RSC Advances

ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/advances

Epigenetic relevant chemical space: A chemoinformatic characterization of inhibitors of DNA methyltransferases

Eli Fernández-de Gortari and José L. Medina-Franco*

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

*Address correspondence to this author. E-mails: medinajl@unam.mx; jose.medina.franco@gmail.com
Tel. +5255-5622-3899, ext. 44458

Abstract

DNA methylation is an epigenetic mechanism mediated by the family of proteins DNA methyltransferases (DNMTs). The misregulation of the covalent modification of DNA through the addition of a methyl group at the carbon-5 position of cytosine residue is common in many diseases including cancer. Recent advances in synthetic and screening technologies for DNMT inhibitors (DNMTi) have made significant contributions to uncover promising candidates for epigenetic drug discovery. The structure-activity information, not available few years ago, is being collected in public molecular databases. However, no systematic chemoinformatic studies that analyze the structural diversity and coverage of the chemical space of DNMTi with experimental activity have been discussed thus far. Herein, we report the assembly and curation of a molecular database of small-molecule DNMTi with a special focus on inhibitors of DNMT1. The compound collection was characterized using a comprehensive chemoinformatic approach that involved physicochemical properties, structural fingerprints, and molecular scaffolds. The availability of activity information enabled to conduct chemotype enrichment analysis and suggest potential privileged epigenetic scaffolds. The structures of inhibitors of DNMT1 were compared to drugs approved for clinical use, compounds in clinical trials, a commercial screening library focused on epigenetic targets, and a general screening collection. The results of this work provided key insights to start characterizing the epigenetic relevant chemical space.

Keywords: chemical space, chemoinformatics, molecular scaffolds, structural diversity, structure-activity relationships

Abbreviation List: AUC, area under the curve; CSR, cyclic systems retrieval; DNMT, DNA methyltransferase; DNMTi, DNA methyltransferase inhibitor; ECFPs, Extended Connectivity Fingerprints; EF, enrichment factor; GpiDAPH3, Pharmacophore Graph Triangle; HBA hydrogen bond acceptors; HBD, hydrogen bond donors; HEMD, Human Enzyme and Modulator Database; MOE, Molecular Operating Environment; MW, molecular weight; PC, principal component; PCA, principal component analysis; RB, rotatable bonds; SlogP, partition coefficient octanol/water; SMILES, Simplified Molecular Input Line-Entry System; TGD, Typed Graph Distance; TPSA, topological polar surface area; TTD, Therapeutic Target Database.

1. Introduction

DNA methylation is an epigenetic modification involving the addition of methyl group at position 5C of a cytosine residue. This process plays a key role in mammal's development and in cancer cell growth. The methylation process is mediated by an enzymatic family called DNA methyltransferases (DNMTs). In humans, this family includes DNMT1, DNMT2, DNMT3A and DNMT3B.[1] DNMT1 and DNMT3B exhibit larger activity, which can be inferred from the strong reduction in DNA methylation in cell lines with double knock-out. DNA methylation represents one of the main mediation of epigenetic regulation. Therefore, the identification of novel DNMT inhibitors (DNMTi) is a promising research avenue to develop novel therapies against cancer and other diseases associated with epigenetic alterations.[2-4]

Currently, 5-aza and 5-aza-2´-deoxycytidine are two drugs approved for clinical use for the treatment of myelodysplasia (Figure 1). 5-Aza and 5-aza-2´-deoxycytidine are nucleoside analogues which, after its incorporation into DNA, cause depletion of the DNMTs. However, these drugs have

high toxicity, low bioavailability and low chemical stability, coupled with an uncertain mechanism of antitumor activity.[5] For this reason, research efforts to discover non-nucleoside DNMTi with greater specificity and lower toxicity are needed.

<Insert Figure 1 here>

One of the main advantages of non-nucleoside DNMTi is that they do not need to be incorporated into the DNA. This characteristic contributes to the possible development of selective inhibitors against different DNMTs with the consequent decrease of unwanted side effects. Thus far, several non-nucleoside inhibitors have been identified such as SGI-1027, procainamide, tea polyphenol (-)-epigallocatechin 3-gallate, genistein, NSC401077, hydralazine, among others.[6,7] The first-generation of inhibitors showed low activity and selectivity against DNMTs. However, new generations of inhibitors with increased activity and selectivity profile have been developed, such as analogs of SGI-1027 (Figure 1).[8] Nevertheless, these compounds have low potency especially in cells and lack of selectivity towards different DNMTs. Figure 1 shows representative inhibitors of DNMT1 and molecules associated with demethylating properties.

The increased research efforts to develop more potent and specific DNMTi have augmented notoriously the number of screening data. Compounds from different sources including general screening collections and synthetic compounds have been tested for inhibition of DNMTs, in particular DNMT1. The results are being reported and collected not only in research papers but also in compound databases available in the public domain. Examples of such databases are the Human Epigenetic Enzyme and Modulator Database (HEMD),[9] developed at the Shanghai Jiao Tong University, and EpiDBase,[10] a recently manually curated epigenetic database that contains 11,422 small molecules with activity against different epigenetic targets.

Compound collections either in-house or publicly available are an essential part on lead identification and drug development efforts.[11,12] Chemoinformatic characterization of chemical libraries is a major first step towards the virtual and or experimental screening to identify new molecules with biological activity. As discussed in detail elsewhere, analysis of the structural diversity and analysis of

distribution in chemical space of compound databases can be a crucial early step in virtual screening.[13,14] For example, diverse libraries are appropriate to identify hit compounds with novel chemical scaffolds for a given target such as DNMTs. Focused libraries would be more suitable in lead optimization efforts.[13] Chemoinformatic analysis of chemical libraries has found several applications not only in the research area of epigenetic-related libraries discussed in this work but also in drug and probe discovery based on natural products, combinatorial libraries and food-related chemicals.[11,13,15]

As part of our ongoing efforts towards the development of DNMTi, herein we compiled and curated a molecular database of DNMTi. Specifically, we collected structure and activity information currently available in public molecular databases and recent scientific literature for compounds tested as inhibitors of DNMT1. Part of this data set was recently analyzed using an activity landscape approach to identify 'activity cliff generators', that is, compounds that are structurally similar to other molecules but have different activity profiles.[16] As explained in detailed in that report, the activity landscape study was based on systematic pair-wise comparisons of the structure and activity similarity. The structure-activity relationships of the activity cliff generators were explained, at the molecular level, using docking studies.[16] In contrast, in this work the structural diversity and coverage of the chemical space[14] of the assembled library was characterized using multiple criteria including physicochemical properties, structural fingerprints, and molecular scaffolds. The herein collection of DNMTi was compared to several reference compound databases commonly used in drug discovery campaigns such as approved drugs, compounds in clinical trials, a general screening collection, and a commercial screening library focused on epigenetic compounds. It has been demonstrated that chemoinformatic characterization of compounds collections gives rise to valuable insights that guide the discovery and development of bioactive molecules.[12,14,17,18] Of note, the present chemoinformatic analysis is focused on the structural aspects of the chemical structures of a large collection of inhibitors of DNMT. This work complements other recent reviews and research reports that discuss in detail three-dimensional aspects of the protein-ligand interactions of inhibitors of DNMT1.[19-21]

A substantial advance to the field of this work relative to previous chemoinformatic analysis of epigenetic-related databases[22] is that herein we analyze a curated data set with experimental DNMT inhibitory activity. This is in sharp contrast with a chemoinformatics study published four years ago for a commercial DNMT-focused library but with no experimental activity. As discussed above, the growing interest of the scientific community on epigenetic drug discovery has boosted the availability of screening data until recent years.

2. Methods

2.1 Compound Databases

2.1.1 Database of inhibitors of DNMT1

A molecular database of inhibitors of DNMT1 was assembled collecting information from four major sources including three public compound databases and literature searching. The three public databases were ChEMBL[23] using the query text 'DNMT1' in the target browser; HEMD[9] with the information located in the enzyme browser, option DNA and submenu DNA (cytosine-5)-methyltransferase 1; and Binding Database,[24] in the IC_{50} menu and submenu DNA methyltransferase. The search retrieved 265, 106, and 337 molecules from ChEMBL, HEMD, and Binding Database, respectively. In order to identify additional compounds not reported in the three major public databases literature was searched using Web of Science (https://isiknowledge.com) focusing the search on papers published from 2013 to the time of writing this manuscript (March 2015). Literature searching retrieved 47 additional molecules. Table 1 summarizes the molecular databases and number of compounds analyzed in this study. Activity data for all compounds (IC_{50} values) were converted to micro molar units.

