De novo design of inhibitors of DNA methyltransferase 1: A critical comparison of ligand- and structure-based approaches

Diana L. Prado-Romero,[1] Fernanda I. Saldivar-González,[1] Iván López-Mata,[2,3] Pedro A. Laurel-Garcia,[1] Norberto Sánchez-Cruz,[3,4] José L. Medina-Franco[1,*]

[1]*DIFACQUIM Research Group, Department of Pharmacy, School of Chemistry, Universidad Nacional Autónoma de México, Avenida Universidad 3000, Mexico City 04510, Mexico*

[2] *División Académica de Ciencias Básicas, Universidad Juárez Autónoma de Tabasco, Carretera Cunduacán-Jalpa de Méndez, Km 1, Cunduacán, Tabasco 86690, Mexico*

[3] *Instituto de Química, Unidad Mérida, Universidad Nacional Autónoma de México, Carretera Mérida-Tetiz Km. 4.5, Ucú, Yucatán 97357, Mexico*

[4] *Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas Unidad Mérida, Universidad Nacional Autónoma de México, Sierra Papacál, Yucatán 97302, Mexico*

[*]Correspondence author: medinajl@unam.mx; Tel.: +52-55-5622-3899

Abstract

Designing and developing inhibitors against the epigenetic target DNA methyltransferase (DNMT) is an attractive strategy in epigenetic drug discovery. DNMT1 is one of the epigenetic enzymes with significant clinical relevance. Structure-based *de novo* design is a drug discovery strategy used in combination with similarity searching to identify a novel DNMT inhibitor with a novel chemical scaffold and warrants further exploration. This study aimed to continue exploring the potential of *de novo* design to build epigenetic-focused libraries targeted toward DNMT1. Herein, we report the results of an in-depth and critical comparison of ligand- and structure-based *de novo* design of screening libraries focused on DNMT1. The newly designed chemical libraries focused on DNMT1 are freely available on GitHub at https://github.com/DIFACQUIM/De-Novo_DNMT1.

Keywords: chemoinformatics; drug discovery; docking; epigenetics; Epigenetic Target Profiler; focused libraries; fragments; library design; open-access.

Abbreviations: 3D, three-dimensional; AUC, area under the curve; CDP, Consensus diversity plots; CSP3; ESOL, estimated solubility; HBA, hydrogen bond acceptors; HBD, hydrogen bond donors; DNMT, DNA methyltransferase; ECFP4, Extended Connectivity Fingerprint of radius 2; IC_{50}, half-maximal inhibitory concentration; LE, ligand efficiency; logD, distribution coefficient; logP, partition coefficient; MOE, Molecular Operating Environment; MW, molecular weight; PC, principal component; nRotB, number of rotatable bonds; PCA, principal component analysis; PDB, Protein Data Bank; QED, quantitative estimate of drug-likeness; RECAP, retrosynthetic combinatorial analysis procedure; RMSD, root mean square deviation; RO3, Rule of Three; SAH, *S*-Adenosyl-*L*-homocysteine; SAscore, synthetic accessibility score; TPSA, topological polar surface area; tSNE, t-distributed stochastic neighbor embedding; Vina, AutoDock Vina.

1. Introduction

The primary goal of *de novo* design is to generate new chemical entities with desired properties [1–4]. Generating bioactive compounds within the physicochemical-relevant chemical space is highly desirable in drug discovery. Thus, *de novo* design is an attractive approach to generate focused libraries with a desired or predicted bioactivity towards a biochemical or molecular target. Indeed, other than the common and traditional general screening compound libraries, chemical vendors and commercial companies are developing focused libraries of chemical compounds, as well as fragments (privileged fragments) focused on various targets of therapeutic relevance [5,6]. Although the physical samples (compound material) of the compounds are readily available for experimental screening in, for example, high-throughput or medium-throughput mode, the chemical structures are publicly accessible and can be used for benchmarking computational studies.

DNA methyltransferases (DNMTs) are one of the main epigenetic target families with clinical relevance [7–9]. DNMTs include two *de novo* methyltransferases, DNMT3A and DNMT3B, and the maintenance methyltransferase DNMT1 (the most abundant DNMT), which is in charge of duplicating the pattern of DNA methylation during replication, and it is necessary for adequate mammalian development. Two nucleoside inhibitors of DNMT methylation, azacitidine and decitabine, have been approved by the FDA to treat myelodysplastic syndrome [10]. Despite their high efficacy, nucleoside drugs suffer from undesirable pharmacokinetic profiles, chemical instability, and toxic side effects [11]. Due to DNA methylation being a fundamental mechanism for gene regulation, the design and development of non-nucleoside DNMT modulators is a promising strategy for developing novel epigenetic-based therapies. Among the earliest non-nucleoside inhibitors identified for DNMT1 is (-)-epigallocatechin-3-gallate (EGCG) (ref) (Figure 1). This compound and other DNMT1 inhibitors sourced from natural origins have been subject to extensive review [12,13].

In epigenetic drug discovery, including DNMT1, efforts are increasing to design and analyze focused libraries of DNMT inhibitors. Our research group recently reported a comprehensive analysis of the chemical content, diversity, and chemical space coverage of eleven commercial epigenetic-focused libraries with over 50,000 molecules. In that study, the most and least diverse chemical libraries were identified [14]. Moreover, separate research endeavors have identified five compounds, including glyburide and panobinostat (Figure

1), as DNMT1 inhibitors [15]. Also highlighted is the quinazoline UNC-0646 (Figure 1), initially identified as a G9a inhibitor, which has been recently discovered to also effectively inhibit DNMT1 at the nanomolar level [16].

