ACS OMEGA

Open Access



http://pubs.acs.org/journal/acsodf

Article

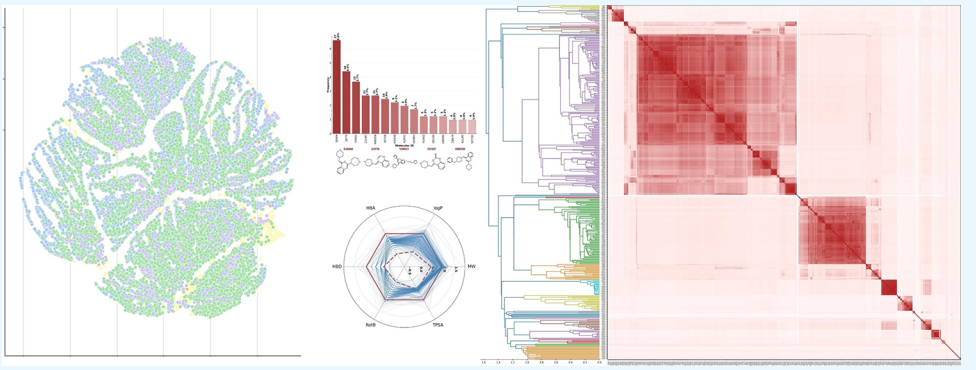
# Diversity and Chemical Space Characterization of Inhibitors of the Epigenetic Target G9a: A Chemoinformatics Approach

Raziel Cedillo-González and José L. Medina-Franco*

Cite This: *ACS Omega* 2023, 8, 30694−30704

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information



**ABSTRACT:** G9a is a histone-lysine methyltransferase that performs the mono- and dimethylation of lysine 9 at histone 3 of the nucleosome. It belongs to the SET PKMT family, and its methylations are related to promoter repression and activation. G9a is a promising epigenetic target. Despite the fact that there are several G9a inhibitors under development, there are no compounds in clinical use due to adverse in vivo ADMET (absorption, distribution, metabolism, excretion, and toxicity) issues. The goal of this study is to discuss the exploration, characterization, and analysis of the chemical space of 409 G9a inhibitors reported in a large public database. Exploring the chemical space of the inhibitors led to the quantification of their structural diversity based on molecular scaffolds and structural fingerprints of different designs. As part of the analysis, the G9a inhibitors were compared with commercial libraries focused on epigenetic targets. The findings of this work will help in the development of, in a follow-up study, predictive models to identify G9a inhibitors. This study also points out the relevance of screening commercial libraries to expand the epigenetic relevant chemical space, in particular, G9a inhibitors.

## 1. INTRODUCTION

The term epigenetic can be defined as the heritable changes in genetic information that occur without any change or mutation in the DNA sequence.[1] Epigenetics describes the mechanism of post-translational modifications in the chromatin and associated proteins through which the transcription is regulated.[2] These modifications have an important role in regulating the expression and transcription of genes involved in many human diseases, such as cancer,[3] addiction,[4] and psychiatric[5] and neurodegenerative disorders.[6] One of the most studied post-translational modifications is the covalent and reversible mono-, di-, or trimethylation of histone 3 (H3) and histone 4 (H4) at the definite $\varepsilon$-amino group of Lys and Arg and at the terminal imidazole moiety of His.[7] These reactions are catalyzed by the histone methyltransferases (HMTs) and are removed by histone demethylases (HDMs).[8] In particular, the euchromatic histone *N*-methyltransferase 2 (EHMT2) enzyme, also known as G9a or lysine methyltransferase 1C (KMT1C), contains a highly conserved catalytic SET domain that, in the presence of *S*-adenosyl-*L*-methionine (SAM), enables the selective mono-, di-, or trimethylation of lysine 9 in histone 3 (H3K9me1 and H3K9me2, respectively) in euchromatin,[9,10] leading to gene transcriptional repression (di- or trimethylation) or promoter activation (monomethylation).[11] Overexpression of G9a is linked to a variety of unusual biological processes as well as to disease onset, including oncogenic transformation, cancer metastasis,[12] HIV-1 latency, cognitive disturbances, and neurodegenerative disorders.[13] G9a inhibition is effective against the development of carcinogenic cells.[14,15]

Promising G9a inhibitors with a quinazoline scaffold have been reported, but none have passed clinical trials due to their poor pharmacokinetic properties. Therefore, it has been crucial to identify better and safer G9a inhibitors that may act as novel

Received: June 26, 2023
Accepted: August 2, 2023
Published: August 11, 2023

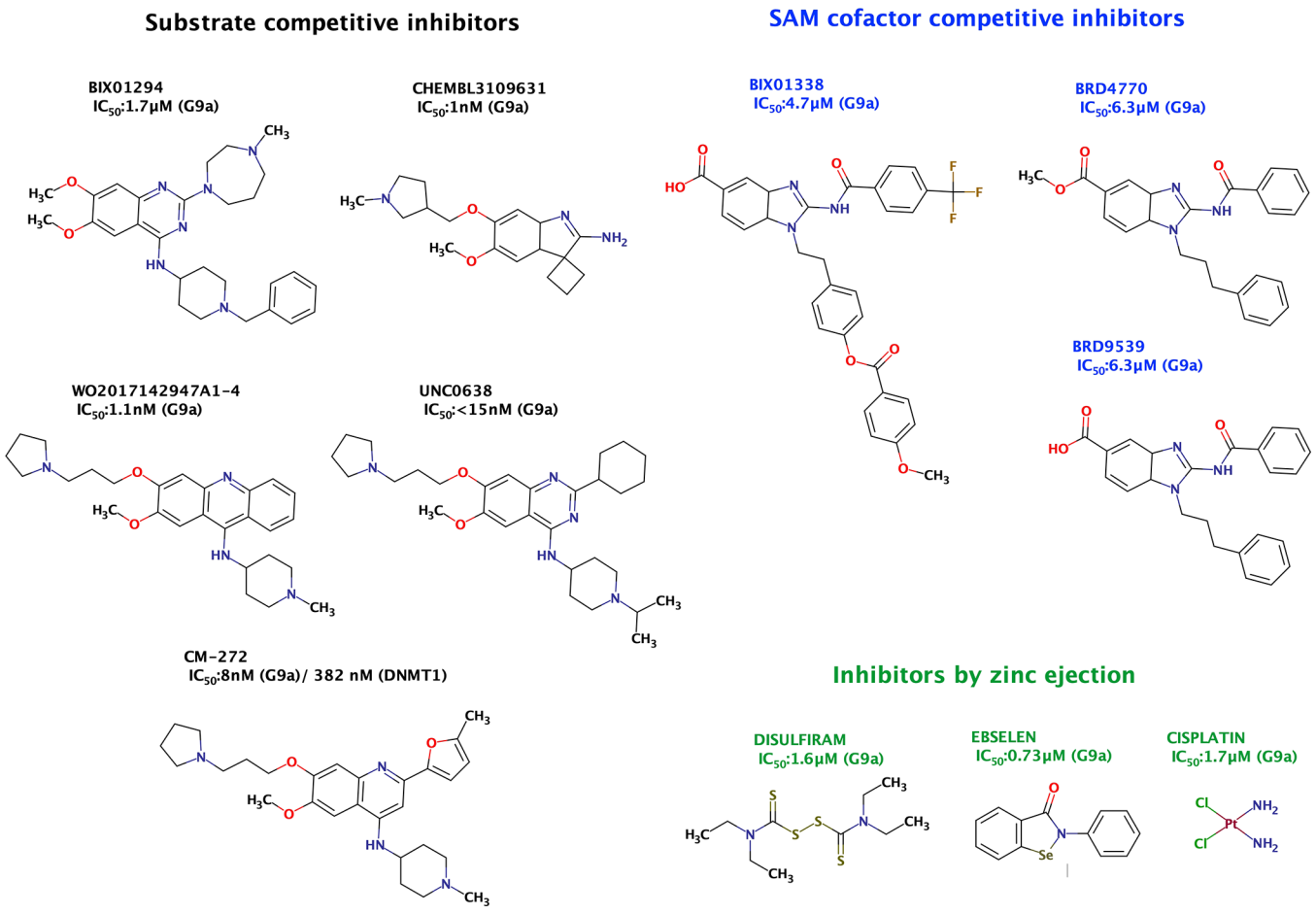



**Figure 1.** Representative compounds with inhibitory activity against G9a are classified into competitive substrate inhibitors (black), SAM cofactor competitive inhibitors (blue), and inhibitors by zinc ejection (green). CM-272 is a substrate-competitive dual inhibitor against the methyltransferase activity of G9a and DNMTs.