<Insert Table 1 here>

Data curation was carried out following the methodology reported by Fourches *et al*.[25]. Linear notation canonical structure according to simplified molecular input line-entry system (SMILES) was obtained for each molecule. Molecules where further prepared using the "wash" module available in

Molecular Operating Environment (MOE)[26] by disconnecting metal salts, remove simple components, rebalance protonation state and deleting compounds with the same structure and activity data. Identical compounds with close but different activity values were kept taking, for only one structure, the mean value of the IC_{50} value. If the difference of the IC_{50} value was too large, the compound was removed from the analysis.

2.1.2 Reference databases

The database of DNMT1 was compared to four reference collections, namely; 1) a general screening collection obtained from Selleckchem (henceforth referred as 'general'), 2) a screening collection focused on epigenetic compounds also obtained from Selleckchem[27-29] (henceforth referred as 'focused'); 3) compounds in clinical trials obtained from the Therapeutic Target Database (TTD) (henceforth referred 'clinical'),[30] and approved drugs obtained from DrugBank (henceforth referred 'approved').[31] Table 2 summarizes the reference collections.

<Insert Table 2 here>

2.2 Structure representation

The compound databases were analyzed using three complementary representations, namely; physicochemical properties, structural fingerprints and molecular scaffolds.[18] Three representations were used to balance the advantages and disadvantages of each one. For example, physicochemical properties are whole molecule descriptors straightforward to interpret and commonly used to build "drug-like" and other similar empirical rules.[32] However, physicochemical properties do not give information about the molecular structural pattern and different structures may present similar or equal physicochemical properties. Similar to physicochemical descriptors, molecular scaffolds are straightforward to interpret and allow an easy communication with medicinal chemists and biologists. Certainly, scaffold analysis has led to concepts widely used in medicinal chemistry and drug discovery such as 'scaffolds hopping' and 'privileged structures'. However, one of the disadvantages of scaffolds analysis is the lack of information due to side chains and the inherent similarity or dissimilarity of the

scaffolds themselves. To overcome this limitation, structural fingerprints can be used since they usually encode the information from the entire chemical structure. Structural fingerprints are widely used and have been successfully applied in many diversity studies and assessment of chemical space.[33,34] Nevertheless, a disadvantage of the most commonly used molecular fingerprints is that they are not straightforward to interpret and the dependence of chemical space with the type of fingerprint used.[35] This limitation can be approached obtaining consensus conclusions obtained from several different fingerprints and representations.

2.2.1 Physicochemical properties

Six physicochemical properties of pharmaceutical interest were calculated for the curated databases using MOE: partition coefficient octanol/water (SlogP), rotatable bonds (RB), hydrogen bond donors (HBD), hydrogen bond acceptors (HBA), topological surface area (TPSA) and molecular weight (MW). The six descriptors include the three important properties of size (MW), flexibility (RB), and molecular polarity. This set of properties is commonly used to compare compound collections for drug discovery.[36] The distribution of the six properties was calculated and summary statistics were obtained including the mean, median, interquartile distances and standard deviation. For each database and properties notched-boxplot and violin plot were generated, as well as distribution type. Homoscedasticity analysis and subsequent hypothesis testing and post-hoc analysis were carried in RGui with package PMCMR.[37] In order to generate a visual representation of the property space (*i.e.*, chemical space based on physicochemical properties), a principal components analyses (PCA) was carried out on the six calculated properties using MOE and DataWarrior.[38]

2.2.2 Structural fingerprints

Compound databases were studied using fingerprints of different design:[18] Molecular ACCess System (MACCS) keys (166-bits), Pharmacophore Graph Triangle (GpiDAPH3), and Typed Graph Distance (TGD), as implemented in MOE, and Extended Connectivity Fingerprints (ECFPs) with radius equal to

four and six as implemented in MayaChem Tools (available at http://www.mayachemtools.org/docs/scripts/html/index.html). MACCS keys were originally developed to increase speed in structure search. Each bit describes a small substructure with maximum ten atoms except hydrogen. GpiDAPH3 consists of a pharmacophore of three points calculated from the two dimensions structure graph; TGD is a distance graph fingerprint type, where each fingerprint represent a set of three elements in the form (u, v, d), where u and v are atoms types like acid, base, hydrogen bond donor and acceptor, among others; and d is the minimum distance between two vertices of the graph. ECFP is a circular topology fingerprint with variable diameter distance. Despite the fact MACCS keys where designed to augment molecular diversity and conduct similarity searching, they have proven to be very useful to describe chemical space of compound databases.[39] Since different fingerprints provide complementary information, it was possible to derive a consensus assessment of the structural diversity of the compound databases.

The structural similarity was computed using the Tanimoto similarity coefficient:[40,41]

$$T(a,b)=\frac{c}{a+b-c} \qquad (1)$$

where *a* and *b* are the number of fragment bits corresponding to the *i-th* and *j-th* molecules and *c* is the number of fragments bits common to both molecules. For each similarity matrix, random samples of 5000 similarity values off the diagonal were extracted to calculate statistics (e.g., mean, median, interquartile distances and standard deviation) and generate plots of the cumulative distribution functions. The distribution type, homoscedasticity, hypothesis testing and post-hoc analysis were conducted using PCMCR package in RGui.[37]

2.2.3 Molecular scaffolds

In this study the scaffolds also called cyclic systems were generated by systematically removing the side chains from the molecules *i.e.*, removal of the vertex with degree one, with the program Molecular Equivalent Indices (MEQI).[42] The cyclic systems are part of the chemotypes defined in the

methodology developed by Johnson and Xu. For each cyclic system a unique chemotype identifier (chemotype code) with five characters is assigned. The cyclic systems represent equivalent classes and molecules classified in a cyclic system do not fall into other chemotype class. MEQI has been broadly used for the scaffold analysis of a large number of compound databases.[43-45]

In order to measure the scaffold diversity of the inhibitors of DNMT1 the number of cyclic systems was recorded along with singletons (cyclic systems that contain only one compound). The fraction of cyclic systems relative to the size of the data set, the fraction of singletons relative to the size of the data set and the number of cyclic systems were computed and compared to values reported in the literature for other data sets.[46] The distribution of the molecular scaffolds was further characterized using the cyclic systems retrieval (CSR) curve which is fully discussed elsewhere.[46,47] 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 of the inhibitors of DNMT1 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 fraction of cyclic systems that contain 50% of the inhibitors of DNMT1 and the area under the curve (AUC).[46]

2.3 Active scaffolds in the database of inhibitors of DNMT1

The molecular scaffolds containing 'active' DNMT1 inhibitors were also analyzed. For this analysis a compound was considered 'active' for IC_{50} values was lower than 10μM. DNMT1 active scaffolds were detected using well-established measures[43] briefly described hereunder.

The background activity Act(C) is the fraction of active compounds in the data base and was calculated with the expression:

$$Act(C) = [C^*] / [C] \qquad (2)$$

where [C] is the total number of compounds, and [C*] is the total number of active compounds.

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

$$Act(C_\lambda) = [C_\lambda^*] / [C_\lambda] \quad (3)$$

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 computed with the expression:

$$EF(C_\lambda) = Act(C_\lambda) / Act(C) \quad (4)$$

$EF(C_\lambda)$ measured the proportion of active molecules of a particular chemotype relative to the proportion of active compounds in the data set. Therefore, the molecular scaffolds with the highest EF were 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.[43] Chemotype enrichment plots have been used in the scaffold analysis of compound databases.[43, 44]

3. Results and discussion

The analysis of the database of inhibitors of DNMT1 is organized in two major parts: diversity assessment using different representations and exploration of the cyclic systems with DNMT1 inhibitory activity.

3.1 Analysis of compounds tested as inhibitors of DNMT1

This section is further divided in three major parts, each one focused on a different representation.

3.1.1 Physicochemical properties

Figure 2 shows the distributions of the pharmaceutically relevant physicochemical properties calculated with MOE. For all data sets the distributions are summarized as notch box plots. In these plots the

boxes represent the interquartile distance and enclose data points with values within the first and third quartile. The bold black line denotes the median of the distribution, and the lines above and below indicates the maximum and minimum value excluding the outliers. The open circles denote the outliers and notch represents the 95% confidence interval of the median. Summary statistics of the distributions are presented below each plot.

<Insert Figure 2 here>

None of the distributions showed a normal distribution as measured using the Shapiro test implemented in R.[37] The statistical difference between the distributions was assessed using the non-parametric Kruskal-Wallis analysis and a post hoc Nemenyi test implemented in R package PMCMR.[37] Results of the statistical analysis are presented in Figure S1 in the Supporting Information.