Although in epigenetic drug discovery *de novo* design has been used to identify inhibitors of a bromodomain [17] and other epigenetic targets, such technique has been reported scarcely for designing DNMT inhibitors [18]. Recently, ligand-based *de novo* design based on compounds with reported activity for DNMT1 and a diverse subset of screening compounds were used to generate novel candidate DNMT1 inhibitors with drug-like properties. Then, the compounds designed *de novo* were used as reference for fingerprint-based similarity searching in a commercial epigenetic-focused library. The most similar compounds were acquired and tested in an enzymatic inhibition assay, identifying a DNMT1 inhibitor (F447-0397) with a novel chemical scaffold and an enzymatic inhibitory concentration in the micromolar range (**Figure 1**) [19]. The identification of F447-0397 as an inhibitor of DNMT1, inspired by *de novo* design, is the first approach to the potential of such drug designing approaches to identify novel active molecules with DNMT1 inhibitory activity. Also, recently Lanka et al. implemented a multi-step virtual screening involving docking a fragment library to yield two potential active compounds with DNMT1 [20].

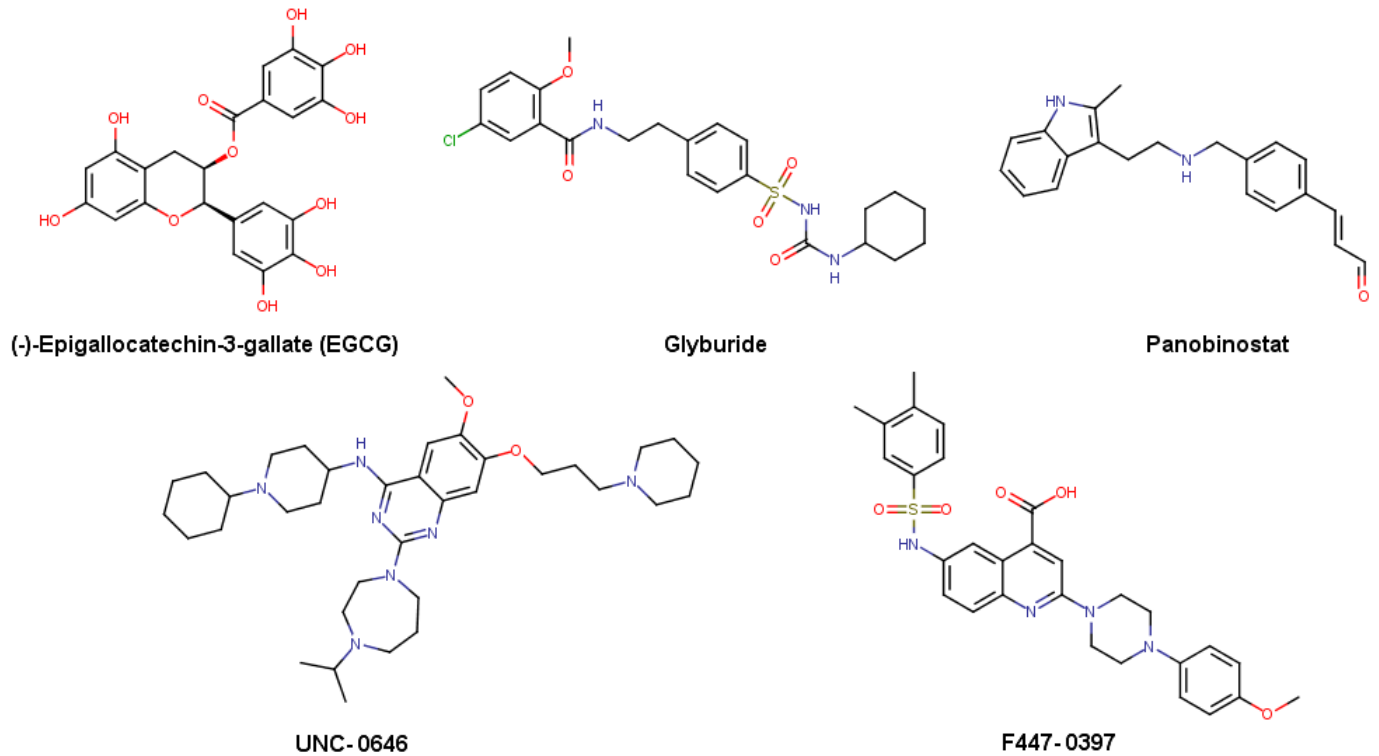

Figure 1. Chemical structures of representative inhibitors of DNA methyltransferases (DNMTs). Notably, F447-0397 was reported recently and has a novel chemical scaffold among DNMT1 inhibitors.

To further explore the potential of *de novo* design to build epigenetic-focused libraries targeted toward DNMT1, herein we discuss a critical assessment of ligand- and structure-based *de novo* design approaches. The new chemical libraries focused on DNMT1 represent an addition to the epigenetic libraries currently available. The newly designed compound collections are freely available and can be used for further virtual and experimental screening.

2. Methods

Figure 2 outlines the main *de novo* design approaches compared in this study. Briefly, the design was carried out using ligand-based (alvaBuilder [21]) and structure-based (LigBuilder [22]) strategies. The fragments' sources were selected to represent similar starting points for both software despite differences in the fragmentation procedure. alvaBuilder requires chemical libraries as training sets for its fragmentation. Then, generated fragments are used as building blocks by the software. Each molecule selected from the training set is split into fragments following the Bemis and Murcko rules [21]. In contrast, LigBuilder employs external fragment libraries provided by the user. In this study, fragment libraries were obtained using the fragmentation rules provided by the RECAP (Retrosynthetic Combinatorial Analysis Procedure) [23] algorithm or fragment libraries that are commercially available.

The fragments' sources included DNMT1 inhibitors, commercial libraries, natural products, and default fragments in LigBuilder. Details of the filtration of chemical libraries and fragments are specified in the following sections. Designed molecules were compared to each other, taking into account the different fragment sources. To unify the comparison criteria, descriptors were calculated with the same methodology. The calculations performed were: Quantitative Estimate of Drug-Likeness (QED) [24], Synthetic Accessibility Score (SAscore) [25], molecular docking with two docking programs, AutoDock Vina (Vina) [26,27] and LeDock [28]. We also profile the newly designed libraries regarding scaffold content and the predictions with a DNMT1 classification model based on a recently developed model to estimate the activity of DNMT1 inhibitors [29] (**Figure 2**). In agreement with Open Science, the newly designed compound libraries and the code for all the analyses reported in this study are freely available on GitHub at https://github.com/DIFACQUIM/De-Novo_DNMT1.