epigenetic therapeutic agents.[12] In addition to quinazoline-containing scaffolds, several other small molecules have been investigated as G9a inhibitors. At the time of writing, the latest version of the ChEMBL database (v. 32)[16] had 608 records related to the bioactivity of compounds tested against G9a. The data set was reduced after a curing and filtering process to 409 compounds that consider a $pCL_{50}$ range from 1.8 to 9.3. Notably, the main compounds in ChEMBL reported to be inhibiting the activity of G9a have been derived from quinolines and indoles.[15] In general, the inhibitors can be roughly classified into three main groups according to their mechanism of action (although there are compounds with unclear mechanism): 1) substrate-competitive inhibitors, 2) SAM cofactor competitive inhibitors, and 3) inhibitors by zinc ejection.[17] Figure 1 shows representative compounds with inhibitory activity against G9a. Of note, José-Enériz et al. designed and synthesized CM-272 (Figure 1), a novel, selective, and potent substrate-competitive dual inhibitor against methyltransferase activity of G9a and DNMTs. Interestingly, CM-272 can potentially inhibit DNMT1, DNMT3A, DNMT3B, and GLP.[18]

Structure−activity relationship (SAR) studies of G9a inhibitors have been explored using activity landscape modeling.[15,19,20] SAR analyses have led to the identification of critical substituents associated with the selectivity and efficacy of the compounds and to the identification of key protein−ligand interactions that drive the inhibition of G9a. However, the past SAR analyses were conducted about 5 years ago, and many more compounds have been tested. Moreover, there has not been a systematic analysis of the structural diversity and property profile of the current hundreds of small molecules tested as inhibitors of this promising epigenetic target and the recently developed commercial libraries focused on G9a.

The goal of this study was to rigorously characterize the chemical content, diversity, and drug-like properties of an epigenetic-focused library containing 409 compounds with reported G9a enzymatic inhibitory activity. To this end, we generated a compound library focused on G9a with the data reported in the latest release of ChEMBL.[16] The chemical space and diversity profile of the compounds tested with G9a were compared to those of several reference libraries, including three commercial screening libraries focused on G9a and a library of approved drugs. The study agrees with the proposal of Flores-Padilla et al. that showed that the content analysis of an epigenetic-focused library is informative before its screening (either computational or experimental) to uncover hits.[21] This work is part of our continued effort to chart the epigenetic-relevant chemical space (ERCS).[22]

## 2. METHODS

The chemoinformatics characterization of the G9a inhibitors was based on the following criteria: distribution of properties



https://doi.org/10.1021/acsomega.3c04566
ACS Omega 2023, 8, 30694−30704

of pharmaceutical relevance, molecular scaffolds (content and diversity), and fingerprint-based structural diversity. In addition to quantitative analysis, visual representations of the chemical space were generated. The data set of G9a inhibitors was compared to three epigenetic-focused libraries and drugs approved for clinical use. Details of each analysis are described hereunder.

**2.1. Data Sets.** This study was based on a curated data set of 409 compounds derived from quinazoline, quinoline, pyrimidine, indole, and purine with reported enzymatic inhibitory activity against G9a ($IC_{50}$). All compounds were retrieved from ChEMBL, release 32, and were categorized according to their $IC_{50}$ values as "active" ($IC_{50}$ < 10 $\mu$M), "inactive" ($IC_{50}$ > 20 $\mu$M), and "unknown" (10 $\mu$M < $IC_{50}$ < 20 $\mu$M) (Figure S1 in the Supporting Information). The SMILES representation of the structures and $pIC_{50}$ (−log $IC_{50}$) values are presented in Table S1 in the Supporting Information. The $pIC_{50}$ values ranged from 1.82 to 9.30.

The epigenetic-focused screening libraries were obtained from three chemical vendors: LifeChemicals,[23] Enamine,[24] and ChemDiv.[25] A set of drugs approved for clinical use was retrieved from DrugBank.[26] To identify how many unique compounds are in all data sets, we generated a "global" data set by putting together all compound data sets. The number of compounds in each data set considered in this work is shown in Table 1, which summarizes the data set name, source, and size (before and after curation). All compound data sets in this study underwent the same preparation process, including SMILES standardization and duplicate removal.

**Table 1. Summary of G9a inhibitors and Reference Compound Data Sets Considered in This Study**

| data set | size | | source |
|---|---|---|---|
| | initial | curated | |
| G9a | 609 | 409 | ChEMBL |
| ChemDiv | 25,883 | 25,883 | Epigenetics Focused Library 25,883 |
| Enamine | 10,560 | 10,542 | Histone Methyltransferase Library |
| LifeChemicals | 3578 | 1114 | Epigenetic Focused Library |
| FDA | 2747 | 2470 | Small molecules from DrugBank |

The data set of G9a inhibitors was analyzed and compared to the reference data sets based on physicochemical properties and scaffolds, as described in the next subsections.

**2.2. Properties of Pharmaceutical Relevance.** The following properties were computed with the RDkit 2023.03.01 library in Python:[27] molecular weight (MW), number of rotatable bonds (RB), hydrogen bond acceptors (HBA), hydrogen bond donors (HBD), topological polar surface area (TPSA), and the octanol/water partition coefficient (log *P*). The data distribution (DD) of each property was analyzed using box plots, and the correlation between the properties was evaluated with Pearson's correlation coefficient (PCC) using the *pearsonr* module from the library SciPy 1.10.1 (Virtanen et al. 2020). Visual representations of the chemical space were done with principal component analysis (PCA) and t-distributed stochastic neighbor embedding (t-SNE) based on the six physicochemical properties (vide supra). In the case of PCA and t-SNE, the variable reduction was generated through *PCA* and *TSNE* modules from the library scikit-learn 1.2.2.[28]

**2.3. Molecular Scaffolds.** We computed the Bemis and Murcko scaffold[29] of each compound in the G9a, FDA, and the three epigenetic-focused screening libraries. The calculation was done with the RDkit 2023.03.01 library in Python.[27] The total amount and frequency of each scaffold were analyzed. In addition to the two diversity measures mentioned above (number and frequency of scaffolds), the specific distribution of compounds in the *n* most populated scaffold was quantified with an entropy-based information metric, the Shannon Entropy (SE).[30,31] The SE of a population of *P* compounds contained in the *n* scaffold is defined as

$$\mathrm{SE} = -\sum_{i=1}^{n} p_i \log_2 p_i,\ p_i = \frac{c_i}{P} \quad (1)$$

where $p_i$ is the relative frequency of the scaffold *i* in a population of *P* compounds containing a total of *n* distinct scaffolds; $c_i$ is the absolute number of molecules containing a particular scaffold *i*. The values of SE range between 0 and $\log_2 n$ and hence depend on *n* but not explicitly on *P*. If SE = 0, then all compounds possess only a single scaffold. If SE = $\log_2 n$, then the *P* compounds are uniformly distributed among the *n* scaffolds, which represents the maximum scaffold diversity. With the purpose of normalizing the SE values for each data set, the scaled SE (SSE) is defined as[30]

$$\mathrm{SSE} = \frac{\mathrm{SE}}{\log_2 n} \quad (2)$$

The values of the SSE range between 0 and 1.0, where all *P* compounds are contained in one scaffold or each scaffold contains an equal number of compounds, respectively. When the SSE values are closer to 1.0, this indicates a large scaffold diversity within the *n* most populated scaffold.

**2.4. Structural Fingerprints.** The fingerprint-based diversity of the compounds in the G9a data set and reference libraries was computed with the RDkit topological fingerprint,[27] extended connectivity fingerprint diameter 4 (ECFP4), and MACCS keys (166-bits) with the Tanimoto coefficient. The fingerprint-based structural diversity of the compounds in the G9a data set was analyzed by means of the similarity matrices and the cumulative distribution functions of the pairwise similarity values. The visualization of the similarity matrix was carried out in Python software v 3.11.3,[32] using a dendrogram and a heatmap plot from *matplotlib* and *scipy* packages. The hierarchical clustering was carried out considering a single linkage, and the delimitation of each cluster was established with a cutoff point of 0.7.

