual representation of the property space. Fig. 3 shows 2D and 3D visual representations of the chemical space of ERCS composed by BRDs, HDACs, and DNMTis. The relative position in the space is compared to other reference collections. As detailed in the Methods section the chemical space was obtained by PCA of the six PCP. For visual clarity, Fig. S1 in the

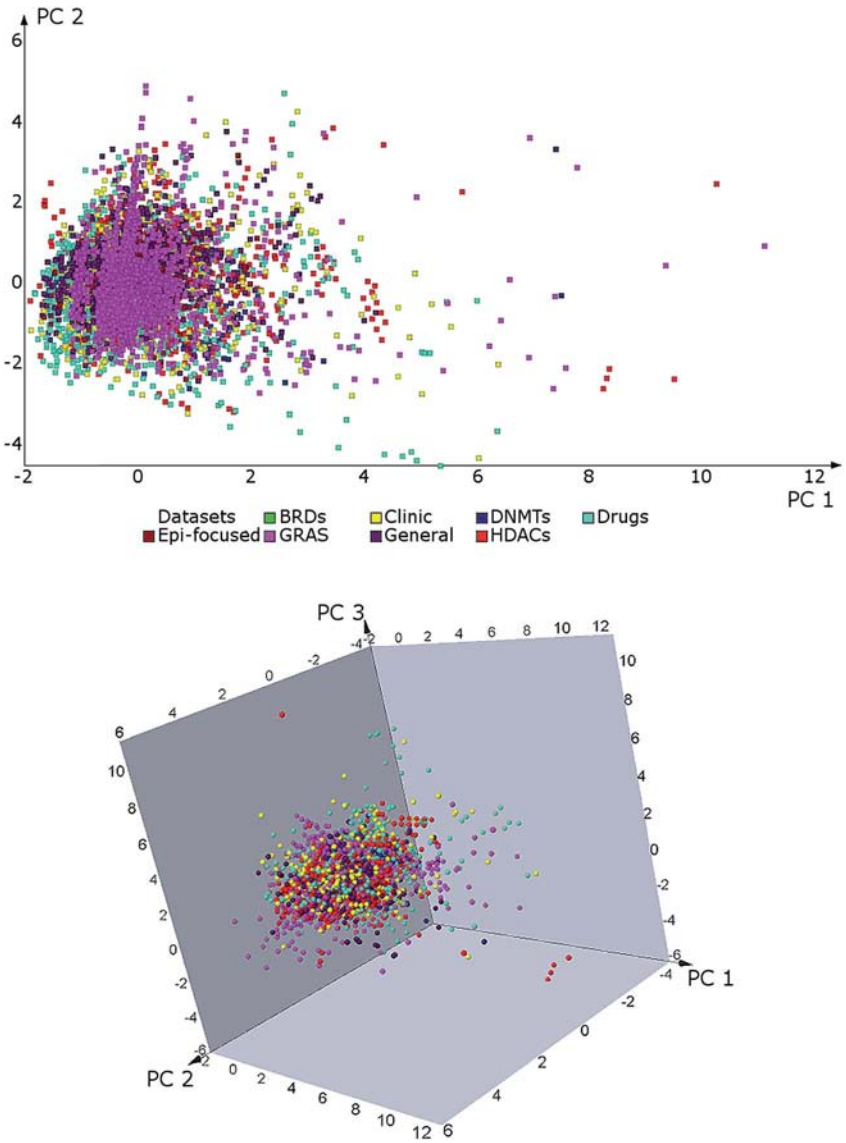


Fig. 3 2D and 3D visual representations of the chemical space of BRDis, HDACis, DNMTis, GRAS and reference data sets. The plots were generated with principal component (PC) analysis of six physicochemical properties of pharmaceutical relevance. The first two PC recover 82% of the variance and the first three, 90%. Data sets are shown separately in Fig. S1 (ESI†). Outliers in HDACs and GRAS sets are not shown for clarity. See main text for details.

ESI† shows the individual position in the chemical space of the BRDs, HDACs and GRAS sets in the same reference coordinates of Fig. 3. The first two principal components (PC) retrieved 82.2% of the variance and the first three PCs recovered 90.3% of the variance. Table S4 in the ESI† summarizes the loadings for the first three PCs of the property space of the eight data sets. HBA and HBD had the higher contributions to PC1 while Slog *P* had the highest contribution to PC2. Of note, all these properties are associated with compound polarity.

Fig. 3 shows that BRDs and HDACs cover similar regions of the property space of DNMTs, 'Drugs', and the other reference collections. In particular, BRDs cover the smallest region of the property space (Fig. S1†), followed by 'Epi-focused', which is nearby in the ERCS. This may be related not only to the fewer number of compounds in these sets (207 compounds and 113, respectively), but also to the type of chemical structures. Statistical analysis (Nemenyi values shown on Table S3†) confirmed that BRDs and 'Epi-focused' are significantly similar in properties to each other. Compounds tested for HDAC inhibition show a broader distribution on the property space. The outliers of the HDACs set are mainly associated with peptides. Overall, this type of compounds are more flexible and more polar than small molecules used in drug discovery and may have, depending on the nature of the peptide, a large MW.

Analysis of the property distribution and visual representation of the chemical space of GRAS revealed remarkable trends. While GRAS shares common regions with the fraction of ERCS studied in this work, the coverage of property space of GRAS is scattered (Fig. 3 and S1†). Most of the outliers in GRAS are similar to the outliers in HDACs. This finding may be attributed

to the nature of food additives as there are flexible molecules as sugars or even peptide derivatives. It is noteworthy the similar distribution of Slog *P* values between GRAS and 'Epi-focused' (Fig. 2). It has been discussed that hydrophobicity as measured by log *P* is one of the most important PCP to develop bioactive compounds.[53] This novel finding that emerged from this comparison is related to the association between food chemicals and epigenetics.

Despite the fact that the analysis of the distribution of the PCP and visual representation of the chemical space based on properties of pharmaceutical relevance is important, they do not provide direct information on the nature of the chemical structures of the epigenetic sets. Structural aspects of the epigenetic-related compounds are addressed in the next section.

3.2 Structural fingerprints

Structural fingerprints enabled the comparison of the epigenetic data sets considering the atom connectivity and chemical structures. As detailed in the Methods section, three fingerprints of different design were employed, namely MACCS keys (166 bits), GpiDAPH3, and TGD. These representations have been used in the chemoinformatic characterization of the structural diversity of a number of compound libraries, including compounds tested with DNMT1 activity.[20,49,54] Fingerprint-based representation enabled an analysis of the intra- and inter-compound set similarity that are discussed in the next two sub-sections.

3.2.1 Intra-compound set similarity. For the BRDs, HDACs and each reference data set the intra-compound set similarity was assessed calculating all pairwise structural comparisons using the Tanimoto coefficient and three different fingerprints. Details are summarized in the Methods section. Fig. 4 shows the distribution of the similarity values for each data set using CDFs. Table S5 in the ESI† summarizes the statistics of each data set for the different fingerprint representations. Overall, the magnitude of the similarity values computed with GpiDAPH3 and TGD fingerprints for BRDs, HDACs and reference data sets were different from the values computed with MACCS keys. That is, for a given data set, the relative order of the similarity values decreased in the order TGD > MACCS > GpiDAPH3 (Table S5†). The relative order of the similarity values has been noticed for many other data sets.[49,54] This result has been attributed to the different design of the fingerprints (detailed in the Methods section).

According to MACCS keys, BRDs was the data set with the highest intra-set similarity followed by HDACs *i.e.*, median similarity values of 0.463 and 0.423, respectively. In other words, these two collections had the lowest structural diversity as compared to all other reference data sets including DNMTs and the 'Epi-focused' collections. For reference, the most diverse sets were GRAS followed by 'Drugs'; both sets had the lowest distribution of MACCS keys/Tanimoto similarity values (*i.e.*, median values of 0.261 and 0.308, respectively). The high structural diversity of approved drugs has been reported[49] and is in line with the fact that approved drugs in DrugBank cover a wide number of molecular targets and mechanisms of action. Interestingly, compounds in the 'Clinic' are less diverse than 'Drugs' and the structural diversity of 'Clinic' is comparable to the diversity of the general screening collection, 'General' (Table S5†). Similar to the conclusions obtained with MACCS keys, BRDs is one of the most similar (less diverse) sets considering GpiDAPH3 and TGD representations. Interestingly, according to GpiDAPH3, BRDs have comparable diversity to 'Drugs'.

The structural diversity of HDACs highly depended on the molecular representation (Fig. 4 and Table S5†). According to

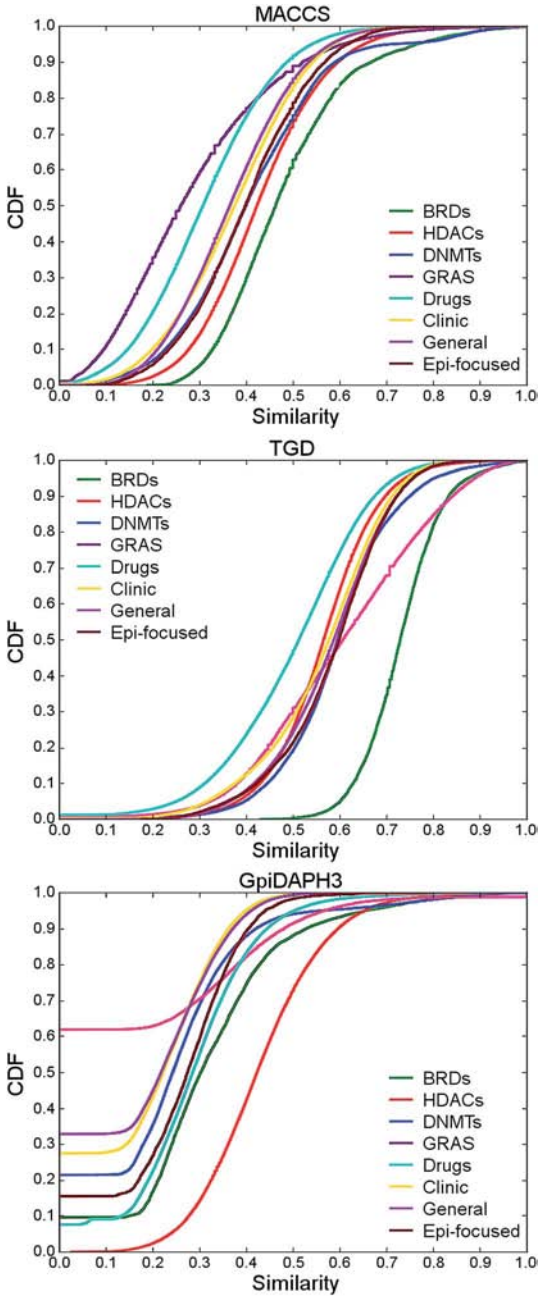


Fig. 4 Intra-library similarity: cumulative distribution function (CDF) of the pairwise similarity values for BRDs, HDACs and reference data sets using MACCS keys, GpiDAPH3, and TGD fingerprints. The statistics of each distribution is summarized in Table S5 of the ESI.†

GpiDAPH3, the HDACs set is the most similar (less diverse). But considering MACCS keys, HDACs is one of the most diverse sets with similarity comparable to 'Clinic'.