The database of inhibitors of DNMT1 showed, overall, a slightly larger number of HBA and HBD than approved drugs, compounds in clinical trials and the general screening collection. The distribution of both properties of the inhibitors of DNMT1 was comparable with the collection focused on epigenetic targets (*i.e.,* median HBA and HBD values of 5 and 2, respectively, and a significant pairwise differences p values of 0.35 and 0.79, respectively (see Figures 2 and S1). Inhibitors of DNMT1 also had higher TPSA values than other reference databases. These results indicated that the compounds screened as inhibitors of DNMT1 are, in general, more polar than other compounds considered in this study.

The distribution of SlogP and MW values of inhibitors of DNMT1 was similar to the other reference collections except approved drugs. Actually, the collection of approved drugs showed slightly lower SlogP and MW values than the other databases including compounds in clinical trials.

Finally, the distribution of RB of inhibitors of DNMT1 was similar to the other reference collections (Figure 2). In particular the statistical analysis showed a very similar distribution of the RB values (p=0.99) of the DNMT1 compounds and screening molecules focused on epigenetic targets. Also, the general and focused screening collections showed high similarity (p=0.88) of the distributions of RB values (Figure S1). These results indicated comparable flexibility as measured by this property.

3.1.1.1 Visual representation of the property space

The six physicochemical properties were subjected to a principal component analysis (PCA). Table 3 summarizes the percent of covariance that captures each of the first four principal components (PCs). These values indicate that the first two PCs captured 77.3% of the variance while the first three PCs captured 89.4 of the variance indicating that 2D and 3D PCA plots are reasonable visual representations of the property space generated for these libraries. Table 3 also summarizes the loading values of each property for the first four PCs. For the first PC, HBD and HBA had the highest loadings (0.134 and 0.106, respectively). SlogP was the property with the highest contribution to the second PC (loading value of 0.236), and RB was the property with the main contribution to the third PC (loading value of 0.220).

<Insert Table 3 here>

Figure 3 shows a 2D representation of the chemical space scatter plot of the first two PCs. As discussed above, this PCA plot captures 77.3% of the variance. This figure represents a 2D visual representation of the chemical space of the database of compounds tested as DNMT1 as compared to other four reference databases. All plots are in the same coordinate system. The panel at the top left shows all databases while the reminder of the panels depicts each data set separately. Figure S2 in the Supporting Information shows a 3D representation of the chemical space generated by plotting the first three PCs.

<Insert Figure 3 here>

The visual representation of the chemical space shows that, as expected, the general screening (data points in yellow), clinical (blue), and approved drugs (cyan) cover a wide area of the property space. Similar conclusions have been obtained in other studies comparing the property space of general screening collections with currently approved drugs.[18] In contrast, the collection focused on epigenetic targets (green) covers a more restricted area of the space. The compounds tested as inhibitors of DNMT1 (red) also cover a broad area of the property space. The more densely populated area is also

occupied by drugs, compounds in clinical trials and the general screening collection. However, the DNMT1 collection has molecules (data points) that expand a wide range of scores along PC1 (x-axis). As discussed above, HBD and HBA are the properties with the largest contributions to PC1. This is in agreement with the results of the distinct distributions of these properties for DNMT1 compounds (*vide supra*, Figure 2). Note, however that while interpreting the visual representation of the chemical space one needs to bear in mind that this is an *approximation* of the chemical space that is fully represented with the six properties.[48] Putting together the results of the visual representation of the property space is possible to conclude that most of the compounds tested as DNMT1 inhibitors cover the traditional medicinal property space and there are molecules that expand the traditional space.

3.1.1.2 Active vs. inactive inhibitors of DNMT1

We also investigated the distribution in the property space of the most active compounds tested as inhibitors of DNMT1, namely; 378 compounds with IC_{50} values lower than 10μM and 32 molecules with IC_{50} values lower than 1μM. The distribution of the six properties was compared with the properties of the remaining inactive compounds *i.e.,* molecules with higher IC_{50} values. Results are summarized in Figure S3 in the Supporting Information. HBA, HBD, and TPSA showed similar distributions for actives and inactives. Overall, active compounds showed higher MW and SlogP values suggesting that they are bigger and less hydrophilic that the inactive molecules as captured by these properties. Also, active compounds had slightly higher values of RB suggesting that active compounds are more flexible than the inactives.

3.1.2 Molecular fingerprints

3.1.2.1 Structural diversity of all data sets

First the structural diversity of the five compound databases was evaluated using four fingerprint representations of different design as described in the Methods section. As discussed above, different structural representations were used to address the dependence of chemical space with the

representation.[49] Figure 4 illustrates the cumulative distribution function of the similarity values computed with MACCS keys and TGD. Figure S4 in the Supporting Information shows summary statistics of the similarity distributions obtained with the Tanimoto coefficient and the four fingerprints.

<Insert Figure 4 here>

Results indicated that each representation has different magnitude of the similarity values. Overall, for any given compound database, similarity values calculated with TGD had the highest values (*e.g.*, mean similarity values between 0.54 and 0.64) followed by MACCS keys (*e.g.*, mean similarities between 0.31 and 0.41), GpiDAPH3 (e.g., mean values between 0.13 and 0.26) and ECFP4 (*e.g.*, mean similarity values between 0.06 and 0.07). The same relative order has been noted for other compound databases which can be associated with the design of the fingerprint.[18] For instance, the very low similarity values computed with ECFPs is associated with the high resolution of this fingerprint.

The distributions of the similarity values (Figure 4 and S4) showed that, in general, the set of compounds tested as inhibitors of DNMT1 are structurally diverse. The structural diversity of this set is comparable to the diversity of the set focused on epigenetic targets according to TGD, MACCS, and GpiDAPH3. For example, for MACCS keys which is highly used to compare compound databases,[11,17,18] the median similarities for the DNMT1 and focused set are 0.398 and 0.394, respectively. Similar conclusions can be obtained from other statistics for MACCS keys and other fingerprint representations used in this study.

The molecular fingerprint analysis indicated that the data set of approved drugs showed the highest structural diversity also according to TGD, MACCS keys and GpiDAPH3 fingerprints (for example, with median similarity values of 0.56, 0.30, and 0.14, respectively). Similar conclusions can be obtained with other statistics. The second and third most diverse libraries were the set of compounds in clinical trials and general screening collections, respectively.

Despite the fact ECFP4 has many successful applications in similarity-based virtual screening and activity landscape studies,[16,50] the very low similarity values made difficult to obtain quantitative conclusions in diversity analysis. Indeed, in a recently published activity landscape study of inhibitors

of DNMT1, extended connectivity fingerprints were useful to uncover activity cliff generators. However, the methods used in that work[16] were based on systematic pair-wise comparisons of the structure and activity similarity giving different insights of the structural diversity studies reported in this manuscript.

3.1.2.2 Structural diversity of active vs. inactive inhibitors of DNMT1

We also compared the structural diversity of the 'actives' (*i.e.*, 378 molecules with IC_{50} < 10 μM) and 'inactive' compounds in the data set of inhibitors of DNMT1 as computed with the Tanimoto coefficient and TGD and MACCs keys. Both sets of compounds showed similar distributions (data not shown) suggesting that, in general, the active compounds are also diverse and may be comprised of several different scaffolds. The molecular scaffold analysis, described in the next section, enabled to test this hypothesis.