**2.5. Analysis and Visualization of Chemical Space.** The chemical space of the six data sets (including, in some cases, a global data set for visual purposes) was analyzed by DD, PCC, PCA, and t-SNE based on the six molecular properties of pharmacological interest described in Section 2.2 (HBA, HBD, RB, TPSA, log *P*, and MW). The modules and packages used for these analyses were carried out in Python v 3.11.3,[32] and the main tools for visualization were *matplotlib* and *seaborn* packages.

The DD was performed with the module *describe* from Python, and the visualization was done through two plots; the first was a boxplot showing the statistical distribution of the data sets for each property, and the second was a pairplot plotting the pairwise relationships by each property of the data sets. These visualizations were made using the submodules *boxplot* and *pairplot* from the *seaborn* package.

In order to define the correlation between the six properties for each data set, the *corr* submodule of the *pandas* package


https://doi.org/10.1021/acsomega.3c04566
ACS Omega 2023, 8, 30694−30704

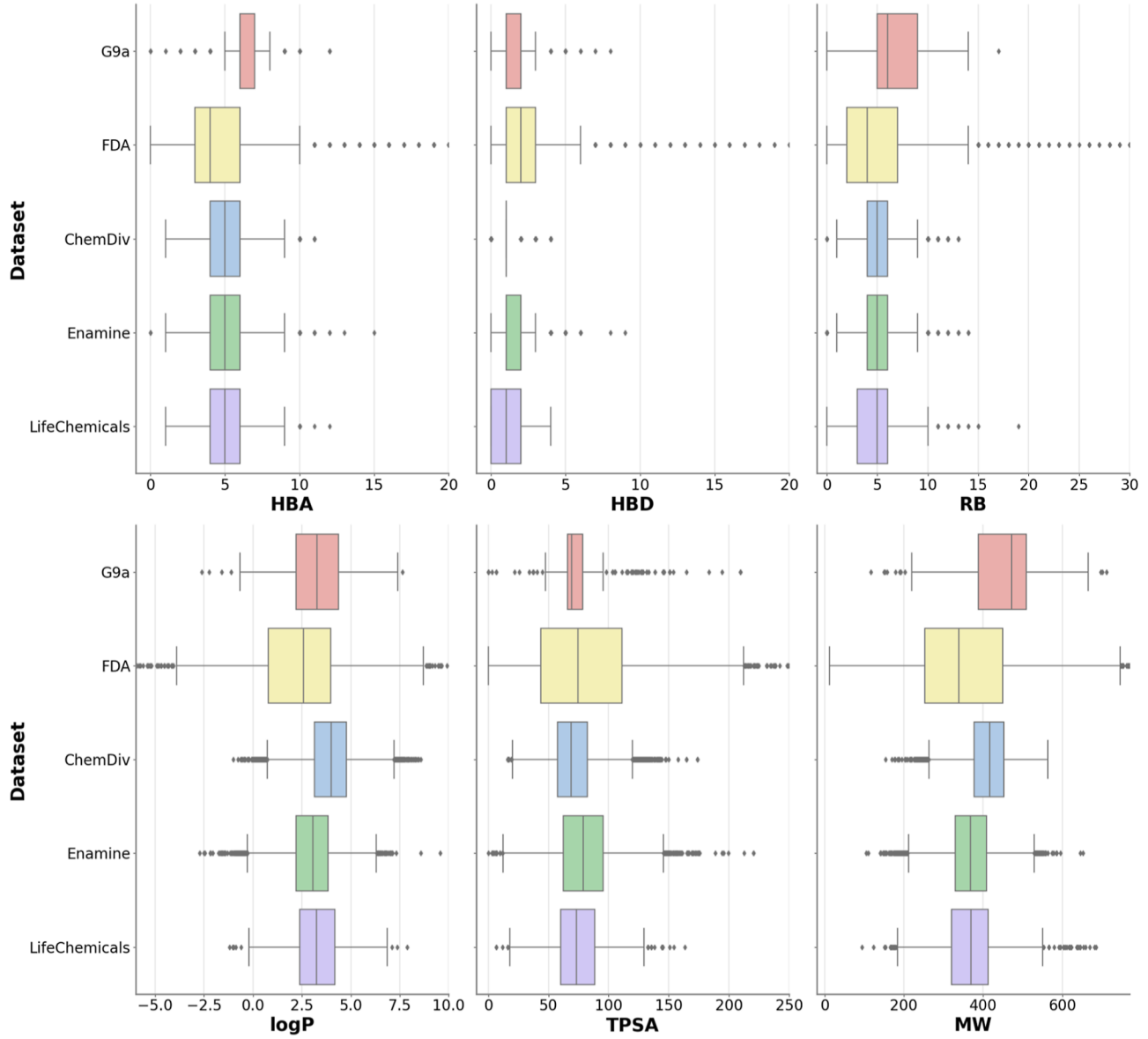



**Figure 2.** Box plots of the six properties of pharmaceutical relevance for all data sets considered in this study. G9a inhibitors are shown in red, FDA in yellow, ChemDiv in blue, Enamine in green, and LifeChemicals in purple.

was used for PCC, and it was charted through a heatmap using the submodule heatmap from the *seaborn* package. For PCA and t-SNE, the values for each property were normalized across compounds before the reduction of dimensions. For both PCA and t-SNE, the number of components considered to reduce the dimensionality of the data set was two components. The *scatterplot* submodule from *seaborn* was used to plot the chart of PCA. For t-SNE, the perplexity and the number of iterations were set at 40 and 300, respectively.

## 3. RESULTS AND DISCUSSION

Identifying inhibitors of epigenetic targets, including G9a, is an area of active research area. Screening compound libraries and optimization of hit and lead compounds have led to the population of ERCS.[33] Remarkably, more and more chemical libraries have been experimentally tested, and the information has been deposited in public libraries such as ChEMBL and other chemogenomic databases.[34] Consequently, the SAR or, more specifically, the structure−epigenetic activity relationship (SEARs) have increased, paving the way for developing predictive models. Similarly, several drug discovery strategies are being successfully implemented and developed to augment the ERCS,[22] including developing screening libraries focused on epigenetic targets.

**3.1. Data Set.** Most of the compounds in the G9a inhibitors data set reported to date in ChEMBL (88%) have relatively low $IC_{50}$ values (<10 $\mu$M) (Figure S1 in the Supporting Information). This observation agrees with the current trend in public repositories to disclose "positive" (aka, "active" compounds). Looking forward to developing robust predictive models that do not rely on decoy data sets, there is a need for the scientific community to disclose "negative" data (i.e., inactive compounds).[35]

**3.2. Properties of Pharmaceutical Relevance.** The drug-likeness definition varies depending on the empirical rules employed within that concept, which has evolved with time. Developed in 1997, Lipinski's rule of five (Ro5) was among the first set of quantitative parameters rules that defined the "drug-


https://doi.org/10.1021/acsomega.3c04566
ACS Omega 2023, 8, 30694−30704

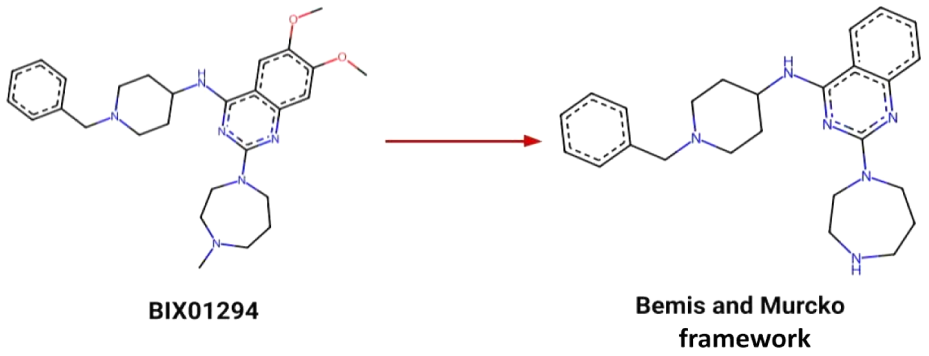



**Figure 3.** Definition of the scaffold used in this study. The scaffold is obtained after iteratively removing the side chains from the entire molecule. For example, the Bemis−Murcko framework of the G9a inhibitor BIX01294 is shown.