Of note, there was not a general relationship between the size of the data sets with structural diversity. It could be anticipated that smaller data sets are less diverse than bigger ones. For instance, the low diversity of BRDs may be associated with the fewer number of molecules (207, see Table 1). Note, however, that 'Epi-focused' has even fewer molecules than BRDs (113) but is more diverse. This result can be rationalized considering that 'Epi-focused' was designed considering several epigenetic targets. A second example is the lower diversity of HDACs as compared to 'Drugs', 'Clinic', and 'Epi-focused' (Fig. 4) despite the fact that the size of the HDACs set (2000 compounds) is bigger than the size of the reference sets (1, 490, 837, and 113 compounds, respectively).

Taking together the results of the molecular diversity using different fingerprints it can be concluded that the BRDs set analyzed in this work does not have a large structural diversity. This result is related to the fact that research groups are developing derivatives of specific type of compounds such as benzodiazepines for BRD inhibition.[11] These results clearly indicated the need to develop new chemical structures as BRDs. This far, compounds cover a limited region of chemical space and there is a large opportunity of increase novelty. The higher structural diversity of HDACs, as compared to BRDs, can be expected from the different type of compounds that have been reported as HDACis. However, one of the most interesting and active compounds are the hydroxamic acid derivatives. For these compounds, specific structural features change but keeping the same pharmacophoric features.[55] This fact can be associated with the fact that HDACs have, overall, higher GpiDAPH3/Tanimoto similarity values but lower MACCS/Tanimoto similarity values (*e.g.*, MACCS keys is more 'sensitive' to the changes in chemical modifications). The relative high diversity of GRAS compounds using different representation is also worth noting. The intra-set similarity results also highlighted the convenience of using multiple molecular representations for a comprehensive assessment of the molecular diversity of compound data sets.[56]

3.2.2 Inter-compound set similarity. In this section of the study we aimed to measure the overlap between ERCS, *i.e.*, the three epigenetic data sets considered in this work: BRDs, HDACs, and DNMTs. As an approach to measure the overlap, we computed the maximum structural similarity of a test collection with all compounds in the reference collection. The distribution of the maximum similarity values were analyzed using CDFs. For this analysis we focused on MACCS keys and TGD as representative structural fingerprints. In this analysis we did not employ GpiDAPH3 because of the relative large number of compounds identified with similarity values of zero in the previous analysis (Fig. 4).

Fig. 5 shows CDF plots of the maximum similarity of five test sets with BRDs, HDACs, and DNMTs as reference sets using MACCS keys and TGDs. The CDFs for other reference sets are shown in Fig. S2 in the ESI.† Summary statistics from the distributions using MACCS keys and TGD are presented in Tables S6 and S7 (ESI†), respectively.

The low values in the CDFs and statistics obtained with MACCS keys and TGD indicated that the epigenetic sets and the reference collections analyzed in this section have, in general, compounds with different chemical structures as compared to BRDs, HDACs and DNMTs. However, the distribution of maximum similarity values showed that there are compounds in the epigenetic-related data sets with similarity value of one. After inspection of pairs of compounds that present MACCS keys and TGD similarity value of one, we found that although there are not identical structures, they share similar motifs. For instance, the presence of biphenyl derivatives with hydroxamic acid moiety on both sets is noteworthy. In fact, a dual active HDAC/bromodomain and extra terminal (BET) small molecule tool inhibitor with a hydroxamic acid has been published.[57] Other moiety shared by the two sets is the phenylsulfonamide.

The CDFs and statistics also showed that 'Drugs' and 'Clinic' are, on average, more similar to BRDs, HDACs and DNMTs than the similarity showed among the three epigenetic sets of compounds. These results further highlights that, in general, the chemical structures tested as inhibitors of BRDs, HDACs and DNMTs, are different.

According to MACCS keys, the relative order or maximum similarity values of GRAS to the epigenetic sets is DNMTs > HDACs > BRDs. In other words, in similarity searching using MACCS keys, it is more likely to identify GRAS molecules similar to DNMTs. However, comparing GRAS to the epigenetic sets using TGD fingerprints, the relative order of maximum similarity values is different: BRDs > HDACs > DNMTs. Taken together, these results suggest that in similarity searching, more than one molecular representation should be employed and then select consensus hits. A detailed discussion of the comparison of all data sets studied in this work (Table 1) with each other is beyond the scope of this work that is focused on BRDs, HDACs and DNMTs.

3.3 Molecular scaffolds

The analysis of the molecular scaffolds is organized in three major groups: scaffold content, diversity, and activity distribution.

3.3.1 Scaffold content. Fig. 6 shows the most frequent cyclic systems (as defined in the methodology section) of the BRDs and HDACs sets. Scaffolds with frequency of at least 10 (BRDs) and 20 (HDACs) molecules are shown (the difference is due to set size). The benzene ring (cyclic system RYLFV) is not included in the figure (*vide infra*). The most frequent scaffolds of DNMTs and the other reference data sets have been published elsewhere.[20]

The most frequent scaffold in the BRDs set (cyclic system 1AWRP) had a frequency of 14 (6.8%) compounds followed by the cyclic system 49ZJ3 with 12 (5.8%) molecules. Most of the frequent cyclic systems had between 3 and 4 rings (Fig. 6). In contrast, the most populated cyclic system as defined by MEQI for HDACs (41 compounds, 2%) had only one ring. Interestingly this cyclic system is the benzimidazole ring which is a sub-

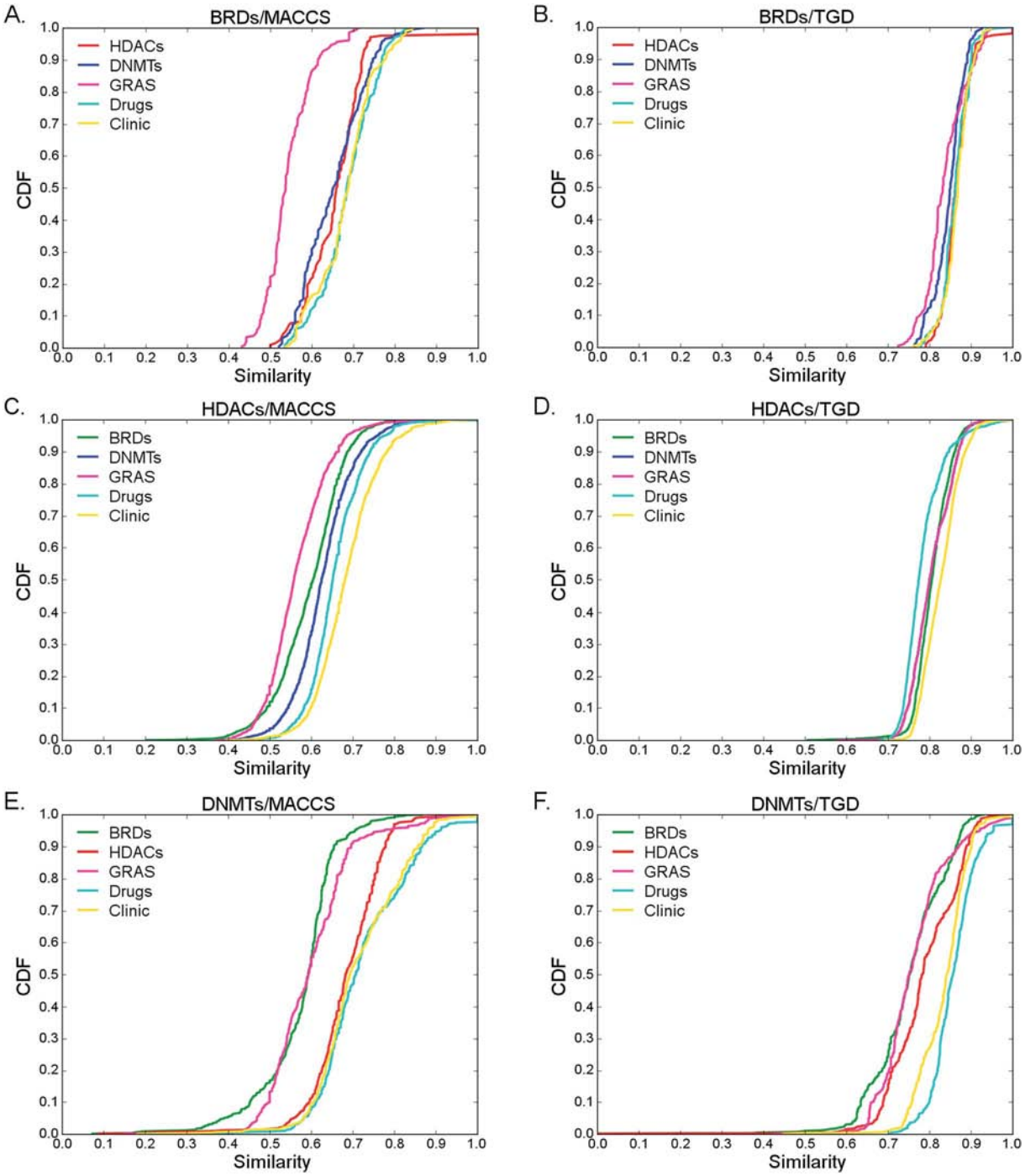


Fig. 5 Inter-library similarity: cumulative distribution function (CDF) of the maximum structure similarity calculated with MACCS keys, TDG and the Tanimoto coefficient comparing three epigenetic data sets, 'Drugs' and 'Clinic' with BRDs, HDACs and DNMTs. The reference set is indicated at the top of each graph. Summary statistics of the CDFs are presented in Tables S6 and S7 in the ESI.†

structure of the most frequent cyclic system in the BRDs set (1AWRP, Fig. 6). The high prevalence of benzimidazole ring in bioactive compounds is well documented.[54,58] Such findings increase the interest to design polyepigenetic drugs using the benzimidazole ring as a key sub-structure.

Not surprisingly, the benzene ring (cyclic system RYLFV) is highly frequent in the HDACs data set and had a frequency of 127 (6.35%) compounds. This cyclic system is highly frequent in approved drugs and several other compound collections.[41,54] Surprisingly, the benzene ring is not present as a cyclic system (*i.e.*, core scaffold) in the BRDs set although it is part of the structure.