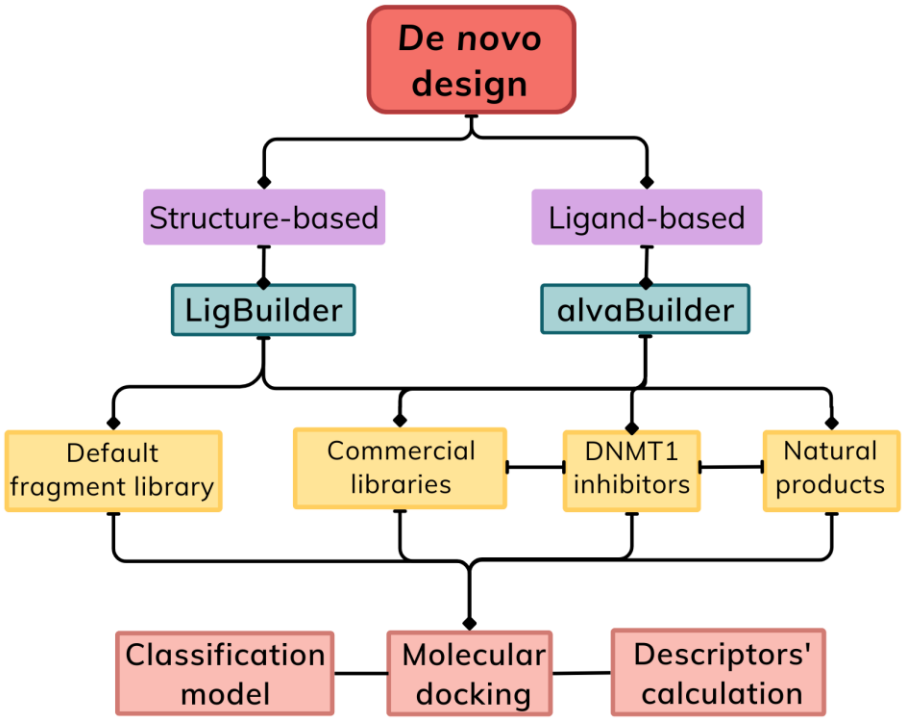

Figure 2. Overview of the *de novo* strategy comparison conducted in this study. Two main *de novo* design strategies, structure-based and ligand-based, were used, and they were performed with the software LigBuilder and alvaBuilder, respectively. Both *de novo* compound libraries were profiled with machine learning classification models, molecular docking, and descriptors to characterize and compare the drug-likeness, synthetic accessibility, and global structural and property diversity, including the molecular scaffolds.

2.1. Data curation

Chemical and fragment libraries and compounds designed *de novo* were curated using the same protocol, previously published [30,31] using RDKit [32] and MolVS [33] Python libraries. Briefly, compounds were standardized, and the largest fragment was kept in cases where molecules had more than one component. Only molecules with the elements: H, B, C, N, O, F, Si, P, S, Cl, Se, Br, and I were selected. The remaining compounds were neutralized and reionized to generate a canonical tautomer and remove duplicates. If present, the stereochemistry information was preserved only for the compounds designed *de novo*.

2.2. Ligand-based *de novo* design

AlvaBuilder v.1.0.10 [21] was used for the ligand-based design. This software uses fragments from the user-selected training set to implement a graph-based construction of molecules. AlvaBuilder fragments the molecules from the training set into ring systems, linkers, and side chains. Eight different chemical libraries were used as training sets, as detailed in section 2.2.1.

The scoring function was established with a range of seven descriptors (values summarized in Table S1 in the Supplementary Material) and a penalized substructure matching for nucleoside scaffolds (Figure S1). The score file with the set of rules is on GitHub at https://github.com/DIFACQUIM/De-Novo_DNMT1/tree/main/alvaBuilder. The ranges were established considering the mean and standard deviation values from molecules with reported biological activity against DNMT1, with $IC_{50} \leq 10$ □M. This threshold has been previously applied in other studies to label active/inactive compounds [34,35].

Seven descriptors were computed with alvaDesc 2.0.10 [36]. They included molecular weight (MW), hydrogen bond donors (HBD) and acceptors (HBA), consensus partition coefficient octanol/water (log P), aqueous solubility (ESOL), synthetic accessibility (SAscore), and topological polar surface area (TPSA). The aggregation method was an arithmetic mean, and the same scoring function was used for all the training sets. The genetic algorithm was set with a population size of 70 and 100 iterations, generating 700 molecules for each training set.

2.2.1. Training chemical libraries

Eight compound libraries summarized in Table 1 were used as training sets for alvaBuilder, including five commercial libraries (two of which are epigenetic-focused), DNMT1 inhibitors, food chemicals from FooDB, and natural products. As the number of compounds available for each chemical library is highly variable, a filtration was made to match the number of 285 DNMT1 active compounds in ChEMBL. The first step after data curation was the selection of the molecules with MW greater than 300 to avoid fragment-like compounds [37]. The MW descriptor was computed and sorted in descending order with alvaMolecule 1.0.4 [38] after representing all the libraries with the Aromatic form available in the software. Then, a diverse subset of 285 molecules was selected with the MaxMin algorithm [39] implemented in Molecular Operating Environment (MOE) 2022.02 [41], using the MACCS Keys fingerprint (166-bits) and the Tanimoto coefficient [40,41]. Details of the remaining number of compounds for each stage are in Table S2 in the Supplementary Material.

Table 1. Training chemical libraries used for alvaBuilder as fragment sources.