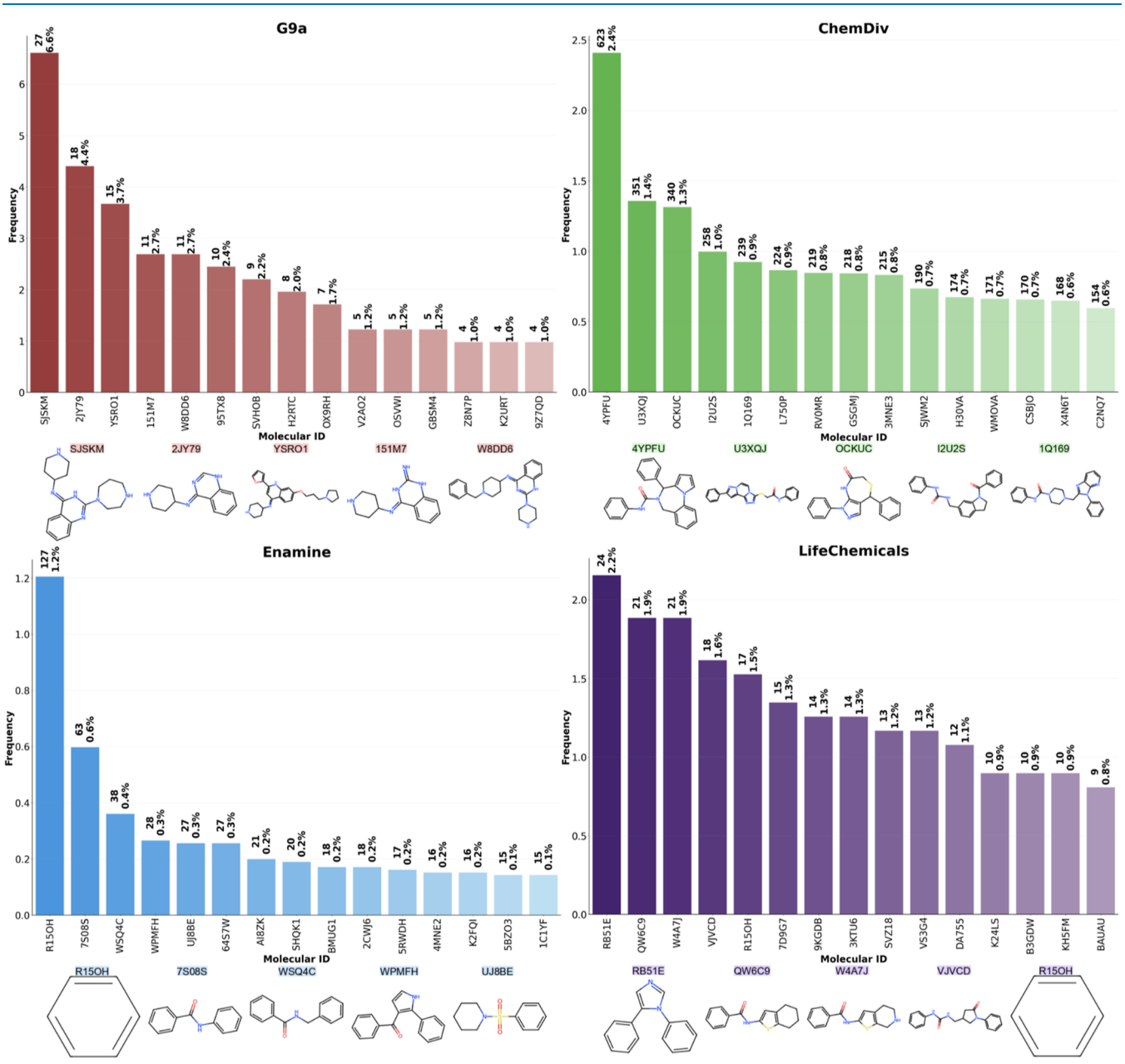



**Figure 4.** Distribution of the 15 most populated scaffolds of the G9a data set and the 3 commercial, focused libraries. The chemical structures of the 5 most frequent scaffolds per data set are shown. The color code is the same as in Figure 2.



https://doi.org/10.1021/acsomega.3c04566
ACS Omega 2023, 8, 30694−30704

Table 2. Summary of the Scaffold Diversity of the Data Sets Studied in This Work[a]

| data set | M | N | N/M | M15 | N15 | N15/M15 | SE15 | SSE15 |
|---|---|---|---|---|---|---|---|---|
| G9a | 409 | 239 | 0.584 | 143 | 15 | 0.105 | 3.647 | 0.933 |
| ChemDiv | 10,542 | 4421 | 0.419 | 3714 | 15 | 0.004 | 3.781 | 0.977 |
| Enamine | 25,883 | 7543 | 0.288 | 466 | 15 | 0.032 | 3.503 | 0.897 |
| LifeChemicals | 1114 | 504 | 0.452 | 221 | 15 | 0.068 | 3.843 | 0.983 |
| FDA | 2470 | 1364 | 0.552 | 674 | 15 | 0.022 | 2.496 | 0.639 |

[a]M: Number of molecules; N: Number of scaffolds; M15: number of molecules contained in the 15 most populated scaffolds; N15: 15 most populated scaffolds; SE: Shannon Entropy; SSE: Scaled Shannon Entropy. The variables with the number 15 correspond to a value of $n = 15$.

likeness" concept.[36] The Ro5 considers a MW less than 500 g/mol, not more than five HBD, not more than ten HBA, and logP less than five.[37] Another widely used set of empirical rules was proposed by Veber et al., who suggested that polarity and molecular flexibility could largely control drug bioavailability independent of MW. Veber's rules consider not more than ten RB and a TPSA greater than 140 Å. It is worth mentioning that if a compound that violates more than one of these rules is flagged as a "high-risk" compound in terms of oral bioavailability[36] although it is known that the "drug-likeness" rules are meant to be used as a guide but not as hard rules.[38] Indeed, Ro5[37] and Veber's rule[39] are often misused. They were originally conceived to aid the development of orally bioavailable drugs and were not designed to guide the medicinal chemistry development of all small-molecule drugs. Oral administration is a desirable objective for the treatment of several diseases like cancer or diabetes, but it is not an absolute requirement.[40]

To compare the G9a data set with the three epigenetic-focused libraries and with an approved drugs data set, we selected six properties of pharmaceutical relevance considered by Lipinski and Veber (MW, HBD, HBA, log *P*, RB, and TPSA). We employed these descriptors to explore the drug-likeness profile of the data sets because they are well-accepted parameters and broadly used in chemoinformatics characterization of compound data sets. Figure 2 shows box plots summarizing the distribution of each of the six properties. Analysis of the figure indicates that the epigenetic-focused libraries and G9a data set have, in general, drug-like properties. Compared to the other data sets analyzed in this work, the compounds tested with G9a have a slightly higher MW and are more flexible, as measured by the number of RB.

The radar plots in Figure S2 of the Supporting Information reinforce the above analysis. This figure shows that most of the compounds in the G9a data set and in the epigenetic-focused libraries fulfill the Ro5 and Vever's rules with a few exceptions. Most of these exceptions are related to logP values. As expected, most compounds in the FDA data set comply with these limits (e.g., the FDA set is included as a general reference in this work). Also, as anticipated, there are well-known exceptions to the Ro5 because, for instance, not all drugs are administered orally. It is important to remark that Ro5 violations do not necessarily indicate a lack of efficacy or safety in a drug.[38,41]

Figure S3 in the Supporting Information shows the correlation coefficient considering the six properties of major pharmaceutical interest for each data set. The PCC of the focused libraries showed that the negative and positive correlations of highest significance were log *P*-TPSA, log *P*-HBD, and log *P*-HBA (negative) and HBA-MW and HBA-TPSA (positive). For approved drugs, the negative and positive correlations of significance are log *P*-HBD and log *P*-TPSA (negative) and TPSA-HBA and TPSA-HBD (positive). Finally, in data set G9a, the negative and positive correlations are log *P*-HBD and log *P*-TPSA (negative) and TPSA-HBD (positive). Overall, these correlations suggest that in the focused libraries, approved drugs, and G9a data set, lipophilicity (as measured by logP) negatively correlates with properties related to hydrogen bonding capabilities (HBD and HBA) and TPSA. Additionally, TPSA shows positive correlations with HBA and HBD. These findings provide insights into the relationships between various molecular properties in the data sets studied.