Acyclic structures (chemotype identifier '00000') are amongst the most frequent structures in the HDACs set with 43 (2.15%)

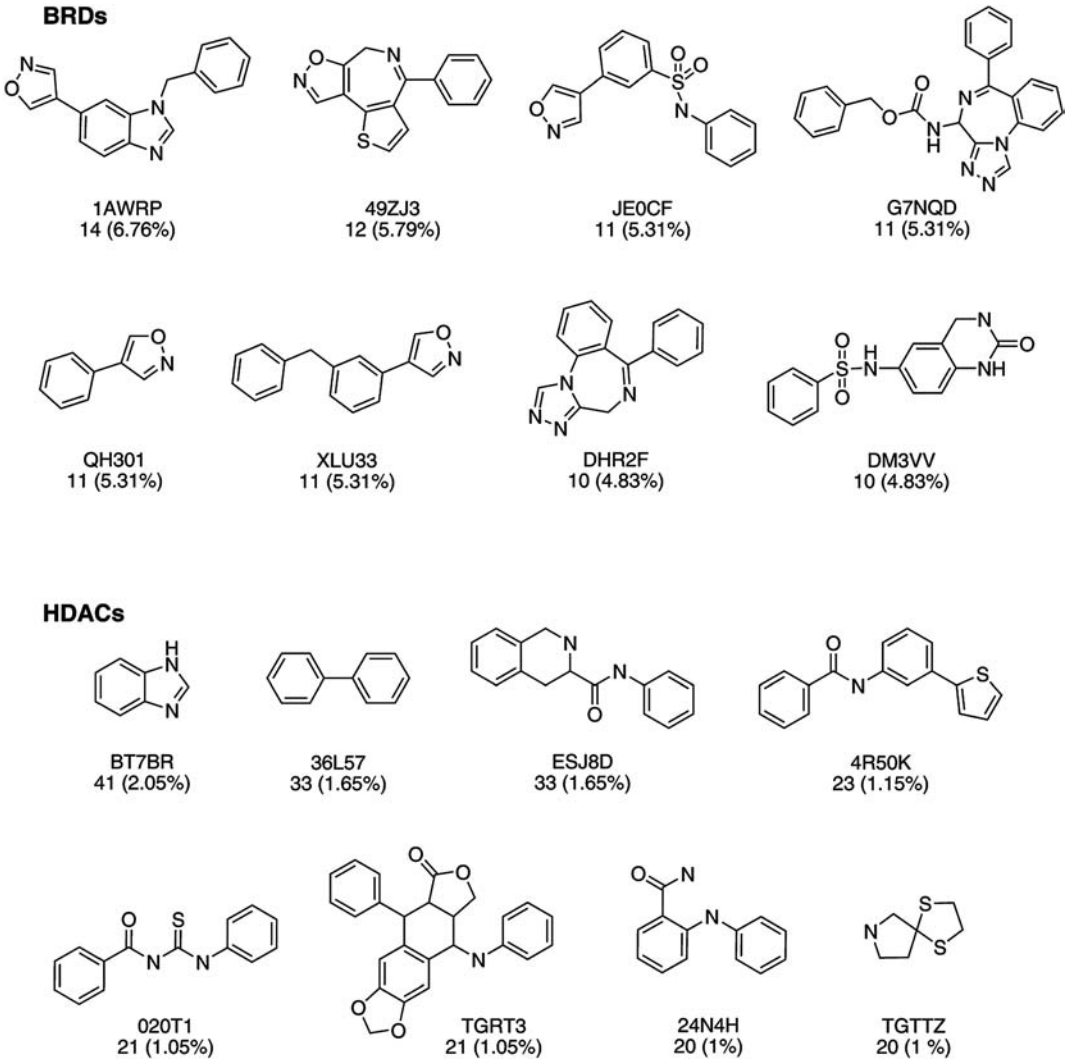


Fig. 6 Most frequent cyclic systems in the BRDs and HDACs sets with a frequency of at least 10 and 15 molecules, respectively. The benzene cyclic system is omitted. The chemotype identifier, number, and percentage of compounds for each cyclic system are shown. See text for details.

molecules. Similar to the benzene ring, the number of acyclic structures is also common in other data sets.[54] In contrast, no acyclic structures were found in the BRDs set.

The most frequent cyclic system identified in the BRDs and HDACS sets are not included in the 'not wanted' list and are not flagged as scaffold that have propensity to form multi-target activity cliffs.[59,60] However one of them (DM3VV) shows a PAINS-like moiety; it has a structure that may break down causing false positives results in biological assays (Fig. 6).[60]

3.3.2 Scaffold diversity. Fig. 7 shows the CSR curves for BRDs, HDACs and reference data sets. The corresponding summary statistics are summarized in Table S8 in the ESI.†

To measure the scaffold diversity with the CSR curves the following rationale was used. The fraction of a given cyclic system is compared to the fraction of the compounds in the data set contained in that group of cyclic systems. If we consider a reference 'most-diverse-set' *i.e.*, a set in which every molecule has its own scaffold, the CSR 'curve' is a diagonal line. As y equals x AUC is the integral:

$$\int_0^1 x\mathrm{d}x = \frac{x^2}{2} \therefore \mathrm{AUC} = 0.5$$

According to this expression, for the maximum diversity AUC = 0.5. It follows that as the AUC value increases (up to a maximum of one), the dataset contains higher number of compound with the same scaffold and the diversity of the data set is lower. Following this rationale, the CSR curves in Fig. 7 indicate that the scaffold diversity of the epigenetic and GRAS data sets decreases in the order: DNMTs > BRDs > HDACs > GRAS. Interestingly, according to the AUC metric, the scaffold diversity of the BRDs and HDACs (AUC = 0.74 and 0.76, respectively) is comparable to the diversity of compound tested as inhibitors of the androgen receptor, estrogen receptor agonist, glucocorticoid receptor, angiotensin-converting enzyme, and acethylcholine esterase that have reported AUC values between 0.74 and 0.76.[43] The scaffold diversity of the DNMT set is similar to compounds tested with

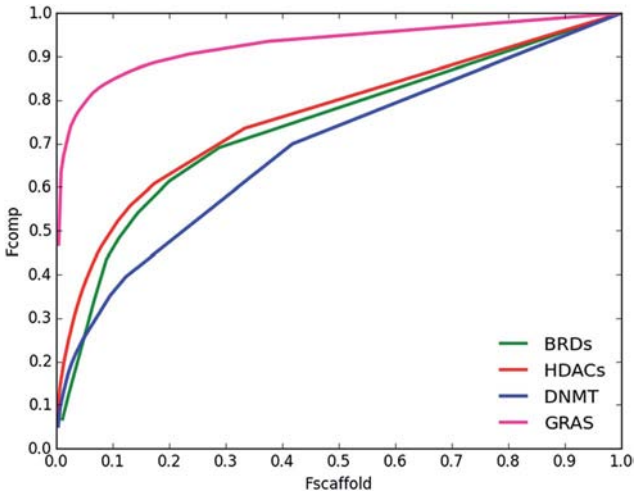


Fig. 7 Cyclic system recovery curves for the epigenetic-relevant data sets and GRAS compounds. Summary statistics are presented in Table S8 of the ESI.†

enoyl-(acyl-carrier-protein) reductase (AUC of 0.69 and 0.70, respectively).[43]

The same relative order of scaffold diversity was obtained with the F_{50} values (Table S8 in the ESI†). Half (50%) of the GRAS compounds were contained in 0.4% of the cyclic systems. In contrast, 50% of the compounds in the DNMTs set were distributed in 22% of the cyclic systems of this set. The F_{50} metric also indicated that the relative order of scaffold diversity of the epigenetic-related data sets is DNMT > BRDs > HDACs. The lower cyclic system diversity of the HDACs set can be influenced by research trends *e.g.*, hydroxamic acid derivatives.

It is noteworthy that the relative order of scaffold diversity does not necessarily match the scaffold diversity measured with structural fingerprints. As discussed elsewhere, the structural diversity evaluated with fingerprints consider the entire structure, including the own nature of the cyclic systems (*e.g.*, size, complexity) and the side chains. A clear example is the relative diversity of the GRAS set: it has a high structural diversity (MACCS keys/Tanimoto in Fig. 4) but it has lower scaffold diversity (CSR curve in Fig. 7) as compared to other data sets. Of note, this is the first study that addresses the scaffold diversity of GRAS.

3.3.3 Activity enrichment of scaffolds. After characterizing the scaffold content and diversity, it was explored whether there are cyclic systems with a larger proportion of active compounds as compared to the proportion of active molecules in the entire data set. To achieve this goal the EF was measured for the most frequent cyclic systems as detailed in the Method section. Fig. S3 in the ESI† shows the corresponding chemotype enrichment plots that represent the chemotype enrichment and frequency for the most frequent scaffolds. For the BRDs and HDCAs sets, all cyclic systems had an EF lower than the unit. This finding suggests that, for the data sets analyzed in this work, there are not cyclic systems (as defined by MEQI) with a significant high proportion of active molecules. For the BRDs, the four most frequent cyclic systems that had the highest values of enrichment factor were 1AWRP, DM3VV, 49ZJ3, and G7NQD (see chemical structures in Fig. 6). For the HDACs set, the cyclic system BT7BR (benzimidazole) had the highest enrichment factor and frequency.

3.4 Molecular complexity

The search for diverse datasets is usually made with the purpose of exploring 'empty voids' of chemical space. The universe of possible existing molecules is huge so the idea of entering uncharted regions may seem easy. What is more challenging is the successful navigation through novel but pharmaceutical relevant areas of chemical space.[61] Specific strategies to venture further in this chemical wilderness have been discussed recently. One of such strategies depends on assessing molecular complexity. To the best of our knowledge, there are not reports that measure the molecular complexity of epigenetic data sets.