Database	Description	Number of compounds
ChemDiv DNMT-targeted library [42]	Small molecules targeting DNMTs.	33936
ChemDiv Epigenetics Focused Set [43]	Drug-like compounds targeting families of epigenetic proteins.	25883
ChemDiv Soluble Diversity Library [44]	Soluble drug-like compounds focused against various biological targets.	15500
ChEMBL actives [45]	Inhibitors with $IC_{50} \leq 10$ μM from ChEMBL 31.	285
FooDB [46]	Food chemicals.	68658
Life Chemicals Diversity Set [47]	A subset of 5,000 drug-like compounds with a wide range of chemical structures.	5120
Life Chemicals Epigenetic Focused Library [48]	Drug-like compounds selected by 2D fingerprint similarity search with known epigenetic modulators.	3578
UNPD-A [49,50]	A diverse subset from the Universal Natural Product Database.	14994

2.3. Structure-based *de novo* design

LigBuilder V3 was used for the structure-based design [22]. LigBuilder constructs molecules by linking and growing fragments from the libraries available. Adding fragments to the default library or changing it completely is possible. The search strategy used in LigBuilder is a genetic algorithm to develop and evolve the molecules. This algorithm mimics the evolution of a population under selection pressure [22,51,52]. For the estimation of the binding affinity of the new molecules, the default empirical scoring function in LigBuilder was used in all cases.

2.3.1. Binding site detection

To begin the calculations with LigBuilder, defining the binding site of the protein of interest is necessary. For this purpose, the software includes the CAVITY module, which analyzes the binding pocket of the protein's three-dimensional (3D) structure of interest [53]. As a result of CAVITY, the required data of the binding pocket is ready for the construction step with the BUILD module (section 2.3.2). The first step was to retrieve DNMT1

3D coordinates from the RCSB Protein Data Bank (PDB). The crystallographic structure of DNMT1 was PDB ID: 4WXX, available online: https://www.rcsb.org/ (accessed on June 30th, 2023) [54]. Chain A from the crystallographic structure of DNMT1, without water molecules, was kept for the calculations, and all water molecules were eliminated. The calculations with CAVITY were in standard ligand mode, with the structure of *S*-adenosyl-*L*-homocysteine (SAH) used as a guide for the search of the binding cavities. The cavity with the best druggability prediction was chosen for the *de novo* design with all the fragment libraries.

2.3.2. Molecule design

Once the binding pocket was defined with CAVITY, the module BUILD was used in Exploring mode to design molecules with potential activity against DNMT1. Default parameters for the construction of molecules were modified for the descriptors: MW (175 – 665), logP (-3 – 7), hydrogen bond donors (0 – 8), and acceptors (1 – 12), to cover the previously observed values from DNMT1 inhibitors found in ChEMBL (minimum-maximum, Table S1). The genetic algorithm was set to a population of 500 and 10 generations. The configuration files modified for the construction of the molecules with LigBuilder are available on GitHub at https://github.com/DIFACQUIM/De-Novo_DNMT1/tree/main/LigBuilder.

2.3.3. Fragment libraries

Eight different fragment libraries were chosen as building blocks for LigBuilder, including the available default library, four commercial libraries (one epigenetic-focused), DNMT1 inhibitors, food chemicals from FooDB, and natural products. These were selected considering the sources of the training sets used before with alvaBuilder (Table 1). Notably, fragments from DNMT1 inhibitors, FooDB, and natural products were obtained from the fragmentation of the corresponding curated chemical libraries with the RECAP algorithm. Table 2 summarizes the sources and number of compounds of the eight libraries.

Since the number of fragments differs from the default library, the same number was selected from the remaining libraries. A workflow in KNIME 4.7.7 [55] was implemented to compute the different stages of the filtering. Fragments H_2O, NH_3, and HCl were not considered. The first stage was to choose all fragments with MW less than 300 Da, according to the Rule of Three (RO3) [37]. This was done with the RDKit Descriptor Calculation node.

Once the first filtering was made, a diverse subset of 400 fragments was selected with the MaxMin algorithm implemented in the RDKit Diversity Picker, using the MACCS Keys fingerprint (166-bits) and the Tanimoto coefficient [40,41]. Then, the distribution coefficient (logD), at pH = 7.4, of the 400 fragments was computed with ADMETlab 2.0 [56]. After that, the data was loaded into the KNIME workflow to keep the fragments with logD less or equal to three. Table S3 summarizes the details of the remaining number of fragments for each step. With the remaining compounds, a final diverse subset of 177 fragments was picked with the MaxMin algorithm, computed with MACCS Keys (166-bits) and the Tanimoto coefficient (Figure 3).

The default library from LigBuilder and the four commercial libraries (Table 2) were used with all the hydrogen atoms as growing points for the molecular assembly. The growing points for the three fragment libraries constructed with RECAP were selected, taking into account the site where the fragmentation was made. The RECAP fragment libraries were analyzed using DataWarrior 05.05.00 [57] to identify the fragmentation site. Using the molecule viewers BIOVIA Discovery Studio Visualizer 24.1.0.23298 [58] and PyMol 2.5 [59], the hydrogen atom or atoms where the fragmentation occurred were identified and marked as the growing site for each fragment.

Table 2. Fragment libraries used for the *de novo* design with LigBuilder.

Database	Description	Number of fragments
LigBuilder default [22]	Common chemical groups and ring frameworks observed in organic compounds.	177
ChemDiv Fragments Library [60]	Fragments with desirable properties, including diversity.	11269
ChemDiv Epigenetics Fragments [61]	Privileged fragments focused on epigenetic regulators.	9196
ChEMBL actives [23,45]	Fragments from ChEMBL actives (Table 1) obtained with RECAP.	1645
FooDB [23,46]	Fragments from FooDB compounds up to 1500 Da obtained with RECAP.	225206
Life Chemicals Soluble Fragments [62]	Diversity-oriented fragment library with experimental solubility data.	1280
Selleckchem [63]	A collection of fragments for Fragment-Based Drug Discovery.	1015
UNPD-A [23,49,50]	Fragments from UNPD-A up to 1500 Da (Table 1) obtained with RECAP.	412110