3.1.3 Molecular scaffolds

The scaffold diversity of the compounds tested as inhibitors of DNMT1 was assessed using frequency counts and CSR curve as detailed in the Methods section. Table 4 includes a summary of the number of the cyclic systems (N), the fraction of cyclic systems relative to the number of molecules in the data set (N/M), and the number of cyclic systems with a single molecule, *i.e.*, singletons (N_{sing}). The fractions of singletons as compared to the number of cyclic systems (58%) and to the number of compounds in the data set (30%) are indicative of the large scaffold diversity of this compound collection. These numbers are comparable to the equivalent fractions of cyclic systems reported in the literature for data sets with activity against different molecular targets.[46] The corresponding CSR curve for the set of compounds tested as inhibitors of DNMT1 is presented in Figure S5 in the Supporting Information. This curve shows and AUC value of 0.67 and F_{50} value of 0.23 and further supported the conclusion that the DNMT1 is a diverse set.[46]

<Insert Table 4 here>

3.2 Active scaffolds in the database of inhibitors of DNMT1

Figure 5 depicts the most frequent cyclic systems for the data set of compounds tested as inhibitors of DNMT1. The chemotype identifier along with the cyclic system frequency and percentage are shown. The most frequent scaffolds can be classified in two major groups, namely; no-nucleosidic and cofactor related analogs. The most frequent scaffolds, with chemotyope identifiers FUIL1 and H8B7P had frequencies between 4 and 5% and are related to the co-factor *S*-adenosyl-L- methionine (SAM). Actually, the cyclic systems with chemotype identifiers FUIL1, H8B7P, KQ2XT, and WU6XX share the same sub-structure (of cyclic system H8B7P). All four cyclic systems together cover 72 compounds (12.7%) of the DNMT1 set. In contrast, the most frequent scaffolds associated with non-nucleosidic compounds (RYLFV, SU70D, 4E1HD, G5AA5, and RNDWX) add up only 40 molecules (7.1%) of the entire set. Other than the ubiquitous benzene (RYLFV) scaffold,[18,51] these results may be related to the historical development of inhibitors of DNMT1 that was initially based on the optimization of the approved drugs that are nucloeosidic compounds and cofactor analogs. However, due to the increased interest to develop non-nucleosidic compounds is expected an increase in the number of compound to be tested as inhibitors of DNMT1.

<Insert Figure 5 here>

3.2.1 Chemotype enrichment

In order to identify the most relevant cyclic systems in the set of DNMT1 compounds, the chemotype EF was calculated for the nine most frequent scaffolds in Figure 5. As discussed in the Methods section, the chemotype EF measures the proportion of active compounds for in a given scaffold (using a pre-established criterion to define an 'active' compound) relative to the proportion of active molecules in the entire data set. Considering this measure, the scaffolds with the highest EF values are the most attractive. Frequency is a second criterion to distinguish the most attractive scaffolds *i.e.*, those with higher frequencies have more reliable SAR information that those scaffolds with fewer compounds. In this context, chemotype enrichment plots are valuable 2D graphs to rapidly classify the

scaffolds by representing the enrichment factor on one axis of the plot and the frequency on the second axis. These plots have been used to identify attractive chemical scaffolds for data sets with different biological activity.[43] However, this this work represents the first analysis of the activity enrichment of chemotypes with DNMT1 inhibitory activity. Of note, a chemotype enrichment analysis was not possible to conduct with limited structure-activity relationship data.

Figure 6 shows the chemotype enrichment plot for the nine most frequent cyclic systems identified for the set of inhibitors of DNMT1. This plot shows that there are four cyclic systems with EF greater than 1.0 and three with EF greater than 1.4 (SU70D, G5AA5, and RNDWX). Interestingly, all four have a non-nuclosidic scaffold. Out of the four, SU70D was the cyclic system with the highest frequency (10 compounds) and, therefore, with the most reliable SAR in this set of scaffolds. The molecules with cyclic systems SU70D, RNDWX, and G5AA5 were obtained from high-throughput screening. The molecules sharing these molecular scaffolds were hits in a Fluorescent Molecular Beacon assay made by Sanford-Burnham Center for Chemical Genomics (PubChem Bioassay ID –AID – 602386).[52] These results encourage the additional exploration of the SAR of the compounds with these cyclic systems (either as potential chemotherapeutic agents or as molecular probes) and test its potential to become 'epigenetic privileged scaffolds' targeting DNMT1. For instance, these selected molecular scaffolds can be used as references (or queries) to conduct sub-structure searching on other compound databases from different origin *e.g.,* natural products or synthetic compounds. Hit compounds of the sub-structure search can be synthesized or purchased if they are commercially available.

<Insert Figure 6 here>

Molecules sharing the 4E1HD chemotype were identified from the study of Chen *et al*.[53] In that study the authors filtered a commercial screening database with 111,121 molecules. The filtered compounds were subject to docking-based virtual screening and clustering. Based on these results 51 molecules were selected for biological testing. Finally, homologous compounds of the hit molecules were acquired for further testing of their activity and selectivity profile with DNMTs.

The chemotype enrichment plot in Figure 6 shows that the cyclic system FU1L1 has an EF close to 1 (EF=0.93). This cyclic system is related to the co-factor SAM and was the scaffold with the highest frequency (29 compounds) in the entire set.

Conclusions and future directions

In this study we present a comprehensive chemoinformatic characterization of herein collected data set of small-molecule inhibitors of DNMT1 retrieved from public repositories. The chemical structures were characterized using complementary approaches. Analysis of the distributions of physicochemical properties indicated that, in general, compounds screened as inhibitors of DNMT1 are more polar than approved drugs, molecules in clinical trials, and screening compounds as measured by the distribution of HBA, HDB, and TPSA. Inhibitors of DNMT1 have similar flexibility as measured by RB. The visual representation of the property space revealed that compounds tested as DNMT1 inhibitors cover the traditional medicinal property space and there are molecules that expand the traditional space. The structural diversity of the databases computed with the Tanimoto coefficient and fingerprint representations revealed that the compounds tested and DNMT1 inhibitors are structurally diverse. In agreement with the diversity studies using different fingerprint representations it was concluded that the compounds focused on epigenetic targets are less diverse than the compounds approved as drugs, in clinical studies and general screening collection. The data set of approved drugs was the most diverse. The chemotype analysis of inhibitors of DNMT1 pointed to four specific molecular scaffolds that are potential ‘epigenetic privileged scaffolds’ and warrant the further acquisition and exploration of the local SAR to test this hypothesis. The four scaffolds are non-nucleosidic which is in agreement with the current trend to develop non-nucleoside inhibitors of DNMT1 as promising therapeutic agents or molecular probes. For representative compounds, including molecules with potential epigenetic privileged scaffolds, it remains to conduct, in addition to the present chemoinformatic analysis, comparative rigid and induce-fit docking (IFD) studies with the three-dimensional structure of

DNMT1. Indeed, IFD studies for selected inhibitors of DNMT1 have been reported aimed to explain, at the molecular level, large changes in the biological activity associated with small changes in the chemical structure.[54]

Acknowledgments

We thank Mark Johnson for providing MEQI. Eli Fernández-de Gortari is grateful to CONACyT for the fellowships granted No. 348291/240072. This work was supported by the National Autonomous University of Mexico (UNAM), grant PAIP 5000-9163 to JLMF.

References

1. K. D. Robertson, *Oncogene*, 2001, **20**, 3139-3155.
2. M. Rius and F. Lyko, *Oncogene*, 2012, **31**, 4257-4265.
3. J. L. Medina-Franco, J. Yoo and A. Dueñas-Gonzalez, in *Epigenetic Technological Applications*, ed. Y. G. Zheng, Elsevier, 2015, ch. 13, pp. 265-290.
4. D. Cheishvili, L. Boureau and M. Szyf, *Br. J. Pharmacol.*, 2015, **172**, 2705-2715.
5. J. P. J. Issa, H. M. Kantarjian and P. Kirkpatrick, *Nat. Rev. Drug Discovery*, 2005, **4**, 275-276.
6. A. Erdmann, L. Halby, J. Fahy and P. B. Arimondo, *J. Med. Chem.*, 2014, **58**, 2569–2583.
7. J. L. Medina-Franco, O. Méndez-Lucio, J. Yoo and A. Dueñas, *Drug Discovery Today*, 2015, **20**, 569-577.
8. S. Valente, Y. W. Liu, M. Schnekenburger, C. Zwergel, S. Cosconati, C. Gros, M. Tardugno, D. Labella, C. Florean, S. Minden, H. Hashimoto, Y. Q. Chan, X. Zhang, G. Kirsch, E. Novellino, P. B. Arimondo, E. Miele, E. Ferretti, A. Gulino, M. Diederich, X. D. Cheng and A. Mai, *J. Med. Chem.*, 2014, **57**, 701-713.
9. Z. Huang, H. Jiang, X. Liu, Y. Chen, J. Wong, Q. Wang, W. Huang, T. Shi and J. Zhang, *PLoS One*, 2012, **7**, e39917.