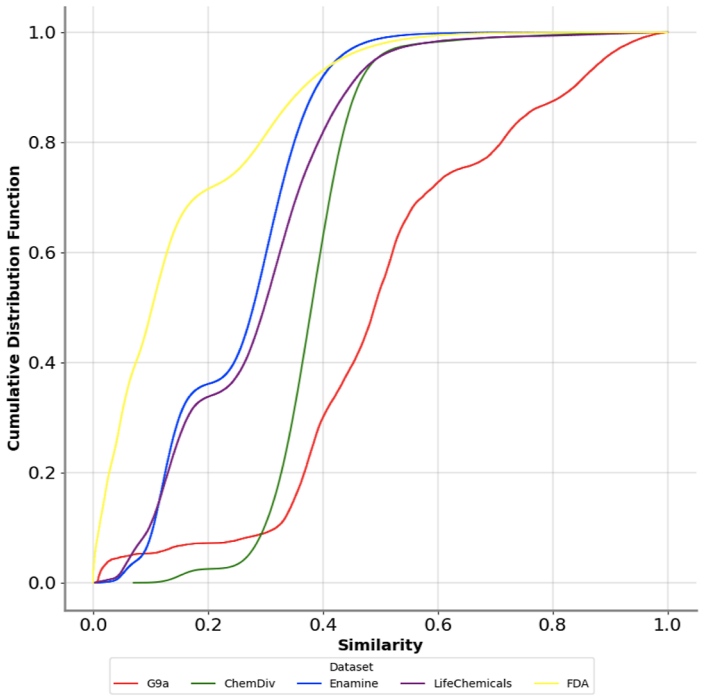
**3.3. Scaffold Analysis.** *3.3.1. Content.* The term molecular scaffold describes the core structure of a molecule.[42] There are different ways to computationally derive the scaffold of a molecule systematically and consistently that have been reported and reviewed elsewhere.[30,42] This study defined the scaffolds as the union ring systems and linkers in a molecule, also known as the atomic framework of Bemis and Murcko[29] (Figure 3).

Figure 4 shows the distribution of the 15 most populated scaffolds in the G9a set, the focused libraries, and the FDA data set. The figure indicates that, other than benzene, which is ubiquitous in bioactive compounds data sets and screening collections,[43] the most frequent scaffolds in the G9a data set correspond to derivatives of quinazoline (10 of 15) and quinoline (2 of 15), as is reported in the literature.[15] The other two frequent scaffolds are pyrimidine (1 of 15) and purine (1 of 15) [see Figures S5 and S6 of the Supporting Information].

In the focused libraries, it is worth highlighting that among the most frequent scaffolds, there are no derivatives of quinazoline or quinoline, enabling the focused libraries to explore other potential sections and scaffolds in the chemical space. There are no shared scaffolds among the 15 most frequent scaffolds in the epigenetics-focused libraries. Still, the Enamine and LifeChemicals data set has derivatives of benzamides between the top three scaffolds. The three most frequent scaffolds for the focused libraries are (ordered by frequency, scaffold ID is shown in parentheses, Figure 4): ChemDiv: *N*,4-diphenyl-4,6-dihydrobenzo[*c*]cyclopenta[*e*]-azepine-5(3H)-carboxamide (4YPFU) > *N*-phenyl-2-((9-phenylpyrazolo[1,5-*a*][1,2,4]triazolo[3,4-*c*]pyrazin-3-yl)thio)-acetamide (U3XQJ) > 1,4-diphenyl-4,8-dihydro-1*H*-pyrazolo-[3,4-*e*][1,4]thiazepin-7(6*H*)-one (OCKUC); Enamine: *N*-phenylbenzamide (7S08S) > *N*-benzylbenzamide (WSQ4C) > phenyl(2-phenyl-1*H*-pyrrol-3-yl)methanone (WPMFH); LifeChemicals: 1,5-diphenyl-1*H*-imidazole (RB51E) > *N*-(4,5,6,7-tetrahydrobenzo[*b*]thiophen-2-yl)benzamide (QW6C9) > *N*-(4,5,6,7-tetrahydrothieno[2,3-*c*]yridine-2-yl)-benzamide (W4A7J).

*3.3.2. Diversity.* The scaffold diversity of four data sets of active compounds and compounds registered in focused libraries targeting G9a (Table 1) was assessed using frequency,


https://doi.org/10.1021/acsomega.3c04566
ACS Omega 2023, 8, 30694−30704




| Dataset | ECFP4 | | | Topologic | | | MACCS keys (166-bits) | | |
|---|---|---|---|---|---|---|---|---|---|
| | Mean | Median | Stdev | Mean | Median | Stdev | Mean | Median | Stdev |
| G9a | 0.289 | 0.275 | 0.178 | 0.509 | 0.490 | 0.213 | 0.666 | 0.704 | 0.200 |
| ChemDiv | 0.172 | 0.159 | 0.069 | 0.381 | 0.379 | 0.084 | 0.514 | 0.511 | 0.113 |
| Enamine | 0.143 | 0.138 | 0.046 | 0.253 | 0.278 | 0.115 | 0.434 | 0.432 | 0.107 |
| LifeChemicals | 0.153 | 0.141 | 0.070 | 0.280 | 0.298 | 0.144 | 0.445 | 0.440 | 0.123 |
| FDA | 0.096 | 0.094 | 0.051 | 0.150 | 0.103 | 0.143 | 0.301 | 0.296 | 0.143 |

**Figure 5.** Cumulative distribution function of the pairwise similarity values of the G9a data set and reference compound collections considered in this study. The similarity was computed with the RDkit topological fingerprint and the Tanimoto coefficient. The table shows the summary statistics of the similarity values computed with the Tanimoto coefficient and other fingerprints with different designs.

SSE values, and other well-established metrics used to analyze the scaffold diversity quantitatively, as detailed in the Methods section. SSE was computed to quantify the diversity of the most populated scaffolds. SSE values closer to 1 indicate that compounds are uniformly distributed in different scaffolds, i.e., maximum diversity. If SSE is closer to 0, all of the compounds share the same scaffold, i.e., less diversity. Table 2 summarizes the number of molecules (M), the number of scaffolds (N), and the fraction of scaffolds relative to the number of molecules in the data set (N/M) in each data set. The table also summarizes the SSE values for each data set's 15 most populated scaffolds. The results indicate that, overall, the G9a data set and LifeChemicals have the largest diversity as measured by the total fraction of scaffolds (N/M) and SSE15. Regarding the focused library studies, ChemDiv is the second most diverse, followed by Enamine (the latter with the lowest fraction of total scaffolds and SSE15 values of 0.288 and 0.897, respectively). As a reference, the total scaffold diversity of the FDA set is also quite large (N/M = 0.552). However, the diversity of the 15 most populated scaffolds is not quite large (SSE15 = 0.639) due to the large fraction of compounds with the benzene scaffold (11%), followed by the compounds with a steroidal scaffold (9.2%), as shown in Figure S5 in the Supporting Information.

Taken together, these results point to the large diversity of scaffolds tested with G9a activity. Out of the focused libraries, LifeChemicals is the most suitable for finding novel scaffolds and can be prioritized in a virtual screening campaign to identify G9a inhibitors.