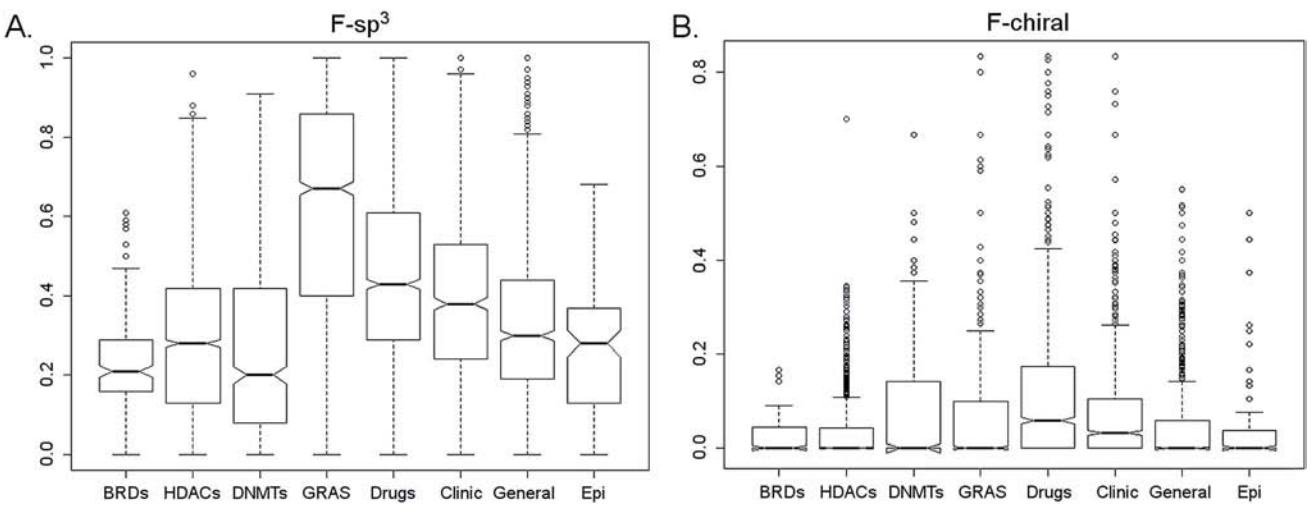


Fig. 8 Box plots of the distribution of the fraction of sp^3 carbon atoms (F-sp^3) and fraction of chiral centers (F-chiral) for the BRDs, HDACs, DNMTs and reference data sets. Summary statistics are in Table S9 of the ESI.†

Fig. 8 shows the distribution of the fraction of chiral and sp^3 carbon atoms for the BRDs, HDACs, DNMTs and other reference data sets. Summary statistics are presented in Table S9 in the ESI.† The distributions were compared with Nemenyi tests for pairing and analysis of raw data. Analysis of the results revealed that HDACs and 'Epi-focused' have similar distribution of F-sp^3 values suggesting that it is likely equal to find 3D structures (less flat compounds) in both data sets. The overall lower F-sp^3 values of BRDs than HDACs (and other reference sets, except DNMTs), indicate that the structures of BRDs currently contained in ChEMBL are, in general, more flat.

BRDs and HDACs had comparable distributions of F-chiral values indicating similar stereochemical complexity. Moreover, their distribution of F-chiral values was also comparable to 'General' and 'Epi-focused'. DNMTs showed broader range of F-chiral values.

In agreement with previous analyses, 'Drugs' had overall, higher F-chiral and F-sp^3 values than commercial screening library 'General'. Notably, GRAS compounds had even higher F-sp^3 values than 'Drugs'. These results suggest that it is more likely that GRAS compounds have 3D structures as compared to approved drugs and any other data sets analyzed in this work. Indeed, it is known that the activity of several approved drugs are stereospecific as only one enantiomer is active, while the other may not be active or even toxic. GRAS on the other hand, contains flavor molecules: many sugars and sweeteners are optically active. Also flavors may show "property cliffs", that is, flavor may drastically change with small structure changes (*i.e.*, limonene gives both their flavors to lemons and tangerines, the difference is just a chiral center).[62]

Taken together, the results of this analysis suggested that the compounds currently tested as inhibitors of BRDs, HDACs and DNMTs have comparable or less stereochemical complexity, and are less flat than currently approved drugs.

4. Conclusions

Epigenetics is an emerging therapeutic approach and has proven to be a very important aspect of our daily lives.[63,64] As such, it is quite relevant to understand the chemical space currently covered. As part of our research program to further advance epi-drug and epi-drug discovery using computational methods,[17,65] herein we discuss for the first time the characterization of the chemical space of BRDis and HDACis using structure–activity data available in two major public databases: ChEMBL and Binding Database. It was concluded that the chemical space of data sets with activity against HDACs and BRDs showed a significant overlap with the property space of drugs approved for clinical use, compounds in clinical development and commercial screening libraries. Complementary measures of structural diversity revealed that there is a need to develop and test more diverse compounds as BRDis.

Based on the diversity analysis of the screening data analyzed of HDACis, BRDis and DNMTis it was found that the development of poly-epigenetic drugs has not been extensively explored. However, it was shown the feasibility of develop at least dual epigenetic inhibitors. This conclusion has been supported experimentally. These observations open up a unique avenue to develop potentially more efficient epigenetic therapies of compounds targeting multiple epigenetic targets.

Surprisingly, despite the fact that several different chemical scaffolds have been explored for BRDs and HDACs, there is not yet a unique molecular scaffold (as defined by Johnson and Xu) with a large enrichment factor. These quantitative results highlight the need to continue increasing the SAR of the most promising chemical scaffolds identified in this work.

From the quantitative analysis of the structural complexity it was concluded that, in general, the chemical structure of inhibitors of BRDs, HDACs, and DNMT have a limited structural complexity. These results encourage the development of new inhibitors with increased complexity that may lead to improved selectivity. Taking all this together, this work represents a significant contribution to further advance the understanding of the ERCS.

5. Perspectives

The first step in our ERCS odyssey was to analyze compounds targeted to several different types of BRDs (2, 3, 4) and HDACs (1, 2, 3, 5, 8 and SIRT1). The next logical step is to assess the fraction of ERCS of different BRDs and HDACs individually and expand the analysis to other public and more specialized repositories such as Chromohub.[66] As part of this analysis, an in-depth characterization of the overlap between ChEMBL/Binding Database with Chromohub can be made. Furthermore, it will be of high relevance to explore the chemical space of compounds with and without cellular activity, as well as to determine structure–selectivity relationships of HDACs, BRDs, and other epigenetic modulators.

During the course of this study it was found that GRAS compounds share a similar property space as the ERCS region; in particular, share similar hydrophobicity values which is one of the most important PCP related to bioactive compounds. These novel results support the systematic exploration of flavor chemicals with potential health benefits as epigenetic modulators. Furthermore, the chemoinformatic study uncovered that GRAS chemicals have a larger number of 3D (less flat) structures as compared to other general screening and 'Epi-focused' commercial collections. It is anticipated that GRAS chemicals may show selectivity towards epigenetic targets and may act as 'master key' epigenetic-compounds. Further experimental studies are required to assess this hypothesis.

List of abbreviations

BRDs	Bromodomains
DNA	Deoxyribonucleic acid
DNMT	DNA-methyl transferase
EF	Enrichment factor
ERCS	Epigenetic relevant chemical space
GRAS	Generally recognized as safe
HDACs	Histone lysine deacetylase
HBA	Hydrogen bond acceptor

HBD	Hydrogen bond donor
RTB	Number of rotatable bonds
Slog *P*	Partition coefficient water/octanol
MW	Molecular weight
PCP	Physicochemical properties
PC	Principal component
PCA	Principal component analysis
SMILES	Simplified molecular input line entry
SAM	*S*-Adenosyl-methionine
TPSA	Topological surface area

Acknowledgements

Fernando Prieto-Martínez and Eli Fernández-de Gortari are grateful to CONACyT for the fellowships granted No. 660465/576637 and 348291/240072, respectively. Authors thank Tatiana Enríquez Gómez and Jonathan Guerra Pérez for carefully proofreading the manuscript. This work was supported by the Universidad Nacional Autónoma de México (UNAM), grant PAPIIT IA204016. We also thank the institutional program Nuevas Alternativas de Tratamiento para Enfermedades Infecciosas (NUATEI) of the Instituto de Investigaciones Biomédicas (IIB) UNAM and Programa de Apoyo a la Investigación y el Posgrado (PAIP) 5000-9163, Facultad de Química, UNAM, for financial support.

References

1 D. C. Dolinoy, in *Comprehensive Toxicology*, ed. C. A. McQueen, Elsevier, Oxford, 2nd edn, 2010, pp. 293–309.
2 J. C. Jiménez-Chillarón, R. Díaz, M. Ramón-Krauel and S. Ribó, in *Transgenerational Epigenetics*, ed. T. Tollefsbol, Academic Press, Oxford, 2014, pp. 281–301.
3 N. Carey, *MedChemComm*, 2012, **3**, 162–166.
4 K. P. Nightingale, in *Epigenetics for Drug Discovery*, The Royal Society of Chemistry, 2016, pp. 1–19.
5 Q. Gao, J. Tang, J. Chen, L. Jiang, X. Zhu and Z. Xu, *Drug Discovery Today*, 2014, **19**, 1744–1750.
6 A. O. Arguelles, S. Meruvu, J. D. Bowman and M. Choudhury, *Drug Discovery Today*, 2016, **21**, 499–509.
7 A. Tarakhovsky, *Nat. Immunol.*, 2010, **11**, 565–568.
8 C. Robert and F. V. Rassool, in *Advances in Cancer Research*, ed. G. Steven, Academic Press, 2012, vol. 116, pp. 87–129.
9 S. Majumder and A. Advani, *Journal of Diabetes and its Complications*, 2015, **29**, 1337–1344.
10 F. L. Cherblanc, R. W. M. Davidson, P. Di Fruscia, N. Srimongkolpithak and M. J. Fuchter, *Nat. Prod. Rep.*, 2013, **30**, 605–624.
11 B. Pachaiyappan and P. M. Woster, *Bioorg. Med. Chem. Lett.*, 2014, **24**, 21–32.
12 C.-Y. Wang and P. Filippakopoulos, *Trends Biochem. Sci.*, 2015, **40**, 468–479.
13 S.-Y. Wu and C.-M. Chiang, *J. Biol. Chem.*, 2007, **282**, 13141–13145.
14 C. W. Lindsley, *ACS Chem. Neurosci.*, 2014, **5**, 1142.
15 J. L. Medina-Franco, M. A. Giulianotti, G. S. Welmaker and R. A. Houghten, *Drug Discovery Today*, 2013, **18**, 495–501.
16 G. Franci, M. Miceli and L. Altucci, *Epigenomics*, 2010, **2**, 731–742.
17 O. Méndez-Lucio, J. Tran, J. L. Medina-Franco, N. Meurice and M. Muller, *ChemMedChem*, 2014, **9**, 560–565.
18 J. L. Medina-Franco, K. Martinez-Mayorga and N. Meurice, *Expert Opin. Drug Discovery*, 2014, **9**, 151–165.
19 K. Martinez-Mayorga, T. L. Peppard, F. López-Vallejo, A. B. Yongye and J. L. Medina-Franco, *J. Agric. Food Chem.*, 2013, **61**, 7507–7514.
20 E. Fernández-de Gortari and J. L. Medina-Franco, *RSC Adv.*, 2015, **5**, 87465–87476.
21 J. L. Medina-Franco, *Epi-Informatics: Discovery and Development of Small Molecule Epigenetic Drugs and Probes*, Academic Press, London, UK, 2016.
22 X. Lucas, B. A. Grüning, S. Bleher and S. Günther, *J. Chem. Inf. Model.*, 2015, **55**, 915–924.
23 J. B. Baell, *J. Chem. Inf. Model.*, 2013, **53**, 39–55.
24 G. Papadatos and J. P. Overington, *Future Med. Chem.*, 2014, **6**, 361–364.
25 T. Q. Liu, Y. M. Lin, X. Wen, R. N. Jorissen and M. K. Gilson, *Nucleic Acids Res.*, 2007, **35**, D198–D201.
26 C. H. Arrowsmith, C. Bountra, P. V. Fish, K. Lee and M. Schapira, *Nat. Rev. Drug Discovery*, 2012, **11**, 384–400.
27 M. Wegner, D. Neddermann, M. Piorunska-Stolzmann and P. P. Jagodzinski, *Diabetes Res. Clin. Pract.*, 2014, **105**, 164–175.
28 *Molecular Operating Environment (MOE), version 2013*, Chemical Computing Group Inc., Montreal, Quebec, Canada, available at http://www.chemcomp.com.
29 D. Fourches, E. Muratov and A. Tropsha, *Nat. Chem. Biol.*, 2015, **11**, 535.
30 J. L. Medina-Franco, K. Martínez-Mayorga, T. L. Peppard and A. Del Rio, *PLoS One*, 2012, **7**, e50798.
31 F. Lovering, J. Bikker and C. Humblet, *J. Med. Chem.*, 2009, **52**, 6752–6756.
32 *R Development Core Team, R Foundation for Statistical Computing*, Vienna, Austria, 2011.
33 T. Sander, J. Freyss, M. von Korff and C. Rufener, *J. Chem. Inf. Model.*, 2015, **55**, 460–473.
34 P. Jaccard, *Bull. Soc. Vaudoise Sci. Nat.*, 1901, **37**, 547–579.
35 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.
36 J. L. Medina-Franco, G. M. Maggiora, M. A. Giulianotti, C. Pinilla and R. A. Houghten, *Chem. Biol. Drug Des.*, 2007, **70**, 393–412.
37 M. P. scripts, Matplotlib Python scripts, available at http://dx.doi.org/10.5281/zenodo.15423.
38 Y. J. Xu and M. Johnson, *J. Chem. Inf. Comput. Sci.*, 2002, **42**, 912–926.
39 Y. Xu and M. Johnson, *J. Chem. Inf. Comput. Sci.*, 2001, **41**, 181–185.
40 F. López-Vallejo, A. Nefzi, A. Bender, J. R. Owen, I. T. Nabney, R. A. Houghten and J. L. Medina-Franco, *Chem. Biol. Drug Des.*, 2011, **77**, 328–342.