10. S. Loharch, I. Bhutani, K. Jain, P. Gupta, D. K. Sahoo and R. Parkesh, *Database*, 2015, **2015**.
11. F. López-Vallejo, M. A. Giulianotti, R. A. Houghten and J. L. Medina-Franco, *Drug Discovery Today*, 2012, **17**, 718-726.
12. F. Ntie-Kang, P. A. Onguene, M. Scharfe, L. C. O. Owono, E. Megnassan, L. M. Mbaze, W. Sippl and S. M. N. Efange, *RSC Adv.*, 2014, **4**, 409-419.
13. J. L. Medina-Franco, *Drug Dev. Res.*, 2012, **73**, 430-438.
14. J.-L. Reymond, *Acc. Chem. Res.*, 2015, **48**, 722-730.
15. J. L. Medina-Franco, K. Martínez-Mayorga, T. L. Peppard and A. Del Rio, *PLoS One*, 2012, **7**, e50798.
16. J. J. Naveja and J. L. Medina-Franco, *RSC Adv.*, 2015, **5**, 63882-63895.
17. 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.
18. 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.
19. J. Hu, S. Chen, X. Kong, K. Zhu, S. Cheng, M. Zheng, H. Jiang and C. Luo, *Curr. Med. Chem.*, 2015, **22**, 360-372.
20. J. Yoo, S. Choi and J. L. Medina-Franco, *PLoS One*, 2013, **8**, e62152.
21. J. Yoo and J. L. Medina-Franco, *Curr. Med. Chem.*, 2012, **19**, 3475-3487.
22. J. Yoo and J. L. Medina-Franco, *Comp. Mol. Biosci.*, 2011, **1**, 7-16.
23. A. Gaulton, L. J. Bellis, A. P. Bento, J. Chambers, M. Davies, A. Hersey, Y. Light, S. McGlinchey, D. Michalovich, B. Al-Lazikani and J. P. Overington, *Nucleic Acids Res.*, 2012, **40**, D1100-D1107.
24. T. Q. Liu, Y. M. Lin, X. Wen, R. N. Jorissen and M. K. Gilson, *Nucl. Acids Res.*, 2007, **35**, D198-D201.
25. D. Fourches, E. Muratov and A. Tropsha, *J. Chem. Inf. Model.*, 2010, **50**, 1189-1204.

26. Molecular Operating Environment (MOE), version 2010, Chemical Computing Group Inc., Montreal, Quebec, Canada. Available at http://www.chemcomp.com.
27. F. Mei, S. P. J. Fancy, Y.-A. A. Shen, J. Niu, C. Zhao, B. Presley, E. Miao, S. Lee, S. R. Mayoral, S. A. Redmond, A. Etxeberria, L. Xiao, R. J. M. Franklin, A. Green, S. L. Hauser and J. R. Chan, *Nat Med*, 2014, **20**, 954-960.
28. A. Moisan, Y.-K. Lee, J. D. Zhang, C. S. Hudak, C. A. Meyer, M. Prummer, S. Zoffmann, H. H. Truong, M. Ebeling, A. Kiialainen, R. Gérard, F. Xia, R. T. Schinzel, K. E. Amrein and C. A. Cowan, *Nat Cell Biol*, 2015, **17**, 57-67.
29. A. Bouslimani, L. M. Sanchez, N. Garg and P. C. Dorrestein, *Nat. Prod. Rep.*, 2014, **31**, 718-729.
30. C. Qin, C. Zhang, F. Zhu, F. Xu, S. Y. Chen, P. Zhang, Y. H. Li, S. Y. Yang, Y. Q. Wei, L. Tao and Y. Z. Chen, *Nucl. Acids Res.*, 2014, **42**, D1118-D1123.
31. V. Law, C. Knox, Y. Djoumbou, T. Jewison, A. C. Guo, Y. Liu, A. Maciejewski, D. Arndt, M. Wilson, V. Neveu, A. Tang, G. Gabriel, C. Ly, S. Adamjee, Z. T. Dame, B. Han, Y. Zhou and D. S. Wishart, *Nucl. Acids Res.*, 2014, **42**, D1091-D1097.
32. M. D. Shultz, *Bioorg. Med. Chem. Lett.*, 2013, **23**, 5980-5991.
33. H. Eckert and J. Bajorath, *Drug Discovery Today*, 2007, **12**, 225-233.
34. P. Willett, *Mol. Inf.*, 2014, **33**, 403-413.
35. V. Shanmugasundaram, G. M. Maggiora and M. S. Lajiness, *J. Med. Chem.*, 2005, **48**, 240-248.
36. N. Y. Mok, R. Brenk and N. Brown, *J. Chem Inf. Model.*, 2014, **54**, 79-85.
37. R Development Core Team. R Foundation for Statistical Computing, Vienna, Austria
38. T. Sander, J. Freyss, M. von Korff and C. Rufener, *J. Chem. Inf. Model.*, 2015, **55**, 460-473.
39. J. L. Medina-Franco, K. Martínez-Mayorga, A. Bender, R. M. Marín, M. A. Giulianotti, C. Pinilla and R. A. Houghten, *J. Chem. Inf. Model.*, 2009, **49**, 477-491.
40. P. Jaccard, *Bull. Soc. Vaudoise Sci. Nat.*, 1901, **37**, 547-579.
41. 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.

42. Y. J. Xu and M. Johnson, *J. Chem Inf. Comput. Sci.*, 2002, **42**, 912-926.

43. J. L. Medina-Franco, J. Petit and G. M. Maggiora, *Chem. Biol. Drug Des.*, 2006, **67**, 395-408.

44. 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.

45. J. Pérez-Villanueva, O. Méndez-Lucio, O. Soria-Arteche and J. Medina-Franco, *Mol. Div.*, 2015, in press. DOI: 10.1007/s11030-015-9609-z

46. J. L. Medina-Franco, K. Martínez-Mayorga, A. Bender and T. Scior, *QSAR Comb. Sci.*, 2009, **28**, 1551-1560.

47. 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.

48. A. M. Virshup, J. Contreras-García, P. Wipf, W. Yang and D. N. Beratan, *J. Am. Chem. Soc.*, 2013, **135**, 7296-7303.

49. R. P. Sheridan and S. K. Kearsley, *Drug Discovery Today*, 2002, **7**, 903-911.

50. J. J. Naveja and J. L. Medina-Franco, *Exp.Opin.Drug Discov.*, 2015, in press. DOI: 10.1517/17460441.2015.1073257

51. A. B. Yongye, J. Waddell and J. L. Medina-Franco, *Chem. Biol. Drug Des.*, 2012, **80**, 717-724.

52. Y. L. Wang, J. W. Xiao, T. O. Suzek, J. Zhang, J. Y. Wang, Z. G. Zhou, L. Y. Han, K. Karapetyan, S. Dracheva, B. A. Shoemaker, E. Bolton, A. Gindulyte and S. H. Bryant, *Nucleic Acids Res.*, 2012, **40**, D400-D412.

53. S. J. Chen, Y. L. Wang, W. Zhou, S. S. Li, J. L. Peng, Z. Shi, J. C. Hu, Y. C. Liu, H. Ding, Y. Y. Lin, L. J. Li, S. F. Cheng, J. Q. Liu, T. Lu, H. L. Jiang, B. Liu, M. Y. Zheng and C. Luo, *J. Med. Chem.*, 2014, **57**, 9028-9041.

54. J. Medina-Franco, O. Méndez-Lucio and J. Yoo, *Int. J. Mol. Sci.*, 2014, **15**, 3253-3261.

TABLES

Table 1. Sources of compounds to build the database of inhibitors of DNMT1 studied in this work

Source type	Source	Ref.	Number of compounds
Public database	Binding Database	[24]	265
	ChEMBL	[23]	163
	HEMD	[9]	96
Literature search	Web of Science	https://isiknowledge.com	42
TOTAL	Non-duplicate and curated set of DNMT1 inhibitors		566

Table 2. Reference compound collections considered in this study

Database	Source	Number of compounds
Approved drugs	Drug Bank	1,490
General screening collection	Selleck	1,100
Compounds in clinical trials	Therapeutic Target Database	837
Screening compounds focused of epigenetic targets	Selleck	113

Table 3. Loadings for the first four principal components (PC) of the property space.