**3.4. Fingerprint-Based Diversity.** Figure 5 shows the cumulative distribution functions of the pairwise similarity values computed with the Tanimoto coefficient and three fingerprints of different designs, namely, RDkit topological fingerprint, ECFP4, and MACCS keys (166-bits). The figure also shows the summary statistics of the similarity values. A high value of the Tanimoto coefficient (close to one) indicates high structure similarity (based on that particular fingerprint) and hence a low diversity. The cumulative distribution function and summary statistics of the pairwise similarity values for each data set computed with all three fingerprints indicated that the G9a data set is the least diverse (higher similarity values) compared to the other focused libraries and the FDA data set. Regarding the focused libraries, Figure 5 shows that the ChemDiv data set is less diverse than LifeChemicals. The latter has diversity similar to that of


https://doi.org/10.1021/acsomega.3c04566
ACS Omega 2023, 8, 30694−30704

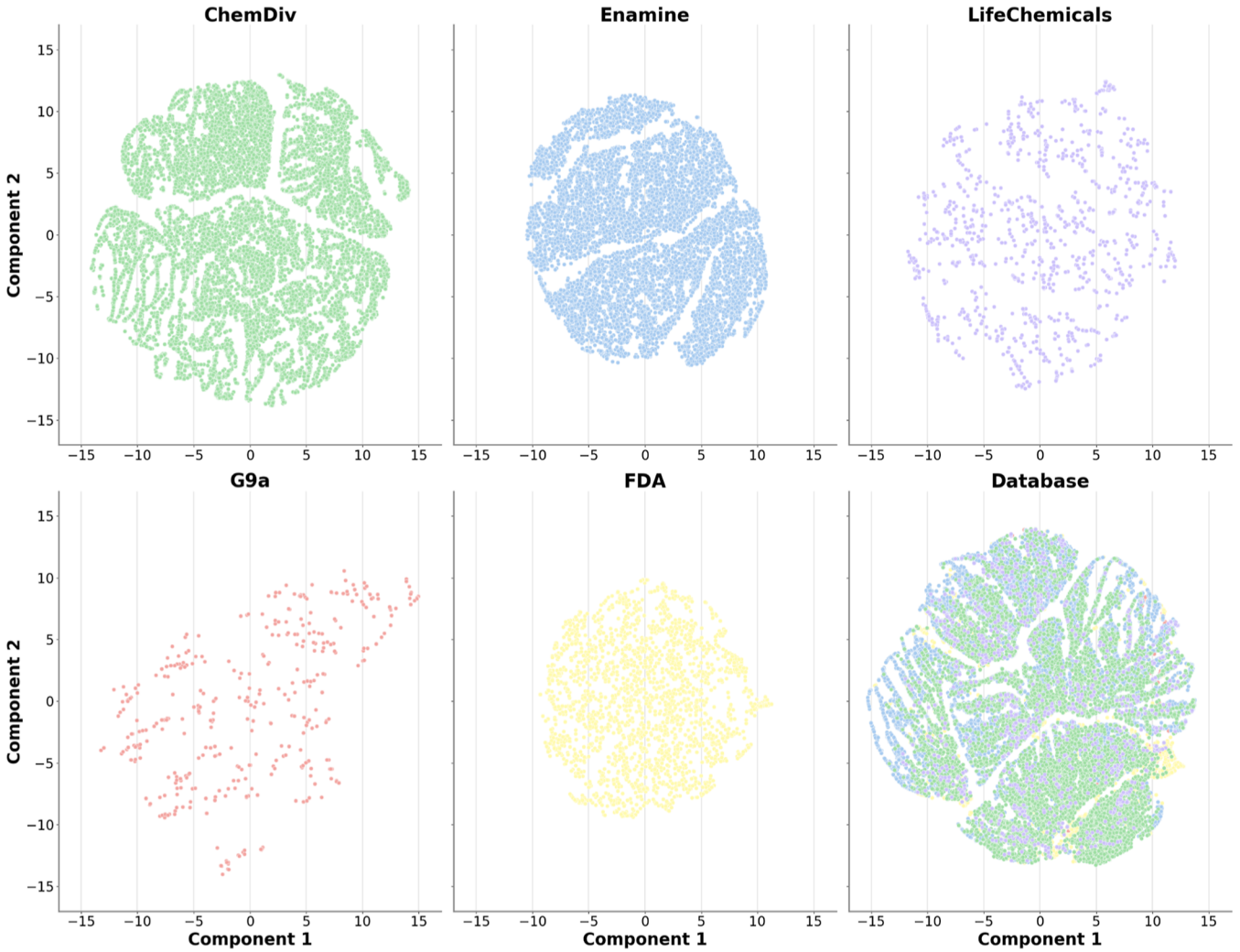



**Figure 6.** Visual representation of the chemical space of the five data sets with a t-distributed stochastic neighbor embedding (t-SNE) of six properties of pharmaceutical relevance. The plot on the bottom-right shows all five compound libraries plotted on the same graph. The color code is the same as in Figure 2.

Enamine. As a reference (or control), the fingerprint diversity for the FDA data set was consistent with previous reports.[44,45] As anticipated, the focused data sets were more diverse than G9a because focused data sets are selected by chemical vendors providing promising insights to identify potential hits in an experimental screening campaign.

The similarity matrix of the G9a data set computed with the RDkit topological fingerprint and the Tanimoto coefficient (detailed in Section 2.4) is shown in Figure S7 in the Supporting Information. The matrix is visualized as a heatmap, and it displays similarity values using a continuous color scale, ranging from zero (white) to one (dark red), where one indicates the highest degree of similarity. The matrix shows that the data set has distinct clusters. Noteworthily, there are two major clusters, which correspond to quinazoline and quinoline derivatives. This observation aligns with what has been pointed out in the literature,[15] namely, the main chemical scaffolds experimentally tested to date are quinazoline and quinoline derivatives. Also, it is worth exploring scaffolds corresponding to the rest of the chemical space reported for G9a or elucidating other scaffolds of potential interest. Identifying these clusters provides valuable insights into the chemical space of G9a and highlights the presence of specific chemical motifs within the data set.

**3.5. Visualization of the Chemical Space.** Figure 6 shows a 2D visual representation of the property space obtained with t-SNE, i.e., the chemical space of the compound libraries as defined by the six properties of pharmaceutical interest. Visual representations of the chemical space using PCA are shown in Figure S8 in the Supporting Information. For PCA, the total variance captured by the first two components is summarized in Table S1 of the Supporting Information. The properties that contributed most to the first two principal components were TPSA and MW (PC1) and logP (PC2).

The visualization of the chemical space in Figures 6 and S8 indicates that the G9a data set is quite diverse, covering a broad region of the space. The visualizations also show that the libraries have comparable and similar properties to the approved compounds designed and prefiltered by the chemical companies selling the libraries. Compounds in the Life-Chemicals collection populate a more constrained region of the chemical space compared with other libraries, such as ChemDiv and Enamine, covering broader areas. As expected by the large variety of mechanisms and targets involved,


https://doi.org/10.1021/acsomega.3c04566
ACS Omega 2023, 8, 30694−30704

compounds in the FDA library populate the most extensive region of the chemical space.

## 4. CONCLUSIONS

Here, we explored the chemical space and structural diversity of reported G9a inhibitors, which was the main goal of this study. Analysis of the properties of pharmaceutical interest indicated that the data set of G9a inhibitors has, in general, drug-like properties. The commercial targeted libraries focused on G9a have a more restricted coverage in the chemical space considering the physicochemical properties (as compared to the G9a data sets and approved drugs) since chemical companies select the molecules to enforce that they have drug-like properties. The scaffold content analysis revealed that among the most frequent scaffolds found in focused libraries, no quinazoline or quinoline derivatives are abundant in the G9a data set, leaving the focused libraries to explore other potential sections and scaffolds in the chemical space. The diversity analysis revealed that the compounds tested with G9a have a large scaffold diversity. The LifeChemicals-focused library can be prioritized in a virtual and/or high-throughput screening campaign to identify G9a inhibitors due to its large scaffold diversity and different scaffolds compared to those tested with G9a. The fingerprint-based diversity analysis suggested that the diversity of the G9a set is not enough to explore the rest of the potential chemical space to elucidate innovative candidates. This aspect is reinforced by the fact that, so far, most of the tested compounds focus on quinazoline and quinoline scaffold derivatives. The previous point can be addressed by screening the focused libraries, where among the most frequent scaffolds the two previous scaffolds and their derivatives are not present, allowing users to navigate across other sectors with the potential of identifying novel candidates in the chemical space of G9a.