41 A. B. Yongye, J. Waddell and J. L. Medina-Franco, *Chem. Biol. Drug Des.*, 2012, **80**, 717–724.
42 A. B. Yongye and J. L. Medina-Franco, *Chem. Biol. Drug Des.*, 2013, **82**, 367–375.
43 J. L. Medina-Franco, K. Martínez-Mayorga, A. Bender and T. Scior, *QSAR Comb. Sci.*, 2009, **28**, 1551–1560.
44 A. H. Lipkus, Q. Yuan, K. A. Lucas, S. A. Funk, W. F. Bartelt, R. J. Schenck and A. J. Trippe, *J. Org. Chem.*, 2008, **73**, 4443–4451.
45 J. L. Medina-Franco, J. Petit and G. M. Maggiora, *Chem. Biol. Drug Des.*, 2006, **67**, 395–408.
46 J. Pérez-Villanueva, O. Méndez-Lucio, O. Soria-Arteche and J. Medina-Franco, *Mol. Diversity*, 2015, **19**, 1021–1035.
47 J. Pérez-Villanueva, J. L. Medina-Franco, O. Méndez-Lucio, J. Yoo, O. Soria-Arteche, T. Izquierdo, M. C. Lozada and R. Castillo, *Chem. Biol. Drug Des.*, 2012, **80**, 752–762.
48 F. Lovering, *MedChemComm*, 2013, **4**, 515–519.
49 F. López-Vallejo, M. A. Giulianotti, R. A. Houghten and J. L. Medina-Franco, *Drug Discovery Today*, 2012, **17**, 718–726.
50 S. H. Bertz, *J. Am. Chem. Soc.*, 1981, **103**, 3599–3601.
51 R. Barone and M. Chanon, *J. Chem. Inf. Comput. Sci.*, 2001, **41**, 269–272.
52 H. Chen, O. Engkvist, N. Blomberg and J. Li, *MedChemComm*, 2012, **3**, 312–321.
53 A. Ganesan, *Curr. Opin. Chem. Biol.*, 2008, **12**, 306–317.
54 N. Singh, R. Guha, M. A. Giulianotti, C. Pinilla, R. A. Houghten and J. L. Medina-Franco, *J. Chem. Inf. Model.*, 2009, **49**, 1010–1024.
55 Y. Chen, H. Li, W. Tang, C. Zhu, Y. Jiang, J. Zou, Q. Yu and Q. You, *Eur. J. Med. Chem.*, 2009, **44**, 2868–2876.
56 A. M. Wassermann, E. Lounkine, J. W. Davies, M. Glick and L. M. Camargo, *Drug Discovery Today*, 2015, **20**, 422–434.
57 S. J. Atkinson, P. E. Soden, D. C. Angell, M. Bantscheff, C.-w. Chung, K. A. Giblin, N. Smithers, R. C. Furze, L. Gordon, G. Drewes, I. Rioja, J. Witherington, N. J. Parr and R. K. Prinjha, *MedChemComm*, 2014, **5**, 342–351.
58 S. R. Langdon, N. Brown and J. Blagg, *J. Chem. Inf. Model.*, 2011, **51**, 2174–2185.
59 Y. Hu and J. Bajorath, *J. Chem. Inf. Model.*, 2010, **50**, 500–510.
60 J. Baell and M. A. Walters, *Nature*, 2014, **513**, 481–483.
61 T. I. Oprea and J. Gottfries, *J. Comb. Chem.*, 2001, **3**, 157–166.
62 J. L. Medina-Franco, G. Navarrete-Vázquez and O. Méndez-Lucio, *Future Med. Chem.*, 2015, **7**, 1197–1211.
63 I. R. Miousse, R. Currie, K. Datta, H. Ellinger-Ziegelbauer, J. E. French, A. H. Harrill, I. Koturbash, M. Lawton, D. Mann, R. R. Meehan, J. G. Moggs, R. O'Lone, R. J. Rasoulpour, R. A. R. Pera and K. Thompson, *Toxicology*, 2015, **335**, 11–19.
64 M. Szyf, *Eur. Neuropsychopharmacol.*, 2015, **25**, 682–702.
65 J. L. Medina-Franco, O. Méndez-Lucio, J. Yoo and A. Dueñas, *Drug Discovery Today*, 2015, **20**, 569–577.
66 L. Liu, X. T. Zhen, E. Denton, B. D. Marsden and M. Schapira, *Bioinformatics*, 2012, **28**, 2205–2206.

Electronic Supplementary Material (ESI) for RSC Advances.

SUPPORTING INFORMATION

A Chemical Space Odyssey of Inhibitors of Histone Deacetylases and Bromodomains

Fernando D. Prieto-Martínez, Eli Fernández-de Gortari, Oscar Méndez-Lucio, José L. Medina-Franco

Facultad de Química, Departamento de Farmacia, Universidad Nacional Autónoma de México, Avenida Universidad 3000, Mexico City 04510, México

Contents **Page**

Table S1. Summary statistics of the distribution of the PCP for the data sets S2

Table S2. Distribution of 'active' compounds and comparison for BRDi compounds S4

Figure S1. Chemical space representation of BRDi, HDACi, DNMT 1, 'Epi-focused', and GRAS S6

Table S3. Nemenyi test values for the PCPs comparisons of eight data sets S7

Table S4. Loadings for the first three principal components of the property space S9

Table S5. Summary statistics of the intra-compound set similarity S10

Figure S2. Inter-library similarity S11

Table S6. Summary statistics of the inter-library comparison using MACCS keys S12

Table S7. Summary statistics of the inter-library comparison using TGD fingerprints S14

Table S8. Descriptive statistics and AUC for CSR curves S16

Figure S3. Chemotype enrichment plots for BRDi and HDACi data sets S17

Table S9. Descriptive statistics for F-chiral and F-sp^3 distributions S18

Table S10. Nemenyi pairing values for F-chiral and F-sp3 distributions S19

Table S1. Summary statistics of the distribution of the PCP for the data sets studied in this work

HBA

	BRDi	HDACi	DNMT 1	GRAS	Drugs	Clinic	Inhibit	Epi-focused
Min	1	1	0	0	0	0	0	0
1st Quart	3	3	4	1	1	2	3	2
Median	4	4	5	1	2	3	4	4
Mean	3.662	4.813	5.357	1	2.919	3.585	4.459	3.416
3rd Quart	4	5	7	2	4	5	5	5
Max	7	37	24	22	18	18	19	8
SD	0.991	2.45	2.5	1.31	2.56	2.37	2.32	1.83

HBD

	BRDi	HDACi	DNMT 1	GRAS	Drugs	Clinic	General	Epi-focused
Min	0	0	0	0	0	0	0	0
1st Quart	1	2	2	0	0	0	1	1
Median	1	3	3	0	1	1	2	2
Mean	1.184	3.265	3.133	0.4666	1.35	1.525	2.298	2.298
3rd Quart	2	4	4	1	2	2	3	3
Max	4	24	12	14	13	13	16	16
SD	0.867	1.511	2.011	1.077	1.559	1.552	1.817	1.294

RTB

	BRDi	HDACi	DNMT 1	GRAS	Drugs	Clinic	General	Epi-focused
Min	1	0	0	0	0	0	0	1
1st Quart	3	6	4	2	2	4	3	4
Median	4	9	5	4	4	6	5	5
Mean	4.734	9.32	5.568	4.291	5.008	6.303	5.433	5.549
3rd Quart	6	11	7	6	7	8	7	7
Max	16	53	27	21	35	35	30	16
SD	2.05	4.03	2.931	3.345	3.779	4.186	3.342	2.84

SLogP

	BRDi	HDACi	DNMT 1	GRAS	Drugs	Clinic	General	Epi-focused
Min	0.183	-14.261	-3.3655	-5.422	-13.6094	-14.04	-8.113	-2.796
1st Quart	3.303	2.559	0.9547	1.519	0.1402	1.089	2.248	1.619
Median	4.161	3.613	2.9577	2.317	1.7112	2.499	3.545	2.588
Mean	4.059	3.594	2.6556	2.317	1.4113	2.274	3.49	2.627
3rd Quart	4.849	4.599	4.2702	3.051	3.0996	3.963	4.729	3.724
Max	6.701	9.65	10.685	8.312	10.4033	11.059	11.059	9.189
SD	1.172	1.69	2.394	1.331	2.91	2.772	1.92	1.715

Table S1. (Continued)