	PC1	PC2	PC3	PC4
Cumulative eigenvalue %	53.86	77.31	89.35	95.11
HBA	0.106	-0.023	-0.134	0.155
HBD	0.134	-0.111	-0.280	-0.619
RB	0.051	0.083	0.220	-0.216
SlogP	-0.011	0.236	-0.213	-0.086
TPSA	0.005	-0.004	0.002	0.007
MW	0.002	0.002	0.000	0.001

Table 4. Measures of scaffold diversity of the compounds tested as inhibitors of DNMT1

	DNMT1 set
N	291
N/M	0.525
N_{sing}	170
N_{sing}/N	0.58
N_{sing}/M	0.30

FIGURES

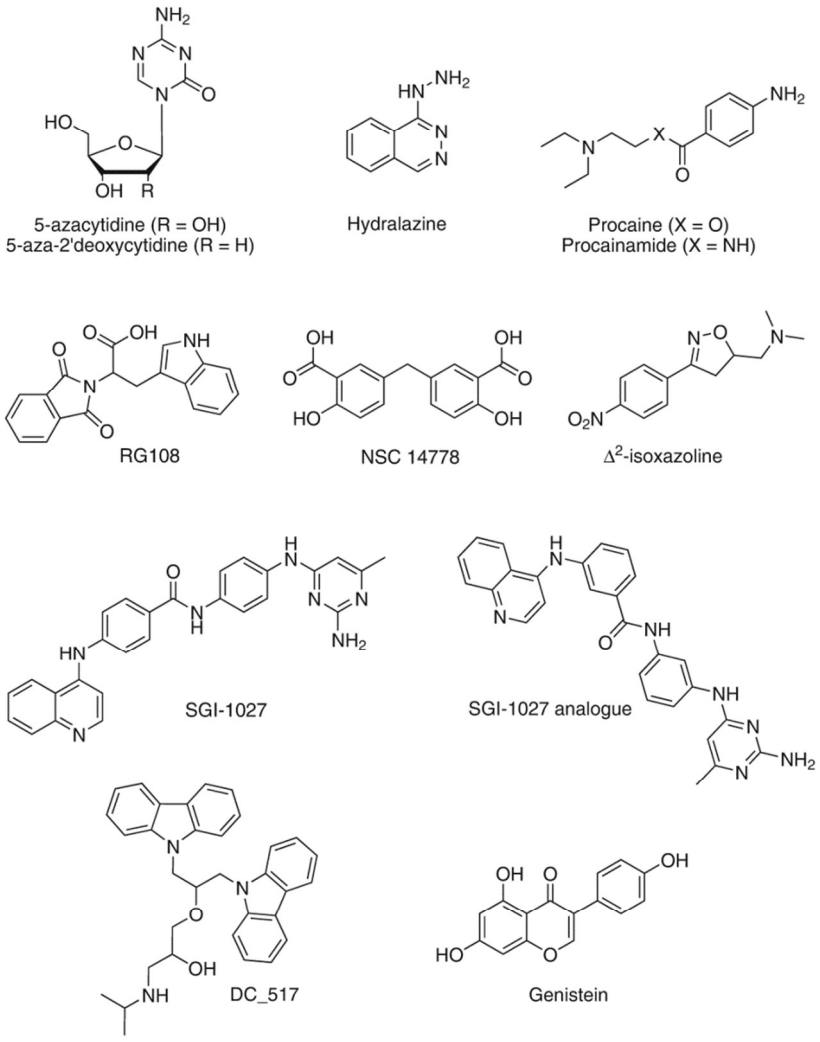
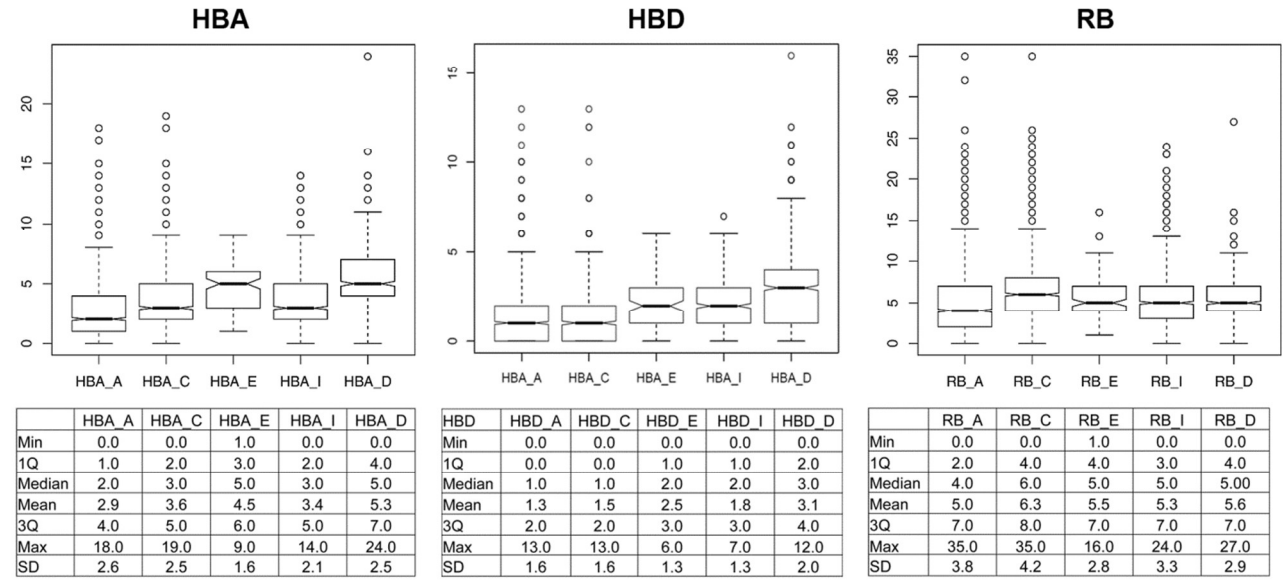


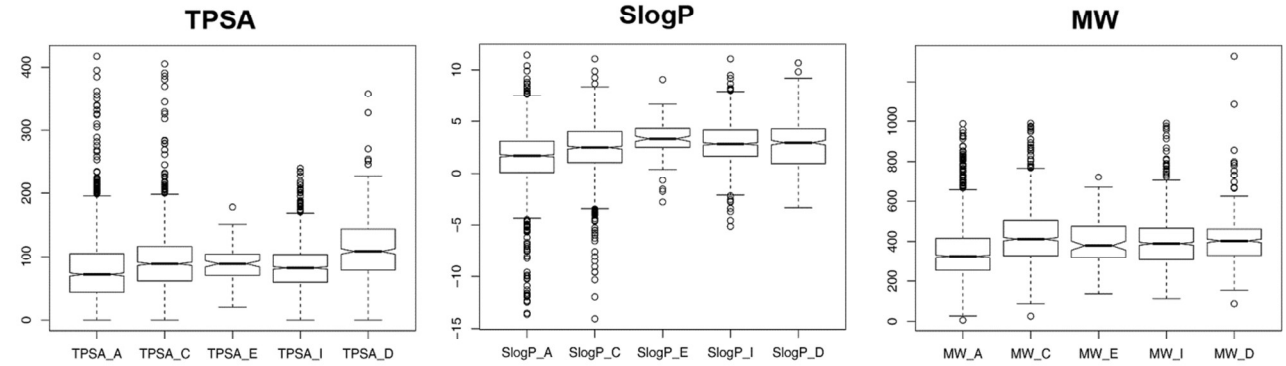
Figure 1. Chemical structures of representative inhibitors of DNMT1 and compounds associated with demethylating properties.



	HBA_A	HBA_C	HBA_E	HBA_I	HBA_D
Min	0.0	0.0	1.0	0.0	0.0
1Q	1.0	2.0	3.0	2.0	4.0
Median	2.0	3.0	5.0	3.0	5.0
Mean	2.9	3.6	4.5	3.4	5.3
3Q	4.0	5.0	6.0	5.0	7.0
Max	18.0	19.0	9.0	14.0	24.0
SD	2.6	2.5	1.6	2.1	2.5

HBD	HBD_A	HBD_C	HBD_E	HBD_I	HBD_D
Min	0.0	0.0	0.0	0.0	0.0
1Q	0.0	0.0	1.0	1.0	2.0
Median	1.0	1.0	2.0	2.0	3.0
Mean	1.3	1.5	2.5	1.8	3.1
3Q	2.0	2.0	3.0	3.0	4.0
Max	13.0	13.0	6.0	7.0	12.0
SD	1.6	1.6	1.3	1.3	2.0

	RB_A	RB_C	RB_E	RB_I	RB_D
Min	0.0	0.0	1.0	0.0	0.0
1Q	2.0	4.0	4.0	3.0	4.0
Median	4.0	6.0	5.0	5.0	5.00
Mean	5.0	6.3	5.5	5.3	5.6
3Q	7.0	8.0	7.0	7.0	7.0
Max	35.0	35.0	16.0	24.0	27.0
SD	3.8	4.2	2.8	3.3	2.9



	TPSA_A	TPSA_C	TPSA_E	TPSA_I	TPSA_D
Min	0.0	0.0	20.3	0.0	0.0
1Q	43.4	61.1	69.6	59.1	78.1
Median	71.2	88.0	88.0	81.4	108.6
Mean	81.9	95.0	89.0	84.1	111.7
3Q	104.9	116.3	104.3	104.0	143.9
Max	417.3	405.4	178.3	240.5	358.2
SD	55.9	52.5	29.2	37.3	47.1