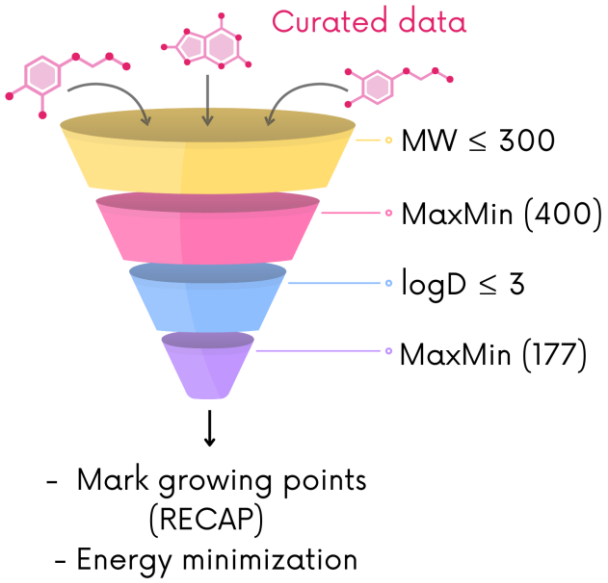

Figure 3. Filtration steps to select the fragments used as building blocks for LigBuilder. All the libraries matched the 177 fragments from the default library; their 3D coordinates were built, and the geometry was energy minimized for each one. The growing sites were specified only for the fragment libraries constructed with the RECAP algorithm.

The 177 fragments of the final libraries were built, and their geometry was energy minimized using the MFF94x forcefield implemented on MOE 2022.02 [64] (Figure 3). Each minimized fragment was stored in an individual .mol2 file. The INDEX file of LigBuilder, to recognize the corresponding fragments for molecular construction, was modified for each of the libraries.

2.4. Visualization of the chemical space

To generate a visual representation of the chemical space of the *de novo* designed libraries and reference data sets, we used t-distributed stochastic neighbor embedding (t-SNE) [65] and principal component analysis (PCA) based on structural fingerprints and continuous properties of pharmaceutical interest. The structural fingerprints used in this study were Extended Connectivity Fingerprints radius two (ECFP4) [66] and MACCS Keys (166-bits), which were computed for all molecules using the RDKit library [32]. Six molecular properties relevant to pharmaceuticals were computed: MW, HBA, HBD, TPSA, logP, and number of rotatable bonds (nRotB).

2.5. Drug likeness and synthetic accessibility

We calculated the QED as an empirical and well-established measure of drug-likeness. QED is based on eight molecular properties relevant to determining drug-likeness: logP, HBA, HBD, PSA, nRotB, number of aromatic rings, and number of structural alerts [24]. The weighted geometric mean of these properties is the resultant QED that goes from zero (worst) to one (best). The suggested threshold for the attractive compounds is 0.67. This value or higher is desirable for drug-like compounds. QED was calculated with RDKit, which is available in the Chem module.

The synthetic accessibility was computed using the SAscore. Synthetic accessibility is a combination of fragment structures comprising historical knowledge of chemical synthesis and a complexity penalty that considers non-standard structural features like stereocomplexity [25]. The score is scaled to 1 (very easy to synthesize) and 10 (very difficult to synthesize). Those molecules with a SAscore greater than six are considered not synthetically feasible. SAscore was calculated using the implementation available in the RDKit library.

2.6. Molecular docking

Our research group previously published the molecular docking protocol used here, with RMSD re-docking values lower than 2 Å for the co-crystallized SAH [19]. This study used two docking programs with different algorithms, LeDock and Vina, to calculate docking scores for all the *de novo-designed* compounds. Relevant details of this protocol are explained in the following lines.

The same crystallographic structure of DNMT1 (PDB ID: 4WXX) used for binding site detection with CAVITY (section 2.3.1) was selected for molecular docking calculations. The protein preparation was made with default settings of the QuickPrep module MOE v. 2022.02 [64]: addition of all the lacking hydrogen atoms, protonation state at pH 7, elimination of water molecules 4.5 Å farther from the protein, addition of missing amino acids residues (breaks of up to ten residues and terminal out gaps of up to five residues) and for larger gaps, neutralization of the endpoints adjoining empty residues and energy minimization. The parameters employed for the energy minimization stage were from the AMBER14:EHT forcefield (ff14SB [67] for the protein, MAB forcefield [68], and AM1-BCC charges for SAH [69]).

Before docking, the corresponding ligands (285 from the reference database and 8066 *de novo* compounds) were built, and their geometry was energy minimized using the MFF94x forcefield implemented on MOE. For every ligand, the most stable tautomer at physiological pH (7.4) was chosen [64].

2.6.1. Docking with Vina

The file with the prepared ligands was split with the LeFrag module [28], and Open Babel v.3.1.1 [70] was used to convert to .pdb format. Protein and ligands were converted to .pdbqt with MGLTools v.1.5.6. The molecular docking was carried out with Vina v.1.2.3 [26,27] with an exhaustiveness of 8 and 5 binding modes to output. The best score for each ligand was selected for further analysis with the code freely available at https://github.com/DIFACQUIM/Docking. The grid box was centered in the coordinates: -47.673, 61.885, 6.256 (x, y, z) with a search space of 17 x 25 x 14 Å.

2.6.2. Docking with LeDock

Docking with Ledock [71] was carried out in the SAH binding site with the default settings: the grid centered 4 Å around the co-crystallized SAH. Twenty docking runs for every ligand and 1 Å for the root mean square deviation (RMSD) clustering. For further data analysis, the best score for every ligand was selected using the code available at https://github.com/DIFACQUIM/Docking.

2.7. Global diversity and scaffold analysis

The total or global diversity of the *de novo* designed libraries and the DNMT1 inhibitors dataset was analyzed, considering multiple structure representations. Specifically, the data sets were compared regarding structural fingerprints, molecular scaffolds, and properties of pharmaceutical interest. Considering the different representations, the total or global diversity of the data sets was analyzed using Consensus Diversity Plots (CDPlots) [72].

2.7.1. Molecular descriptors

The structural fingerprints used in this study to assess the global diversity were ECFP4 and MACCS Keys (166-bits), computed for all molecules using RDKit. Subsequently, a similarity matrix was generated based on these two fingerprints and the Tanimoto coefficient [73]. Values outside the diagonal of the similarity matrix were used to calculate the median MACCS Keys/Tanimoto of the pairwise comparisons.