TPSA

	BRDi	HDACi	DNMT 1	GRAS	Drugs	Clinic	General	Epi-focused
Min	22	26.3	0	0	0	0	0	21.51
1st Quart	67.56	78.43	78.11	18.46	43.37	60.11	61.32	73.25
Median	81.4	101.21	108.65	26.3	71.36	86.19	84.42	89.54
Mean	78.54	110.4	111.7	31.29	82.29	93.97	88.29	91.28
3rd Quart	90.63	129.98	143.87	37.3	105.15	115.15	107.29	105.82
Max	137.07	612.25	358.2	374.13	417.27	391.12	260.04	180.04
SD	19.54	42.897	47.102	24.611	55.74	52.316	40.715	27.983

MW

	BRDi	HDACi	DNMT 1	GRAS	Drugs	Clinic	General	Epi-focused
Min	213.2	111.1	88.11	18.04	43.33	95.12	136.1	134.2
1st Quart	337.8	349.9	329.83	134.18	254.32	323.15	319.2	321.4
Median	374.4	413.5	403.41	160.25	325.41	407.45	408.4	380.4
Mean	378.7	441.6	402.8	169.9	348.4	418.47	411.6	392.5
3rd Quart	423.5	492.8	462.03	196.29	415.51	501.54	480.1	473.6
Max	552.7	1598.8	1326.78	967.02	988.18	975.64	990.2	720.9
SD	64.229	133.525	118.158	61.857	146.822	148.614	129.24	110.161

Table S2. Distribution of 'actives' compounds and comparison for BRDi compounds

Median	BRD2	BRD3	BRD4	IC_{50} (μM)
	4	4	4	1
	4	4	4	10
HBA	4	4	3	≥100
	1	1	1	1
	1	1	1	10
HBD	1	1	1	≥100
	4	4	4	1
	4	4	4	10
RTB	4	4	4	≥100
	4.71362	4.71792	4.54183	1
	4.01176	4.07626	4.2211	10
SLogP	3.782135	3.92028	4.02703	≥100
	81.4	81.4	81.4	1
	81.4	81.4	81.4	10
TPSA	81.4	81.4	78.51	≥100
	415.453	410.441	402.96	1
	376.864	378.493	376.391	10
MW	372.4445	376.864	364.96	≥100

Mean	BRD2	BRD3	BRD4	IC_{50} (μM)
	4	4	3.865854	1
	3.671233	3.821918	3.663317	10
HBA	3.684783	3.829268	3.463115	≥100
	1.1333	1.15625	1.268293	1
	1.041096	1.09589	1.201005	10
HBD	1.065217	1.109756	1.17623	≥100
	4.4	4.1875	4.304878	1
	4.260274	4.39726	4.251256	10
RTB	4.206522	4.353659	3.979508	≥100
	4.735897	4.529324	4.354318	1
	3.986589	3.989324	4.101408	10
SLogP	3.767619	3.820534	3.847727	≥100
	84.46933	83.95625	83.88902	1
	78.75466	82.3137	78.82246	10
TPSA	79.08337	82.40805	75.63689	≥100
	405.6323	398.6574	401.2273	1
	369.7882	383.3375	380.6428	10
MW	365.7007	379.908	361.3758	≥100

Table S2. (Continued)

SD	BRD2	BRD3	BRD4	IC_{50} (μM)
	0.755929	0.718421	1.015463	1
	0.834246	0.733001	0.990975	10
HBA	0.824479	0.699292	1.127049	≥100
	0.743223	0.677251	0.943448	1
	0.610973	0.581291	0.870272	10
HBD	0.625543	0.588073	0.878228	≥100
	0.736788	0.737804	1.95435	1
	1.302001	1.198839	1.704772	10
RTB	1.338673	1.260726	1.849143	≥100
	0.938424	0.915115	0.988773	1
	1.065457	1.118183	1.149915	10
SLogP	1.144782	1.198718	1.291483	≥100
	12.3812	10.72139	19.0008	1
	15.2	12.97325	19.49417	10
TPSA	15.8889	12.55546	20.81594	≥100
	34.6769	39.93311	60.62833	1
	62.87314	49.93638	63.05745	10
MW	63.4056	48.96492	78.06248	≥100

*Note: Statistical comparison towards “active” compounds of HDACi data set was not made because of the following:

Of the 2000 unique compounds; 1135 show $IC_{50} \leq 1$ μM.

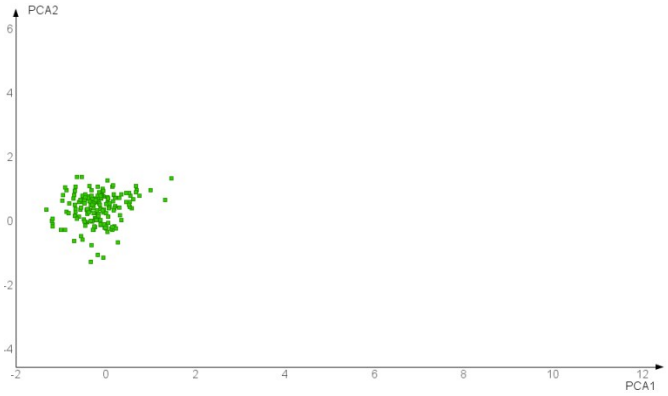
The difference in size makes it difficult to make a good comparison on the same statistic grounds.

It is true that many types of structures have been proposed as HDACs inhibitors; however the most interesting and potent ones are inspired or derived from hydroxamic acid moiety.

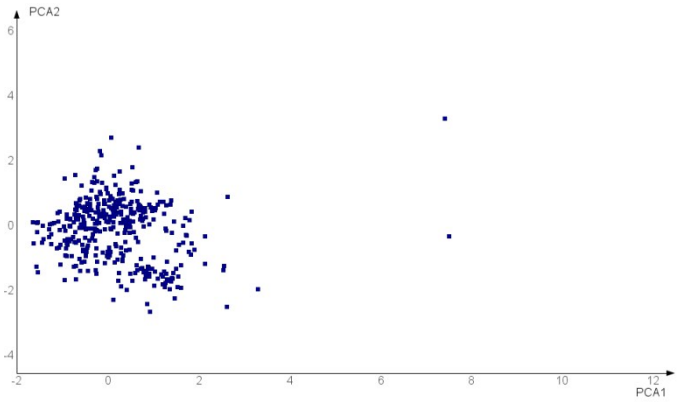
That last fact has produced intensive research of those compounds; therefore a high chance of statistical similarities is plausible.

Finally the pharmacophoric features of such compounds are well understood as of today, making good reference for our probe search.

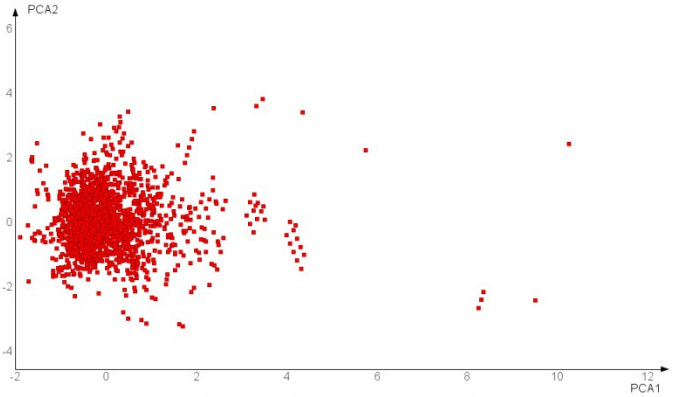
BRDi



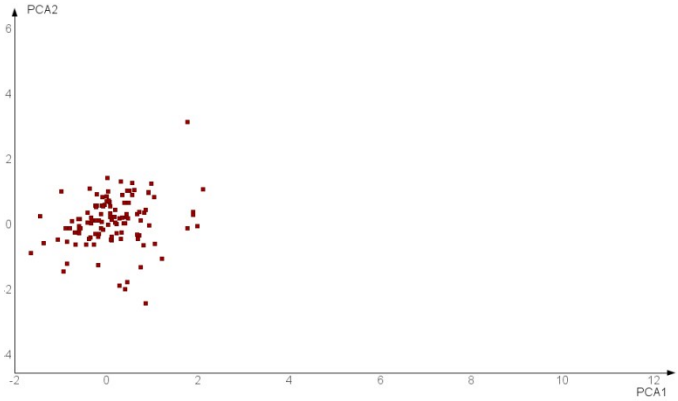
DNMT 1



HDACi



'Epi-focused'



GRAS

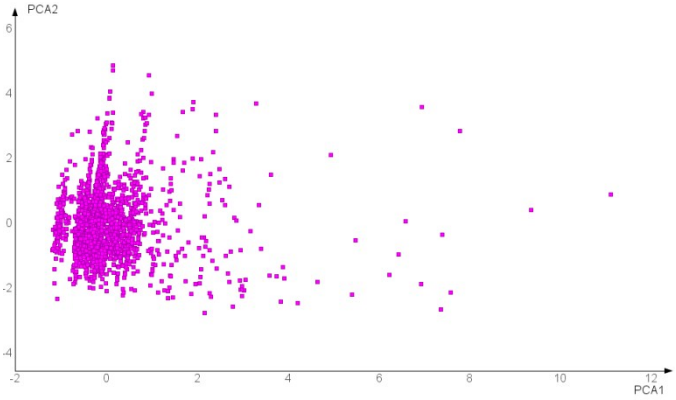


Figure S1. Chemical space representation of BRDi, HDACi, DNMT 1, 'Epi-focused', and GRAS. Data sets are in the same reference coordinates. The chemical space was obtained by principal component analysis of six physicochemical properties of pharmaceutical interest. The outliers in the HDACs and GRAS sets are not shown for clarity). See main text for details.

Table S3. Nemenyi test values for the PCPs comparisons of eight data sets (1 BRDi; 2 HDACi; 3 DNMT 1; 4 GRAS; 5 Drugs; 6 Clinic; 7 General; 8 Epi-focused).