Among the focused libraries, LifeChemicals and Enamine are the most and least diverse data sets in terms of continuous properties, respectively. Based on structural fingerprints, Enamine and ChemDiv are the most and least diverse data sets, respectively. These findings suggest two key points; the first is that the compounds in the LifeChemical data set have a wide range of drug-like properties compared with other data sets and have a minor range of different molecular structures compared with Enamine. In contrast, Enamine is the data set with a limited range of drug-like properties but a large diversity of molecular structures. Finally, ChemDiv is a data set with a medium range of properties of pharmaceutical relevance and less diverse molecular structures. The visual representation of the chemical space indicated that the G9a data set is quite diverse and covers a wide region of space. The commercial targeted libraries focused on G9a have a more restricted coverage in the chemical space, as analyzed using physicochemical properties.

Taking the findings presented above together with the results published in the literature, it is concluded that G9a inhibitors have a large structural diversity and have, overall, drug-like properties. Also, the epigenetic-focused libraries that are commercially available are quite promising for identifying novel inhibitors of G9a. A major goal of this work is to conduct the systematic virtual screening of the focused libraries and test the enzymatic inhibition of selected hits experimentally. Studies are underway in our group and will be reported in due course.

## ASSOCIATED CONTENT

### Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsomega.3c04566.

Distribution of the enzymatic activity of the G9a inhibitors analyzed in this work; Radar plot of drug-likeness in the five data sets used in this study; Pearson's correlation coefficient heat map of the five data sets used in this study considering the six properties of pharmaceutical relevance: chemical space visualizations of the five data sets used in this study comparing the G9a with the reference databases; distribution of the 15 most populated scaffolds of the FDA data set; chemical structures of the 10 most frequent scaffolds per data set; similarity matrix of the 409 compounds in the G9a inhibitors data set; visual representation of the chemical space of the five data sets with principal component analysis of six properties of pharmaceutical relevance; and variance and contribution of each property per the first two principal components (PDF)

## AUTHOR INFORMATION


### Corresponding Author

**José L. Medina-Franco** − *DIFACQUIM Research Group Department of Pharmacy School of Chemistry, Universidad Nacional Autónoma de México, Mexico City 04510, Mexico;* orcid.org/0000-0003-4940-1107; Email: medinajl@unam.mx, jose.medina.franco@gmail.com

### Author

**Raziel Cedillo-González** − *DIFACQUIM Research Group Department of Pharmacy School of Chemistry, Universidad Nacional Autónoma de México, Mexico City 04510, Mexico;* orcid.org/0009-0009-9427-6959

Complete contact information is available at: https://pubs.acs.org/10.1021/acsomega.3c04566



### Funding

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


### Notes

The authors declare no competing financial interest.

## ACKNOWLEDGMENTS


Raziel Cedillo-González is grateful to Consejo Nacional de Ciencia y Tecnología (CONACyT), Mexico, for the scholarship no. 1099206. Helpful discussions with Fernanda I. Saldivar-González, Ana L. Chávez-Hernández, and Diana L. Prado-Romero are also acknowledged.


## ABBREVIATIONS

EHMT2, euchromatic histone *N*-methyltransferase; 2DD, data distribution; ERCS, epigenetic-relevant chemical space; H3,



https://doi.org/10.1021/acsomega.3c04566
*ACS Omega* 2023, 8, 30694−30704

histone 3; H4, histone 4; HBA, hydrogen bond acceptors; HBD, hydrogen bond donors; HDMs, histone demethylases; HMT, histone methyltransferases; KMT1C, lysine methyltransferase 1C; log *P*, octanol/water partition coefficient; MW, molecular weight; PCA, principal component analysis; PCC, Pearson's correlation coefficient; RB, number of rotatable bonds; Ro5, Lipinski's rule of five; SAM, *S*-adenosyl-L-methionine; SAR, structure−activity relationship; SE, Shannon entropy; SEAR, structure−epigenetic activity relationship; SSE, scaled SE; t-SNE, t-distributed stochastic neighbor embedding; DD, data distribution; ERCS, epigenetic-relevant chemical space; H3, histone 3; H4, histone 4; HBA, hydrogen bond acceptors; HBD, hydrogen bond donors; HDMs, histone demethylases; HMT, histone methyltransferases; KMT1C, lysine methyltransferase 1C; log *P*, octanol/water partition coefficient; MW, molecular weight; PCA, principal component analysis; PCC, Pearson's correlation coefficient; RB, number of rotatable bonds; Ro5, Lipinski's rule of five; SAM, *S*-adenosyl-L-methionine; SAR, structure−activity relationships; SE, Shannon entropy; SEAR, structure−epigenetic activity relationships; SSE, scaled SE; t-SNE, t-distributed stochastic neighbor embedding; TPSA, topological polar surface area

## REFERENCES


(1) Baxter, E.; Windloch, K.; Gannon, F.; Lee, J. S. Epigenetic Regulation in Cancer Progression. *Cell Biosci.* **2014**, *4*, 45.

(2) Berger, S. L.; Kouzarides, T.; Shiekhattar, R.; Shilatifard, A. An Operational Definition of Epigenetics Figure 1. *Genes Dev.* **2009**, *23* (7), 781−783.

(3) Zhao, Z.; Shilatifard, A. Epigenetic Modifications of Histones in Cancer. *Genome Biol.* **2019**, *20* (1), 245.

(4) Hamilton, P. J.; Nestler, E. J. Epigenetics and Addiction. *Curr. Opin. Neurobiol.* **2019**, *59*, 128−136.

(5) Richetto, J.; Meyer, U. Epigenetic Modifications in Schizophrenia and Related Disorders: Molecular Scars of Environmental Exposures and Source of Phenotypic Variability. *Biol. Psychiatry* **2021**, *89* (3), 215−226.

(6) Berson, A.; Nativio, R.; Berger, S. L.; Bonini, N. M. Epigenetic Regulation in Neurodegenerative Diseases. *Trends Neurosci.* **2018**, *41* (9), 587−598.

(7) Copeland, R. A.; Solomon, M. E.; Richon, V. M. Protein Methyltransferases as a Target Class for Drug Discovery. *Nat. Rev. Drug Discovery* **2009**, *8* (9), 724−732.

(8) Shilatifard, A. Chromatin Modifications by Methylation and Ubiquitination: Implications in the Regulation of Gene Expression. *Annu. Rev. Biochem.* **2006**, *75*, 243−269.

(9) Milner, C. M.; Campbell, R. D. The G9a Gene in the Human Major Histocompatibility Complex Encodes a Novel Protein Containing Ankyrin-like Repeats. *J. Biochem.* **1993**, *290* (3), 811−818.

(10) Tachibana, M.; Sugimoto, K.; Fukushima, T.; Shinkai, Y. Set Domain-Containing Protein, G9a, Is a Novel Lysine-Preferring Mammalian Histone Methyltransferase with Hyperactivity and Specific Selectivity to Lysines 9 and 27 of Histone H3. *J. Biol. Chem.* **2001**, *276* (27), 25309−25317.

(11) Casciello, F.; Windloch, K.; Gannon, F.; Lee, J. S. Functional Role of G9a Histone Methyltransferase in Cancer. *Front. Immunol.* **2015**, *6*, 487.

(12) Jana, A.; Naga, R.; Saha, S.; Banerjee, D. R. 3D QSAR Pharmacophore Based Lead Identification of G9a Lysine Methyltransferase towards Epigenetic Therapeutics. *J. Biomol. Struct. Dyn.* **2022**, 1−19.

(13) Bellver-Sanchis, A.; Singh Choudhary, B.; Companys-Alemany, J.; Sukanya; Ávila López, P. A.; Martínez Rodríguez, A. L.; Brea Floriani, J. M.; Malik, R.; Pallàs, M.; Pérez, B.; Griñán-Ferré, C. Structure-Based Virtual Screening and in Vitro and in Vivo Analyses Revealed Potent Methyltransferase G9a Inhibitors as Prospective Anti-Alzheimer's Agents. *ChemMedChem* **2022**, *17* (13), No. e202200002.

(14) Jin, Y.; Park, S.; Park, S.-Y.; Lee, C.-Y.; Eum, D.-Y.; Shim, J.-W.; Choi, S.-H.; Choi, Y.-J.; Park, S.-J.; Heo, K. G9a Knockdown Suppresses Cancer Aggressiveness by Facilitating Smad Protein Phosphorylation through Increasing BMP5 Expression in Luminal A Type Breast Cancer. *Int. J. Mol. Sci.* **2022**, *23* (2), 589.