	SlogP_A	SlogP_C	SlogP_E	SlogP_I	SlogP_D
Min	-13.6	-14.0	-2.8	-5.1	-3.4
1Q	0.1	1.0	2.5	1.6	1.0
Median	1.7	2.5	3.3	2.8	3.0
Mean	1.4	2.3	3.4	2.8	2.7
3Q	3.1	4.0	4.3	4.2	4.3
Max	11.4	11.0	9.1	11.0	10.7
SD	2.9	2.8	1.8	1.9	2.4

	MW_A	MW_C	MW_E	MW_I	MW_D
Min	6.9	26.3	136.2	112.2	88.1
1Q	253.3	328.5	320.4	306.4	329.8
Median	325.3	412.0	379.5	389.5	403.4
Mean	347.1	421.8	395.9	394.9	402.8
3Q	415.5	504.6	476.0	466.6	462.0
Max	988.2	990.2	721.9	990.2	1327.0
SD	148.6	152.0	111.4	124.5	118.2

Figure 2. Box notch plots of six physicochemical properties of five compound databases: approved drugs ('A'), compounds in clinical trials ('C'), screening compounds focused on epigenetic targets ('E'), general screening collection ('I'), and inhibitors of DNMT1 ('D'). The properties are hydrogen bond acceptors (HBA) and donors (HDB), rotatable bonds (RB), SlogP, topological polar surface area (TPSA), and molecular weight (MW). 1Q and 3Q, 1st and 3rd quartiles, respectively; SD, standard deviation.

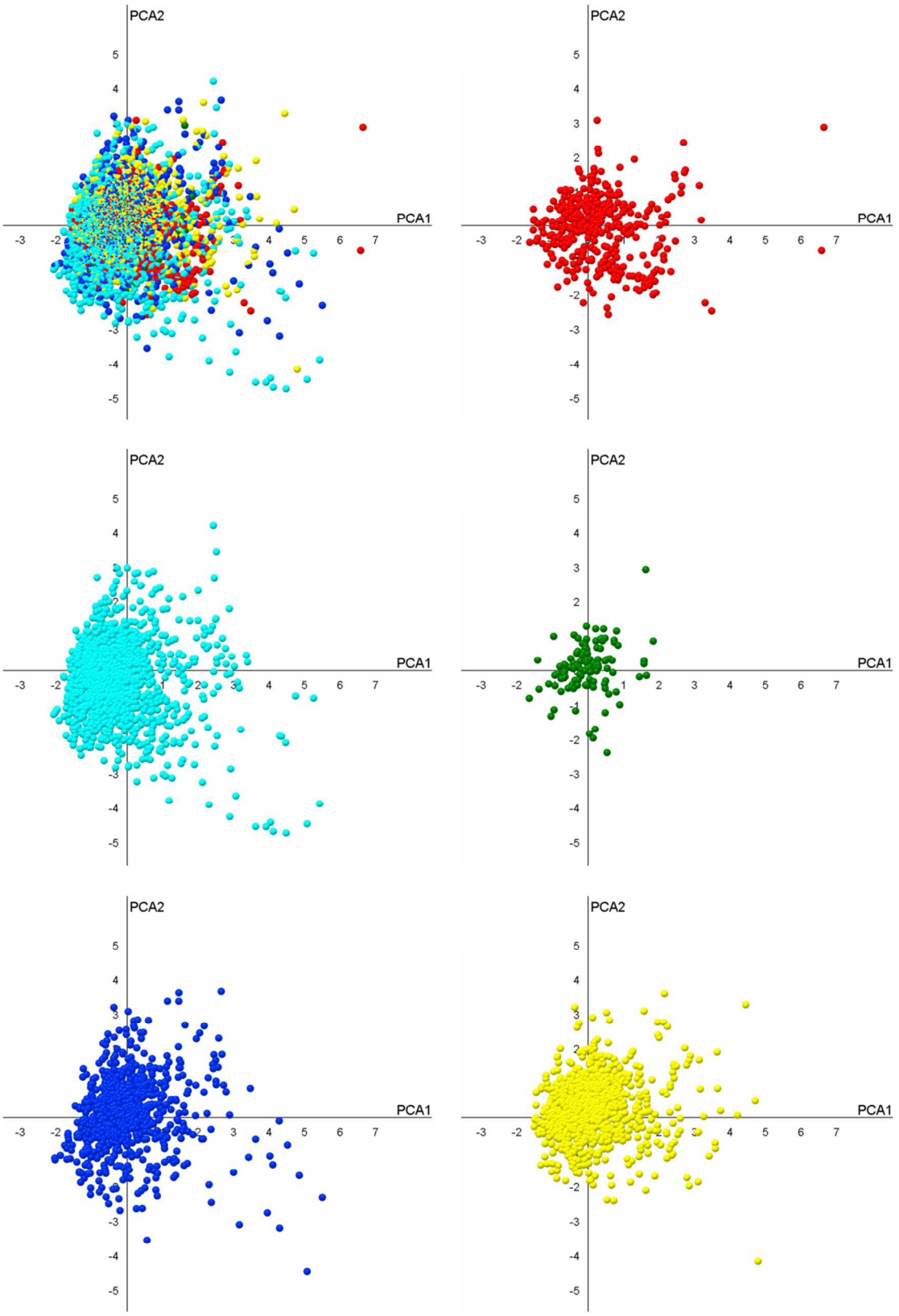


Figure 3. Property space of five data sets obtained by principal component analysis (PCA) of six physicochemical properties. The variance of the first two principal components is 77.3%. DNMT1 inhibitors (red); approved drugs (cyan); focused on epigenetic targets (green); clinical (blue); general screening (yellow). All plots are in the same coordinate system.

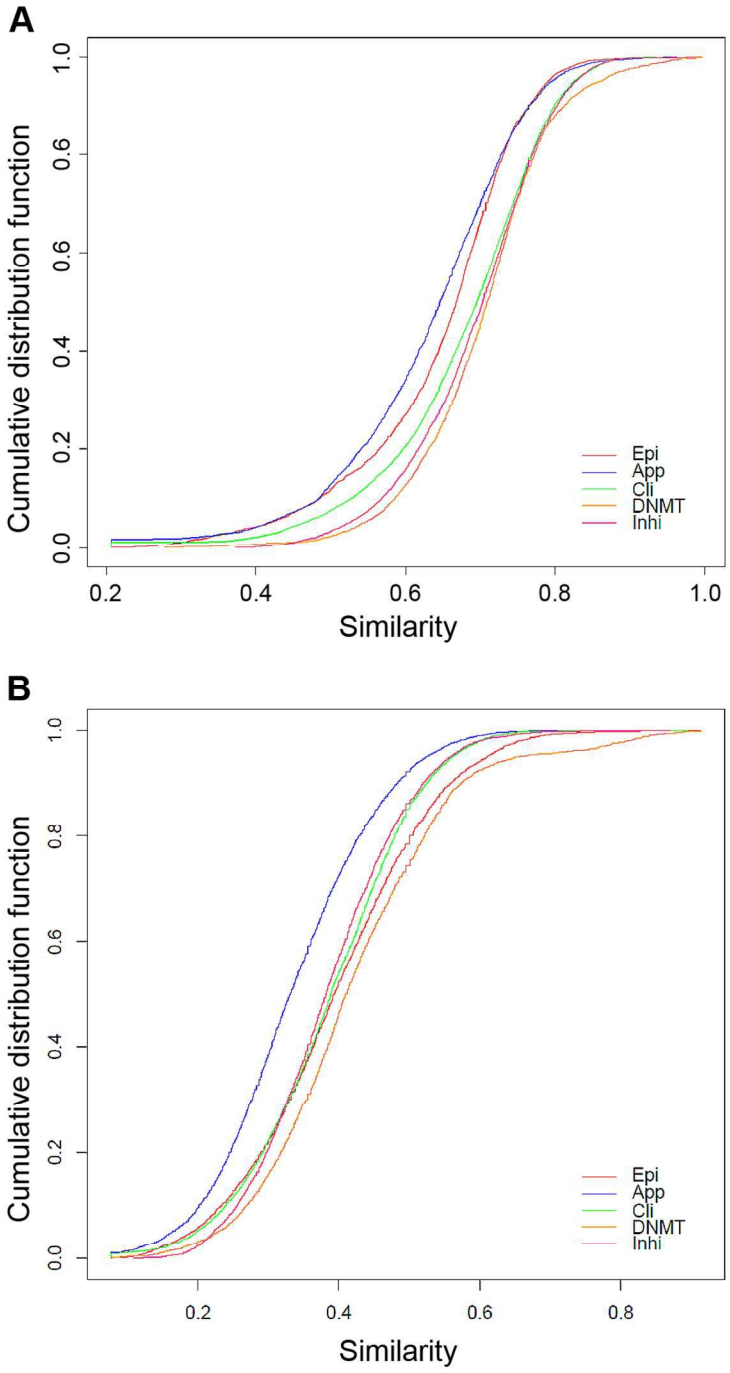


Figure 4. Cumulative Distributions Function (CDF) of pairwise Tanimoto similarity values computed for all data sets with A) TGD and B) MACCS keys fingerprints. Approved drugs ('App'), compounds in clinical trials ('Clin'), screening compounds focused on epigenetic targets ('Epi'), general screening collection ('Inhi'), and inhibitors of DNMT1 ('DNMT').