SlogP

	BRDi	HDACi	DNMT 1	GRAS	Drugs	Clinic	General
2	0.0005	-	-	-	-	-	-
3	9.80E-14	5.20E-14	-	-	-	-	-
4	< 2E-16	< 2E-16	1.10E-07	-	-	-	-
5	< 2E-16	< 2E-16	7.40E-14	< 2E-16	-	-	-
6	< 2E-16	< 2E-16	1.82E-01	< 2E-16	4.50E-05	-	-
7	1.00E-05	0.824	2.90E-13	< 2E-16	< 2E-16	< 2E-16	-
8	2.60E-12	7.50E-07	0.991	0.483	3.00E-04	0.999	0.382

TPSA

	BRDi	HDACi	DNMT 1	GRAS	Drugs	Clinic	General
2	1.10E-09	-	-	-	-	-	-
3	6.90E-09	0.028	-	-	-	-	-
4	< 2E-16	< 2E-16	< 2E-16	-	-	-	-
5	0.0004	< 2E-16	6.50E-14	< 2E-16	-	-	-
6	0.62	0.6	0.999	< 2E-16	7.80E-14	-	-
7	0.6	0.025	0.999	< 2E-16	< 2E-16	1	-
8	0.999	0.303	0.999	< 2E-16	0.003	0.987	0.987

MW

	BRDi	HDACi	DNMT 1	GRAS	Drugs	Clinic	General
2	0.006	-	-	-	-	-	-
3	0.879	0.028	-	-	-	-	-
4	< 2E-16	< 2E-16	< 2E-16	-	-	-	-
5	0.0004	< 2E-16	6.50E-14	< 2E-16	-	-	-
6	0.62	0.6	0.999	< 2E-16	7.80E-14	-	-
7	0.6	0.025	0.999	< 2E-16	< 2E-16	1	-
8	0.999	0.303	0.999	< 2E-16	0.003	0.987	0.987

HBA

	BRDi	HDACi	DNMT 1	GRAS	Drugs	Clinic	General
2	0.001	-	-	-	-	-	-
3	1.10E-06	0.125	-	-	-	-	-
4	< 2E-16	< 2E-16	< 2E-16	-	-	-	-
5	2.50E-13	< 2E-16	< 2E-16	< 2E-16	-	-	-
6	0.0806	< 2E-16	< 2E-16	< 2E-16	9.10E-14	-	-
7	0.342	0.0211	2.00E-06	< 2E-16	< 2E-16	7.70E-14	-
8	0.598	3.40E-06	4.10E-09	6.40E-14	0.004	1	0.0028

HBD

	BRDi	HDACi	DNMT 1	GRAS	Drugs	Clinic	General
2	< 2E-16	-	-	-	-	-	-
3	2.50E-14	0.002	-	-	-	-	-
4	< 2E-16	< 2E-16	< 2E-16	-	-	-	-
5	1	< 2E-16	< 2E-16	< 2E-16	-	-	-
6	0.792	< 2E-16	< 2E-16	< 2E-16	0.04	-	-
7	4.00E-12	< 2E-16	6.10E-10	< 2E-16	< 2E-16	6.00E-14	-
8	0.0002	8.80E-08	0.004	< 2E-16	1.00E-06	0.001	0.382

RTB

	BRDi	HDACi	DNMT 1	GRAS	Drugs	Clinic	General
2	< 2E-16	-	-	-	-	-	-
3	0.027	< 2E-16	-	-	-	-	-
4	0.128	< 2E-16	9.50E-14	-	-	-	-
5	1	< 2E-16	2.80E-06	2.10E-07	-	-	-
6	0.0001	< 2E-16	0.694	< 2E-16	1.00E-13	-	-
7	0.367	< 2E-16	0.565	8.80E-14	0.0007	0.0009	-
8	0.39	7.10E-14	1	0.0001	0.164	0.968	0.989

Table S4. Loadings for the first three principal components of the property space of the eight compound data sets.

Principal component	PC1	PC2	PC3
Eigenvalue	1.9185	1.1174	0.7006
Cumulative eigenvalue (%)	61.344	82.157	90.338
SlogP	-0.0125	0.3832	-0.1515
TPSA	0.00904	-0.00414	-0.000591
MW	0.00284	0.00181	-0.000909
HBA	0.1949	-0.0369	-0.1355
HBD	0.2462	-0.0908	-0.1135
RTB	0.0895	0.0743	0.201

Table S5. Summary statistics of the intra-compound set similarity

MACCS

Dataset	Q1	Mean	Median	Q3	SD
BRDi	0.386	0.48441626	0.463	0.557	0.14001451
HDACi	0.3445	0.4322109	0.4235	0.509	0.12798872
DNMT	0.309	0.41415293	0.4	0.5	0.16258865
GRAS	0.161	0.28927381	0.261	0.387	0.17045788
Drugs	0.22	0.31403743	0.308	0.4	0.1336488
Clinic	0.283	0.37388603	0.375	0.466	0.13127422
General	0.28	0.36957258	0.365	0.453	0.12375299
Epi-Focused	0.313	0.40095354	0.399	0.488	0.12786925

GpiDAPH3

Dataset	Q1	Mean	Median	Q3	SD
BRDi	0.228	0.32056677	0.303	0.403	0.17481642
HDACi	0.34475	0.43225473	0.42375	0.50975	0.12792421
DNMT	0.161	0.24319609	0.243	0.324	0.18191181
GRAS	0	0.15952901	0	0.342	0.22637638
Drugs	0.2095	0.28546563	0.286	0.364	0.13806921
Clinic	0	0.19669936	0.223	0.3	0.13990382
General	0	0.18843588	0.218	0.303	0.14801609
Epi-Focused	0.194	0.25483202	0.275	0.344	0.13709412

TGD

Dataset	Q1	Mean	Median	Q3	SD
BRDi	0.678	0.73223165	0.729	0.782	0.08345325
HDACi	0.3965	0.4277573	0.4455	0.4765	0.06328818
DNMT 1	0.526	0.59578227	0.595	0.663	0.1255546
GRAS	0.471	0.60436491	0.607	0.739	0.18054764
Drugs	0.408	0.50042214	0.513	0.603	0.14793078
Clinic	0.485	0.56071604	0.577	0.652	0.13069903
General	0.504	0.57793955	0.587	0.661	0.11752032
Epi-Focused	0.51875	0.58460793	0.599	0.665	0.11791897

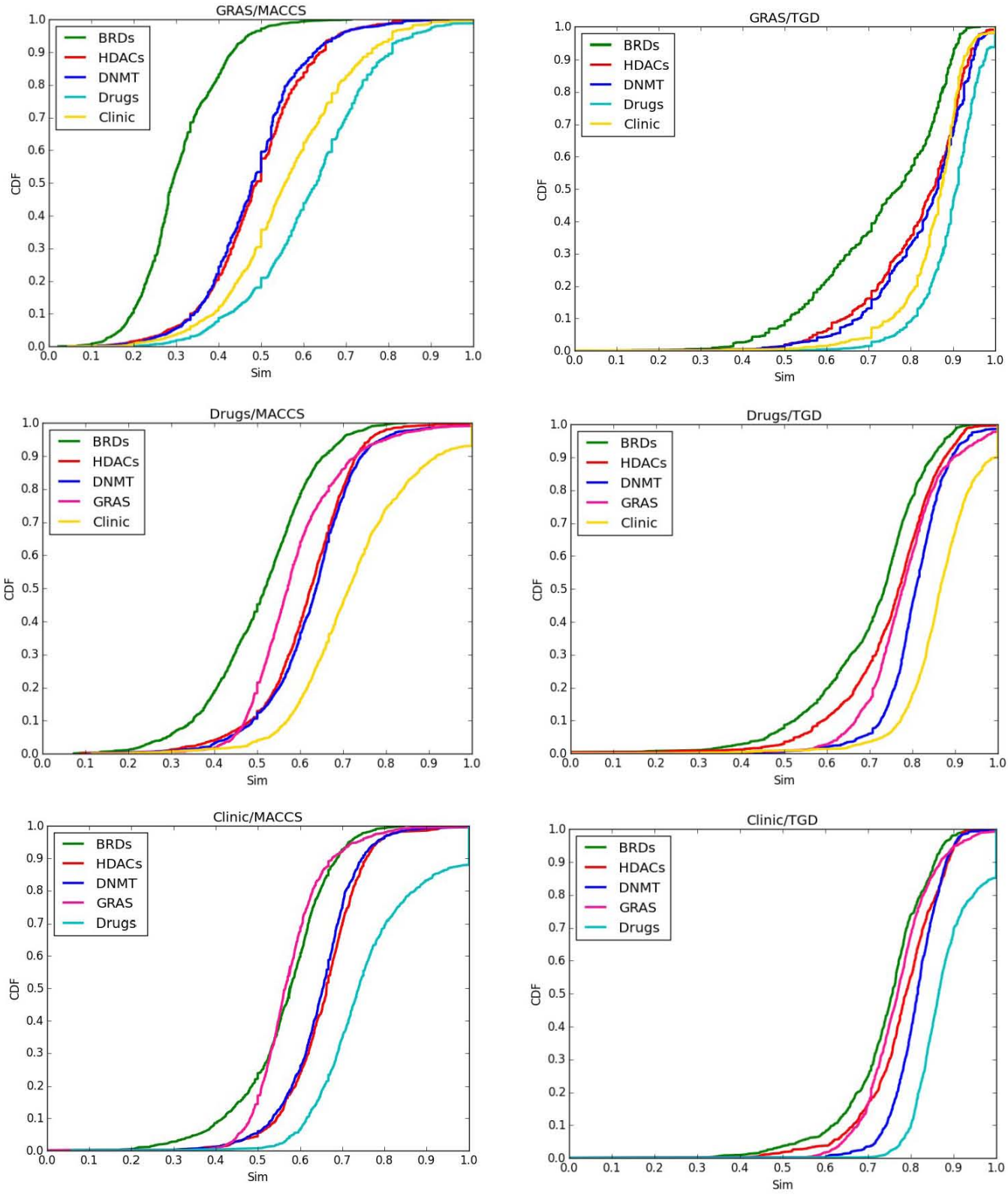


Figure S2. Inter-library similarity: Cumulative distribution function (CDF) of the maximum structure similarity calculated with MACCS keys, TDG and the Tanimoto coefficient comparing different reference and test sets. The reference set is indicated at the top of each graph. Summary statistics of the CDFs are presented in Tables S6 and S7.

Table S6. Summary statistics of the inter-library comparison using MACCS keys; the reference set is shown at the top of each table.