The molecular scaffolds were generated using the Bemis-Murcko definition using RDKit [74]. We calculated the proportion of scaffolds relative to the number of compounds (N/M). Based on Cyclic System Retrieval (CSR) curves, we calculated the area under the curve (AUC) and the fraction of chemotypes that recover 50% of the molecules in the data set (F_{50}). The Shannon entropy of the ten most frequent scaffolds was also calculated [75]. To identify possible unique scaffolds, the unique scaffolds of each database were compared with each other.

Six molecular properties of pharmaceutical interest were computed to assess further diversity: MW, HBD, HBA, TPSA, nRotB, and logP. Subsequently, pairwise intra-database chemical diversity of the six molecular properties was determined using Euclidean distance.

2.8. Classification models

Different classification models were developed to predict the activity class of compounds tested against DNMT1. Their construction process was based on the methodology previously described [29], which is summarized below.

2.8.1. Dataset

Compounds with inhibitory activity against DNMT1 (reported as IC_{50}) were obtained from ChEMBL 33 [76] (last updated May 2023). To build classification models, compounds were labeled as active if they had unequivocally assigned activities lower than or equal to 10 μM and as inactive in the opposite case. Compounds whose labels could not be unequivocally assigned were removed from the dataset (e.g., activity < 100 μM, activity > 1 μM, or duplicated compounds with contradictory labels). This resulted in a dataset containing 225 compounds, with 141 of them labeled as active and the remaining 84 labeled as inactive. This dataset was randomly divided in a stratified manner into two subsets: 80% was used to train different binary classification models and select the best ones using leave-one-out cross-validation (LOO-CV), and the 20% remaining was used as an external test set.

2.8.2. Molecular representations

Three molecular fingerprints of different designs were chosen as descriptors for the construction of the classification models: a) Molecular ACCess System (MACCS) Keys (166-bit) [77], b) Morgan fingerprint with

radius 2 (2048-bit) [66], and c) RDK fingerprint (2048-bit). All of them were generated using the RDKit open-source cheminformatics toolkit, version (2023.09.05) for Python.

2.8.3. Machine learning methods

Binary classification models were built using five machine learning algorithms: k-Nearest Neighbors (k-NN) [78], Random Forest (RF) [79], Gradient Boosting Trees (GBT) [80], Support Vector Machines (SVM) [81], and Feed-Forward Neural Networks (FFNN) [82]. These methods were implemented using the Scikit-learn Python library (v1.4.1) [83]. Training instances were represented by feature vectors (fingerprints) and associated with class labels (active/inactive). Hyperparameters for each model were optimized using leave-one-out (LOO) cross-validation over a limited space. Only select hyperparameters for each algorithm were optimized to keep the search space small. The considered hyperparameters for each model are included in Table S4 in the Supplementary Material, with default values assumed if not explicitly indicated.

Each optimized model was defined as the combination of an algorithm and a fingerprint, whose hyperparameters were optimized using LOO-CV, with balanced accuracy (BA) employed to select the best hyperparameters. To estimate the confidence of the predictions as they differ from the training set, the best-performing models were also compared in terms of their precision and recall computed on a distance-to-model (DM) basis [84,85] using Morgan fingerprints. For that, predictions were categorized into four quartiles according to their mean Jaccard distance to the compounds in the training set. Having the confidence estimation, these models were retrained in the entire dataset and applied to predict the activity class of the compounds generated by *de novo* design.

3. Results and discussion

In the following sections, results from the construction and cheminformatic analysis of two *de novo* design-focused libraries on DNMT1 are described (strategy outlined in Figure 2). Docking scores and results from a classification model are also discussed as an insight into the potential activity of the designed libraries. Since information on the 3D coordinates of DNMT1 and compounds with reported activity against this target are available, we used ligand and structure-based strategies. The comparison of both strategies could disclose the potential differences between the molecular and structural characteristics of the libraries.

The ligand-based strategy was done with alvaBuilder, restricted with the scoring function to propose

molecules with properties within the observed value ranges for the selected descriptors. The ranges were established by considering the values of known DNMT1 inhibitors. The curated library contains 5575 compounds from eight different training sets used as fragment sources. The structure-based strategy was performed with LigBuilder, restricted to a binding site cavity surrounding SAH. The curated library has 2491 compounds from eight distinct fragment libraries.

3.1. Ligand-based *de novo* design with alvaBuilder

One of the desired characteristics of the focused libraries on DNMT1, here constructed with *de novo* design, is to preserve the chemical similarity in terms of physicochemical properties while exploring different regions of the chemical space based on molecular structure. From the chemical space visualizations based on six molecular properties of pharmaceutical relevance (Figures 4 and 5A), it can be seen that compounds designed with alvaBuilder are within the property space of the molecules with reported activity against DNMT1 on ChEMBL. This conclusion can be derived from both visualizations generated with PCA and t-SNE. This observation could be related to the scoring function used in alvaBuilder (section 2.2), based on almost the same molecular properties selected for the chemical space visualizations. Five of six properties were also included in the scoring function, only excluding nRotB. The scoring included ESOL and SAscore results and a penalty for substructures commonly found in analog nucleosides. The latter consideration was utterly relevant since the goal is expanding the chemical space of DNMT1 non-nucleoside inhibitors. The properties of the scoring function for alvaBuilder are secondary constraints for the design and could be causing the visualization of all the design libraries to be more compact in the PCA graphs (Figure 4).