(15) López-López, E.; Rabal, O.; Oyarzabal, J.; Medina-Franco, J. L. Towards the Understanding of the Activity of G9a Inhibitors: An Activity Landscape and Molecular Modeling Approach. *J. Comput. Aided Mol. Des.* **2020**, *34* (6), 659−669.

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

(17) Cao, H.; Li, L.; Yang, D.; Zeng, L.; Yewei, X.; Yu, B.; Liao, G.; Chen, J. Recent Progress in Histone Methyltransferase (G9a) Inhibitors as Anticancer Agents. *Eur. J. Med. Chem.* **2019**, *179*, 537−546.

(18) San José-Enériz, E.; Agirre, X.; Rabal, O.; Vilas-Zornoza, A.; Sanchez-Arias, J. A.; Miranda, E.; Ugarte, A.; Roa, S.; Paiva, B.; Estella-Hermoso de Mendoza, A.; Alvarez, R. M.; Casares, N.; Segura, V.; Martín-Subero, J. I.; Ogi, F.-X.; Soule, P.; Santiveri, C. M.; Campos-Olivas, R.; Castellano, G.; de Barrena, M. G. F.; Rodriguez-Madoz, J. R.; García-Barchino, M. J.; Lasarte, J. J.; Avila, M. A.; Martinez-Climent, J. A.; Oyarzabal, J.; Prosper, F. Discovery of First-in-Class Reversible Dual Small Molecule Inhibitors against G9a and DNMTs in Hematological Malignancies. *Nat. Commun.* **2017**, *8*, 15424.

(19) Rabal, O.; Sánchez-Arias, J. A.; San José-Enériz, E.; Agirre, X.; de Miguel, I.; Garate, L.; Miranda, E.; Sáez, E.; Roa, S.; Martínez-Climent, J. A.; Liu, Y.; Wu, W.; Xu, M.; Prosper, F.; Oyarzabal, J. Detailed Exploration around 4-Aminoquinolines Chemical Space to Navigate the Lysine Methyltransferase G9a and DNA Methyltransferase Biological Spaces. *J. Med. Chem.* **2018**, *61* (15), 6546−6573.

(20) López-López, E.; Prieto-Martínez, F.; Medina-Franco, J. L. Activity Landscape and Molecular Modeling to Explore the SAR of Dual Epigenetic Inhibitors: A Focus on G9a and DNMT1. *Molecules* **2018**, *23* (12), 3282.

(21) Flores-Padilla, E. A.; Juárez Mercado, K. E.; Naveja, J. J.; Kim, T. D.; Alain Miranda-Quintana, R.; Medina-Franco, J. L. Chemoinformatic Characterization of Synthetic Screening Libraries Focused on Epigenetic Targets. *Mol. Inform.* **2022**, *41* (6), No. e2100285.

(22) Prado-Romero, D. L.; Medina-Franco, J. L. Advances in the Exploration of the Epigenetic Relevant Chemical Space. *ACS Omega* **2021**, *6* (35), 22478−22486.

(23) Epigenetic screening libraries. https://lifechemicals.com/screening-libraries/targeted-and-focused-screening-libraries/epigenetic-screening-libraries 2023. (accessed May 05, 2023).

(24) Epigenetics library. https://enamine.net/compound-libraries/targeted-libraries/epigenetics-libraries 2023. (accessed May 05, 2023).

(25) Epigenetics focused set. https://www.chemdiv.com/catalog/focused-and-targeted-libraries/epigenetics-focused-set/ 2023. (accessed May 05, 2023).

(26) Wishart, D. S.; Knox, C.; Guo, A. C.; Cheng, D.; Shrivastava, S.; Tzur, D.; Gautam, B.; Hassanali, M. DrugBank: A Knowledgebase for Drugs, Drug Actions and Drug Targets. *Nucleic Acids Res.* **2008**, *36*, D901−D906.

(27) Landrum, G. RDKit. https://www.rdkit.org 2023. (accessed May 05, 2023).

(28) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Müller, A.; Nothman, J.; Louppe, G.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, É. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825−2830.




https://doi.org/10.1021/acsomega.3c04566
*ACS Omega* 2023, 8, 30694−30704

(29) Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996**, *39* (15), 2887−2893.
(30) 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**, *28* (11−12), 1551−1560.
(31) Weaver, W.; Shannon, C. E. *The Mathematical Theory of Communication*; University of Illinois Press: Urbana, IL, 1949.
(32) Guido, V. R.; Drake, F. L., Jr. Python 3 Reference Manual. Scotts Valley CreateSpace. **2009**.
(33) Gortari, E. F.; Medina-Franco, J. L. Epigenetic Relevant Chemical Space: A Chemoinformatic Characterization of Inhibitors of DNA Methyltransferases. *RSC Adv.* **2015**, *5*, 87465−87476.
(34) Sessions, Z.; Sánchez-Cruz, N.; Prieto-Martínez, F. D.; Alves, V. M.; Santos, H. P.; Muratov, E.; Tropsha, A.; Medina-Franco, J. L. Recent Progress on Cheminformatics Approaches to Epigenetic Drug Discovery. *Drug Discovery Today* **2020**, *25* (12), 2268−2276.
(35) López-López, E.; Fernández-de Gortari, E.; Medina-Franco, J. L. Yes SIR! On the Structure-Inactivity Relationships in Drug Discovery. *Drug Discovery Today* **2022**, *27* (8), 2353−2362.
(36) Mahgoub, R. E.; Atatreh, N.; Ghattas, M. A. Chapter Three - Using Filters in Virtual Screening: A Comprehensive Guide to Minimize Errors and Maximize Efficiency. In *Annu. Rep. Med. Chem.*; Caballero, J., Ed.; Academic Press, 2022; Vol. *59*, pp 99−136.
(37) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings 1PII of original article: S0169-409X(96)00423-1. The article was originally published in Advanced Drug Delivery Reviews 23 (1997) 3−25. 1. *Adv. Drug Delivery Rev.* **2001**, *46* (1−3), 3−26.
(38) Petit, J.; Meurice, N.; Kaiser, C.; Maggiora, G. Softening the Rule of Five–Where to Draw the Line? *Bioorg. Med. Chem.* **2012**, *20* (18), 5343−5351.
(39) Veber, D. F.; Johnson, S. R.; Cheng, H.-Y.; Smith, B. R.; Ward, K. W.; Kopple, K. D. Molecular Properties That Influence the Oral Bioavailability of Drug Candidates. *J. Med. Chem.* **2002**, *45* (12), 2615−2623.
(40) Neidle, S. *Therapeutic Applications of Quadruplex Nucleic Acids*; Academic Press: Boston, 2012, pp 151−174..9 - Design Principles for Quadruplex-Binding Small Molecules
(41) Swinney, D. C.; Anthony, J. How Were New Medicines Discovered? *Nat. Rev. Drug Discovery* **2011**, *10* (7), 507−519.
(42) Brown, N.; Jacoby, E. On Scaffolds and Hopping in Medicinal Chemistry. *Mini Rev. Med. Chem.* **2006**, *6* (11), 1217−1229.
(43) Hu, Y.; Bajorath, J. Structural and Activity Profile Relationships Between Drug Scaffolds. *AAPS J.* **2015**, *17* (3), 609−619.
(44) López-Vallejo, F.; Giulianotti, M. A.; Houghten, R. A.; Medina-Franco, J. L. Expanding the Medicinally Relevant Chemical Space with Compound Libraries. *Drug Discovery Today* **2012**, *17* (13−14), 718−726.
(45) González-Medina, M.; Prieto-Martínez, F. D.; Naveja, J. J.; Méndez-Lucio, O.; El-Elimat, T.; Pearce, C. J.; Oberlies, N. H.; Figueroa, M.; Medina-Franco, J. L. Chemoinformatic Expedition of the Chemical Space of Fungal Products. *Future Med. Chem.* **2016**, *8* (12), 1399−1412.


https://doi.org/10.1021/acsomega.3c04566
*ACS Omega* 2023, 8, 30694−30704