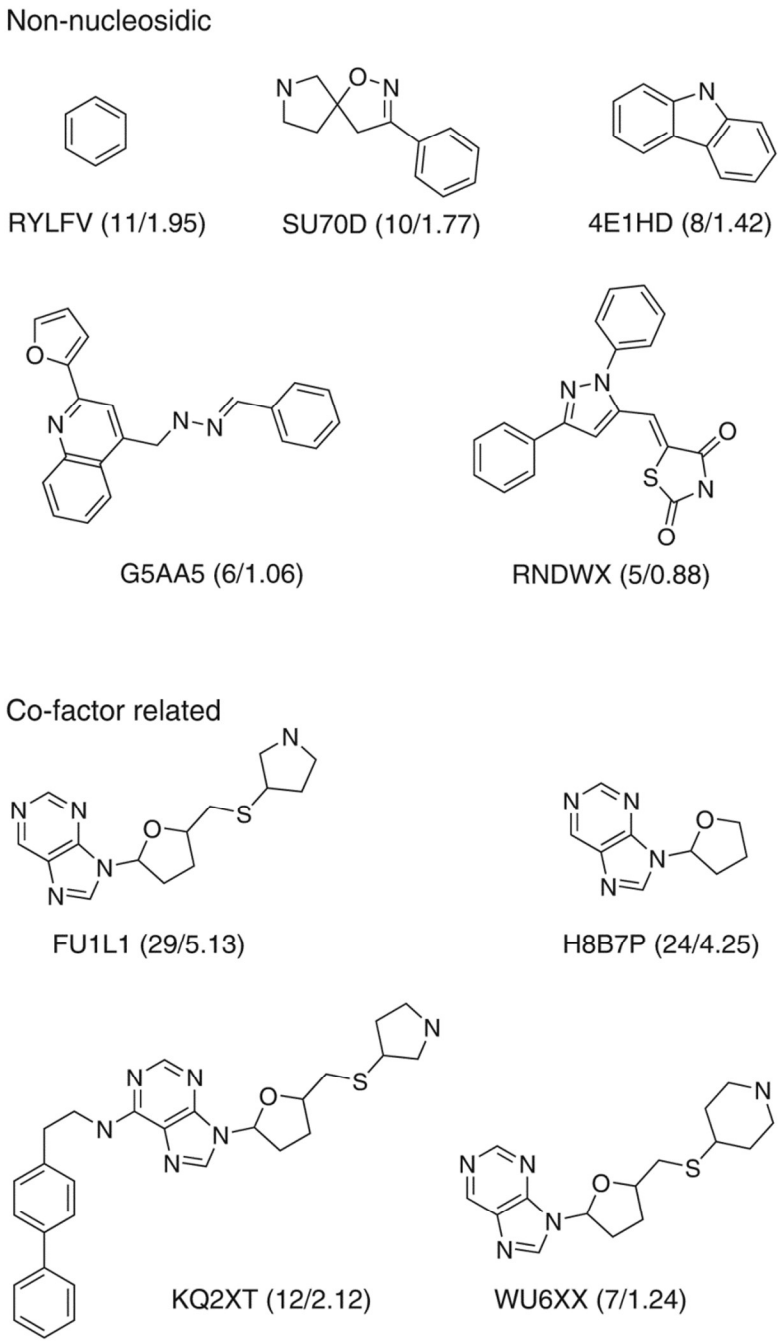


Figure 5. Most frequent cyclic systems identified in the data set of compounds tested as inhibitors of DNMT1. For each scaffold, the frequency and relative percentage are indicated in parenthesis.

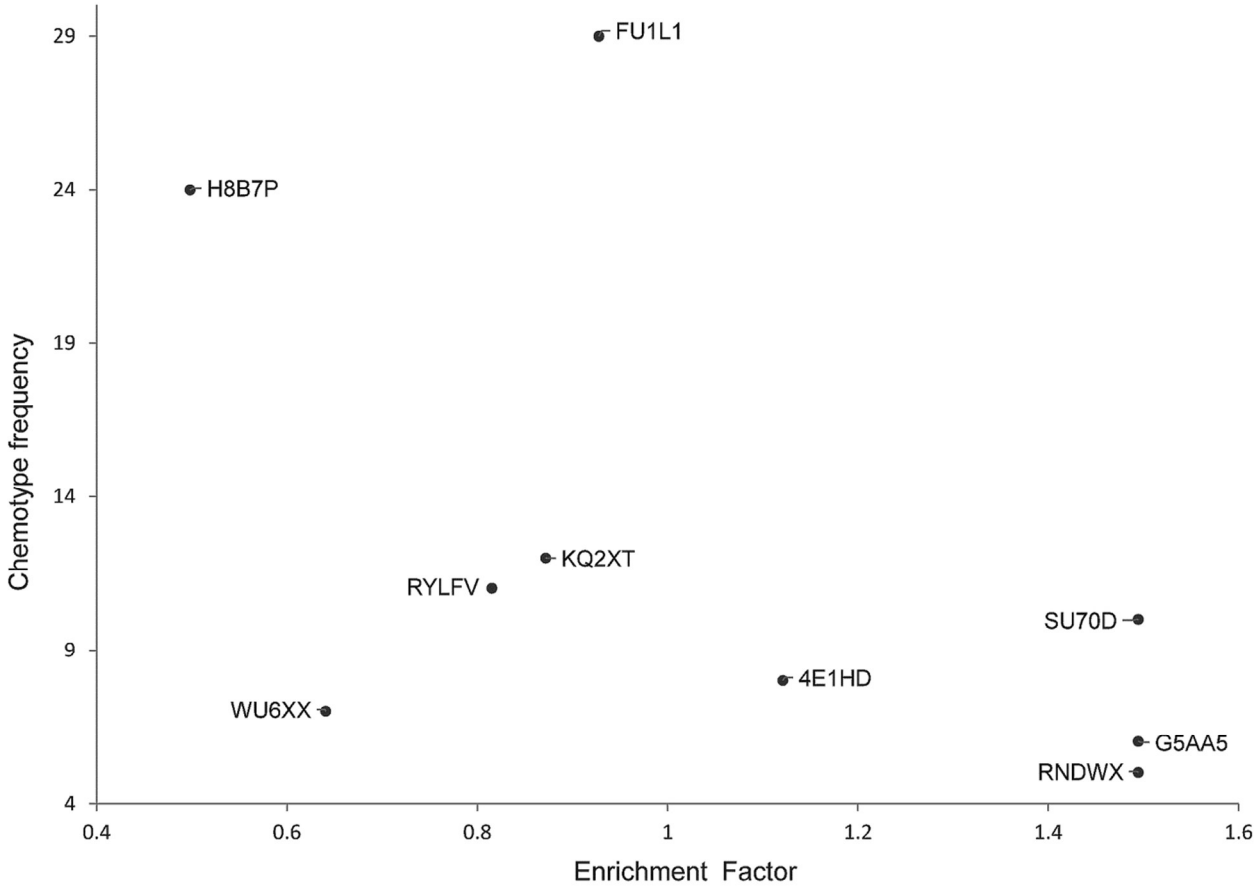


Figure 6. Chemotype enrichment plot for the nine most frequent cyclic systems in the data set of compounds tested as inhibitors of DNMT1. The chemical structures of the cyclic systems are shown in Figure 5.

For Tables of Contents

The first comprehensive exploration of the epigenetic relevant chemical space is reported in this work with a special emphasis on inhibitors of DNA methyltransferases.