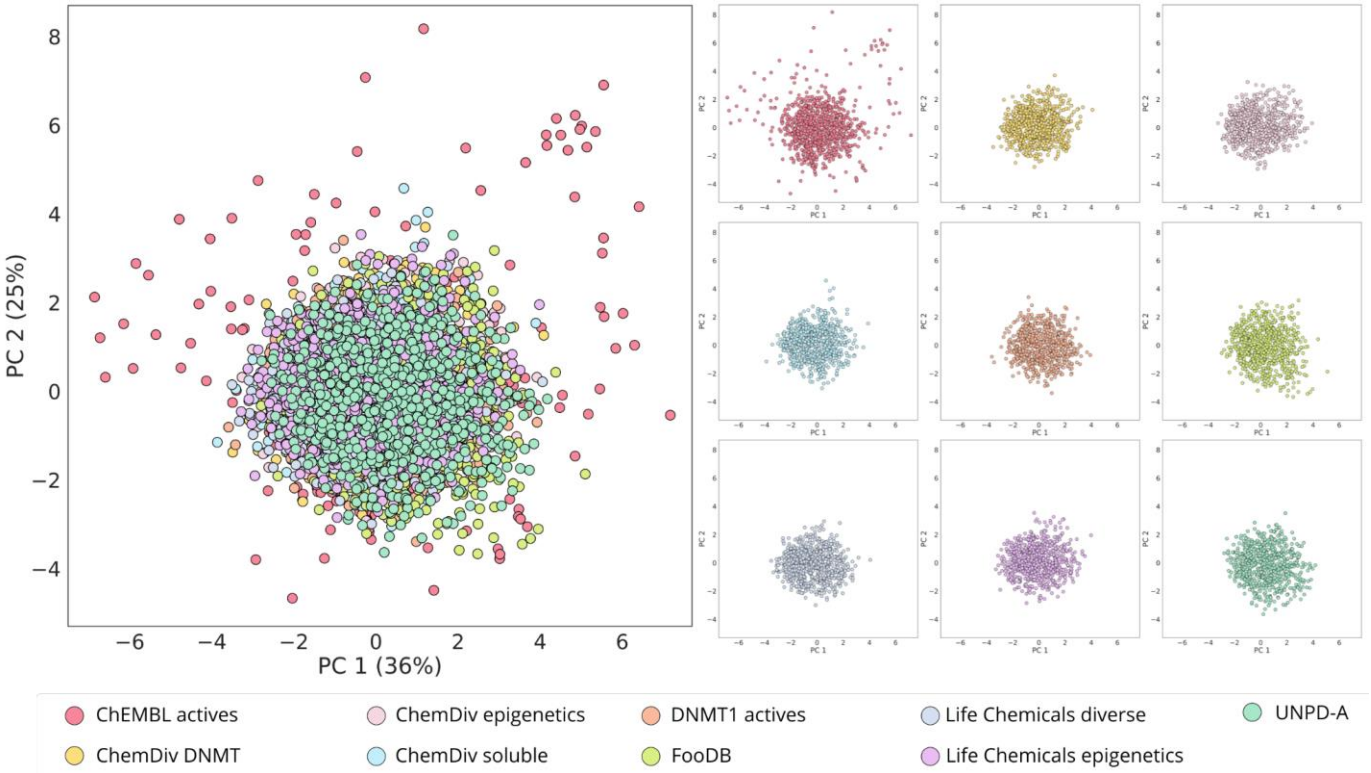

Figure 4. Chemical space visualization of the *de novo* designed libraries with alvaBuilder. The visual representation was made with PCA as dimensionality reduction and six physicochemical properties as molecular representation. On the left are the nine superimposed databases, and on the right are the individual data sets using the same coordinates as the corresponding representations on the left.