BRDi

	HDACi	DNMT 1	GRAS	Drugs	Clinic
Min	0.5	0.521	0.431	0.535	0.536
Q1	0.610	0.590	0.511	0.647	0.642
Median	0.661	0.652	0.537	0.684	0.688
Q3	0.700	0.710	0.578	0.733	0.724
Max	1	0.867	0.712	0.846	0.846
SD	0.079	0.073	0.056	0.066	0.068

HDACi

	BRDi	DNMT 1	GRAS	Drugs	Clinic
Min	0.204	0.392	0.372	0.468	0.395
Q1	0.544	0.586	0.518	0.618	0.636
Median	0.605	0.625	0.559	0.652	0.682
Q3	0.651	0.671	0.611	0.697	0.729
Max	1	1	1	1	1
SD	0.082	0.073	0.073	0.066	0.076

DNMT 1

	BRDi	HDACi	GRAS	Drugs	Clinic
Min	0.073	0.088	0.153	0.333	0.181
Q1	0.538	0.639	0.531	0.651	0.646
Median	0.593	0.683	0.593	0.708	0.694
Q3	0.625	0.742	0.655	0.803	0.790
Max	0.867	1	1	1	1
SD	0.095	0.088	0.098	0.105	0.103

GRAS

	BRDi	HDACi	DNMT 1	Drugs	Clinic
Min	0.022	0.052	0.058	0.2	0.045
Q1	0.351	0.445	0.423	0.411	0.507
Median	0.528	0.632	0.582	0.606	0.670
Q3	0.590	0.683	0.642	0.665	0.733
Max	0.712	1	1	1	1
SD	0.214	0.266	0.244	0.268	0.284

Drugs

	BRDi	HDACi	DNMT 1	GRAS	Clinic
Min	0.073	0.088	0.108	0.105	0.115
Q1	0.429	0.560	0.571	0.514	0.633
Median	0.518	0.625	0.638	0.571	0.714
Q3	0.589	0.684	0.690	0.641	0.805
Max	0.846	1	1	1	1
SD	0.114	0.099	0.098	0.092	0.128

Table S6. (Continued)

	Clinic				
	BRDi	HDACi	DNMT 1	GRAS	Drugs
Min	0	0.022	0	0	0.058
Q1	0.513	0.604	0.600	0.522	0.675
Median	0.576	0.662	0.654	0.566	0.739
Q3	0.630	0.714	0.700	0.616	0.833
Max	0.846	1	1	1	1
SD	0.110	0.095	0.093	0.087	0.129

Q1: First quartile; Q3: Third quartile; SD: standard deviation

Table S7. Summary statistics of the inter-library comparison using TGD fingerprints; the reference set is shown at the top of each table.

BRDi

	HDACi	DNMT 1	GRAS	Drugs	Clinic
Min	0.791	0.764	0.725	0.779	0.761
Q1	0.846	0.826	0.809	0.842	0.849
Median	0.862	0.85	0.832	0.862	0.868
Q3	0.885	0.868	0.873	0.885	0.887
Max	1	0.929	0.963	1	0.948
SD	0.035	0.034	0.048	0.032	0.032

HDACi

	BRDi	DNMT 1	GRAS	Drugs	Clinic
Min	0.506	0.577	0.68	0.693	0.67
Q1	0.769	0.769	0.751	0.793	0.797
Median	0.800	0.800	0.775	0.828	0.830
Q3	0.845	0.845	0.805	0.857	0.867
Max	1	1	1	1	1
SD	0.048	0.048	0.049	0.044	0.045

DNMT 1

	BRDs	HDACs	GRAS	Drugs	Clinic
Min	0	0	0.426	0.703	0.447
Q1	0.701	0.729	0.712	0.825	0.792
Median	0.752	0.783	0.752	0.857	0.842
Q3	0.812	0.861	0.799	0.889	0.871
Max	0.929	1	1	1	1
SD	0.090	0.086	0.079	0.054	0.059

GRAS

	BRDi	HDACi	DNMT 1	Drugs	Clinic
Min	0	0	0	0	0
Q1	0.618	0.748	0.761	0.863	0.825
Median	0.771	0.855	0.863	0.907	0.873
Q3	0.863	0.908	0.913	0.943	0.905
Max	0.963	1	1	1	1
SD	0.153	0.121	0.117	0.075	0.085

Drugs

	BRDs	HDACs	DNMT	GRAS	Clinic
Min	0	0	0	0	0
Q1	0.629	0.692	0.771	0.725	0.821
Median	0.734	0.767	0.808	0.778	0.866
Q3	0.790	0.825	0.850	0.832	0.918
Max	1	1	1	1	1
SD	0.134	0.122	0.081	0.098	0.092

Table S7. (Continued)

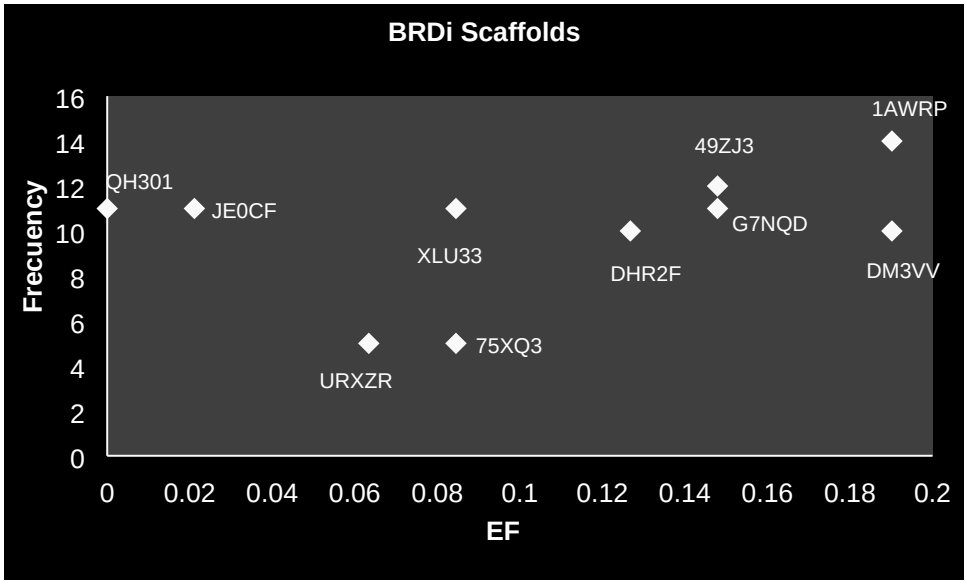
	Clinic				
	BRDi	HDACi	DNMT 1	GRAS	Drugs
Min	0	0	0	0	0
Q1	0.701	0.736	0.778	0.717	0.828
Median	0.757	0.789	0.816	0.769	0.865
Q3	0.808	0.840	0.853	0.816	0.919
Max	0.946	1	1	1	1
SD	0.104	0.095	0.065	0.083	0.076

Q1: First quartile; Q3: Third quartile; SD: standard deviation

Table S8. Descriptive statistics and AUC for CSR curves

Set	Q1	FS_{50}	Q3	AUC	F_{50}
BRDi	0.661	0.783	0.894	0.739	0.122
HDACi	0.669	0.802	0.901	0.762	0.098
DNMT 1	0.525	0.741	0.87	0.689	0.226
GRAS	0.909	0.948	0.974	0.926	0.004

Q1: First quartile; FS_{50} fraction of database recovered by 50% of scaffolds; Q3: Third quartile; AUC: area under the curve; F50: fraction of scaffolds to recover 50% of the compound in the data set.



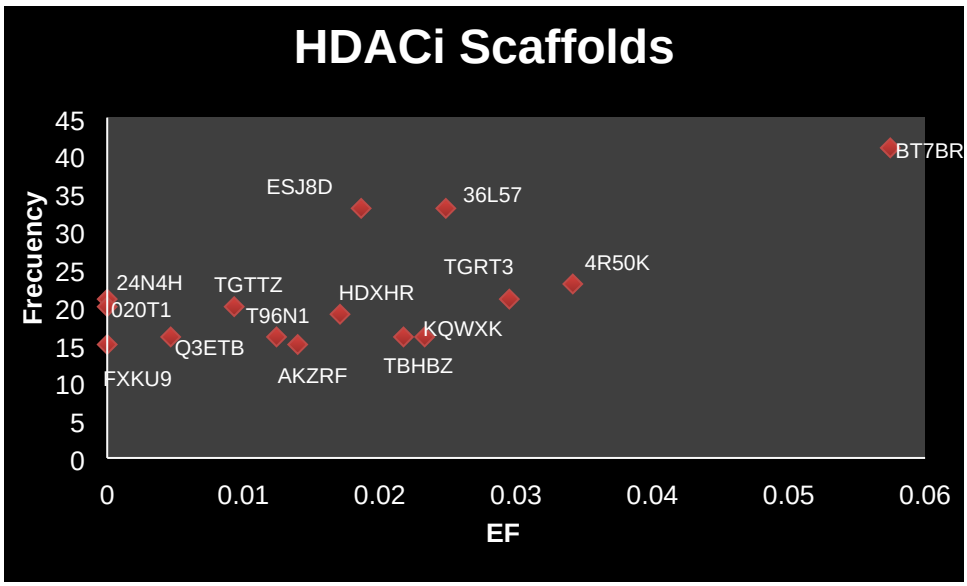


Figure S3. Chemotype enrichment plots for BRDi and HDACi data sets.

Table S9. Descriptive statistics for F-chiral and F-sp^3 distributions

F-chiral

	BRDi	HDACi	DNMT 1	GRAS	Drugs	Clinic	General	Epi-focused
Min	0	0	0	0	0	0	0	0
1st Quart	0	0	0	0	0	0	0	0
Median	0	0	0	0	0.058	0.032	0	0
Mean	0.024	0.034	0.091	0.057	0.107	0.076	0.048	0.034
3rd Quart	0.045	0.043	0.142	0.1	0.173	0.105	0.058	0.038
Max	0.166	0.7	0.667	0.833	0.833	0.833	0.55	0.5

F-sp^3

	BRDi	HDACi	DNMT 1	GRAS	Drugs	Clinic	General	Epi-focused
Min	1	1	0	0	0	0	0	0
1st Quart	3	3	0.08	0.4	0.29	0.24	0.19	0.13
Median	4	4	0.2	0.67	0.43	0.38	0.3	0.28
Mean	3.662	4.813	0.264	0.622	0.458	0.401	0.325	0.271
3rd Quart	4	5	0.42	0.86	0.61	0.53	0.44	0.37
Max	7	37	0.91	1	1	1	1	0.68

Table S10. Nemenyi pairing values for F-chiral and F-sp3 distributions (1 BRDi; 2 HDACi; 3 DNMT 1; 4 GRAS; 5 Drugs; 6 Clinic; 7 General; 8 Epi-focused).

F-chiral

	BRDi	HDACi	DNMT 1	GRAS	Drugs	Clinic	General
2	0.9991	-	-	-	-	-	-
3	0.04151	9.90E-08	-	-	-	-	-
4	0.99064	0.0603	0.0013	-	-	-	-
5	2.60E-11	<2E-16	5.00E-07	< 2E-16	-	-	-
6	0.0044	2.40E-13	0.996	4.10E-07	1.00E-06	-	-
7	0.971	0.08	0.017	0.999	< 2E-16	9.00E-05	-
8	0.912	0.927	0.002	0.372	1.10E-10	0.0003	0.305

F-sp^3

	BRDi	HDACi	DNMT 1	GRAS	Drugs	Clinic	General
2	0.021	-	-	-	-	-	-
3	0.951	9.90E-08	-	-	-	-	-
4	< 2E-16	< 2E-16	< 2E-16	-	-	-	-
5	< 2E-16	< 2E-16	< 2E-16	< 2E-16	-	-	-
6	0.0044	2.40E-13	7.60E-14	< 2E-16	4.50E-05	-	-
7	0.971	0.08	1.40E-05	< 2E-16	< 2E-16	4.20E-09	-
8	0.912	0.927	1	< 2E-16	6.80E-12	1.20E-05	0.382