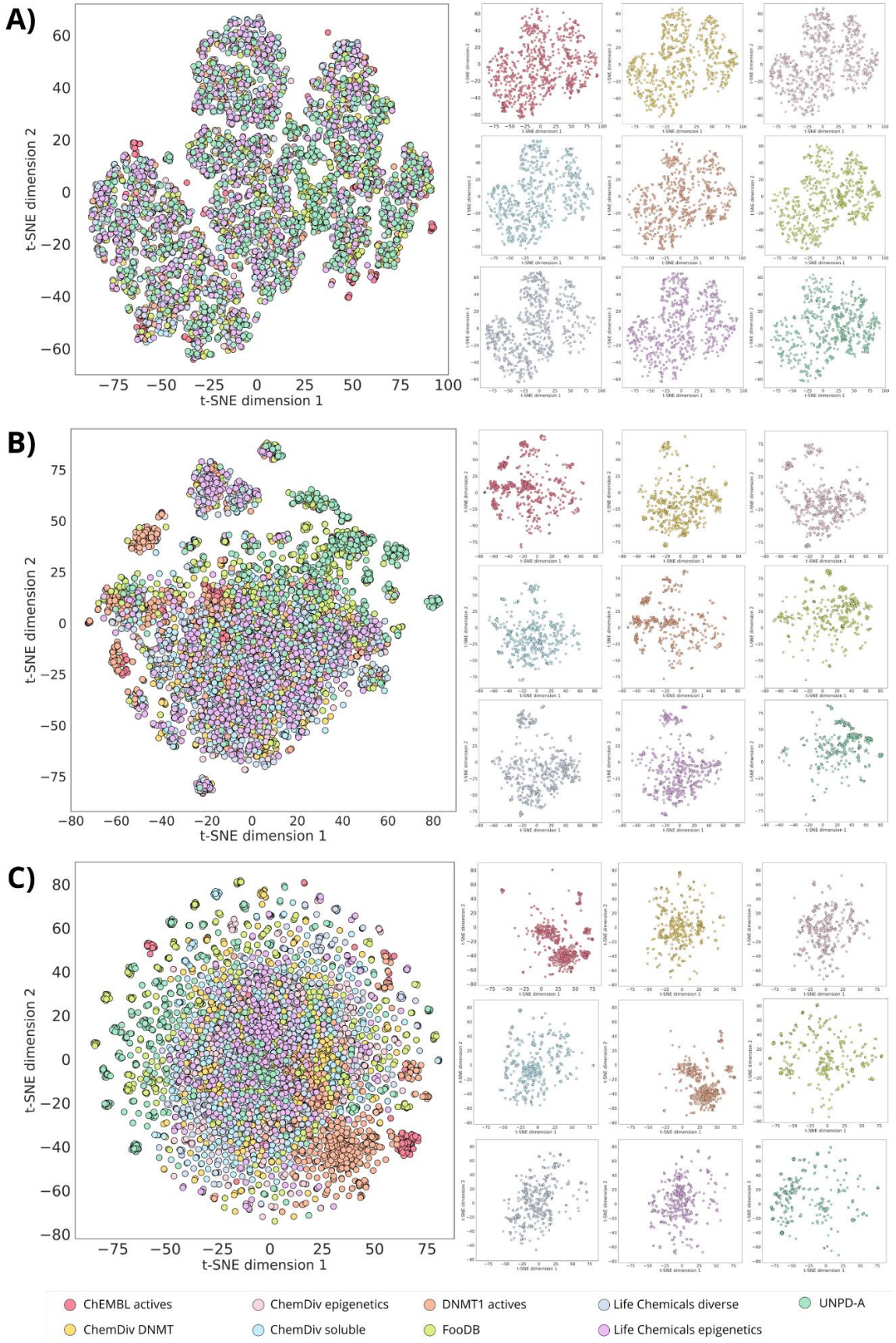

Figure 5. Chemical multiverse visualization of the *de novo* designed libraries with alvaBuilder. The visual representations were generated using t-SNE and **A**) drug-like properties, **B**) MACCS Keys (166-bits), and **C**) ECFP4 (1024-bits) as molecular representations. On the left are superimposed databases, and on the right are the individual data sets using the same coordinates as the corresponding representations on the left.

In contrast, the visualizations based on molecular fingerprints, MACCS Keys, and ECFP4 (Figure 5) suggest that *de novo* compounds designed with alvaBuilder populate different areas of the chemical space, which is more noticeable when compounds are represented with the ECFP4 fingerprint. Interestingly, compounds that were designed with information about compounds published in ChEMBL (DNMT1 actives, depicted in orange) are in the same coordinates of the reference database for both fingerprints.

The compounds designed from the three ChemDiv libraries and the two Life Chemicals libraries (Table 1) cover similar regions of the chemical space with the t-SNE based on MACCS Keys (Figure 5B). These observations are expected because chemical libraries from the same provider could share structural features detected with the MACCS Keys. However, differences are observed with the visualization based on ECFP4 (Figure 5C), associated with the differences in resolution between the two fingerprints. Compounds from FooDB and UNPD-A cover similar areas with both representations; the reason for this could be the account of the rest of the libraries being focused on small, less complex molecules.

Distribution of the QED values (Figure 6) showed that the median values of all databases, even the ChEMBL's DNMT1 inhibitors used as a reference, do not comply with the suggested threshold of 0.67 or higher. In this case, the mean from ChEMBL actives will be more relevant because the aim is to find new hits for DNMT1. Since hits are considered for earlier stages of drug discovery, biological activity normally carries more relevance than drug-like properties or potency. This could explain the lower values from known inhibitors with reported activity in ChEMBL. Nevertheless, QED values are useful for ranking and prioritizing the synthesis of *de novo* compounds or libraries.

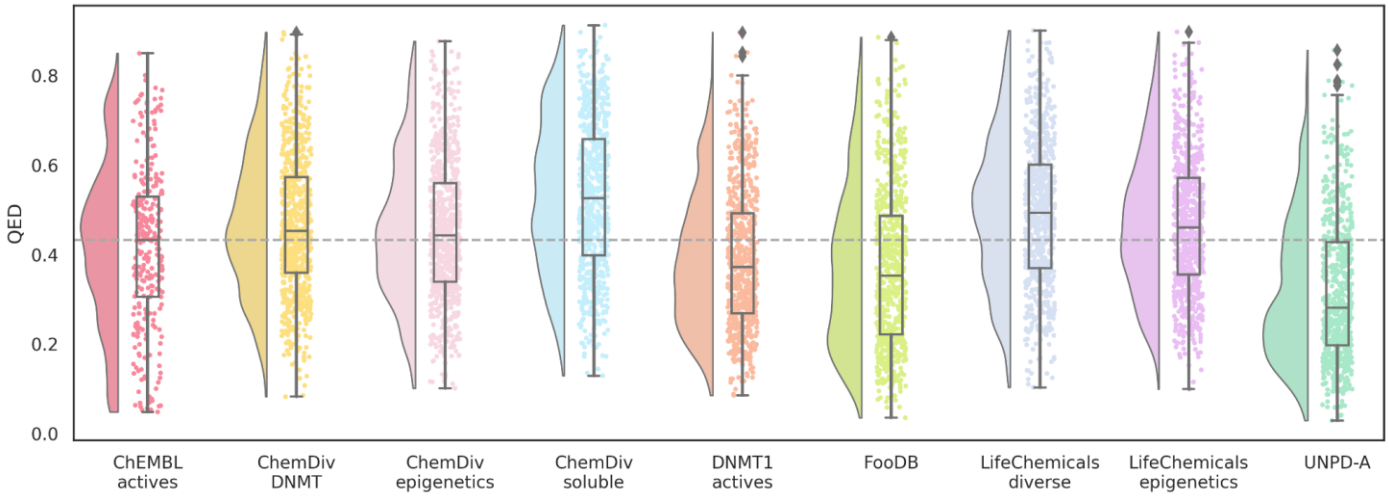

Figure 6. Distribution of QED values of *de novo* compounds designed with alvaBuilder. Each source of fragments is represented in a different color. Active compounds from ChEMBL are used as a reference (pastel red), and the mean value of these compounds is marked with a dotted gray line.

Remarkably, most of the designed databases have similar or better QED mean values than the ChEMBL's DNMT1 inhibitors. Molecules designed with DNMT1 inhibitors as the source of fragments decreased slightly. However, FooDB and UNPD-A molecules have a more pronounced change in their mean values. These could be inherent to the tendency of larger molecular complexity commonly encountered in natural products and food chemicals as compared to small drug-like molecules, represented by all other training sets. Thus, indirect evidence of alvaBuilder compounds inheriting structural characteristics from the building blocks is shown with these results.

Similarly, SAscore values for five datasets are close to the reference; molecules designed from ChemDiv soluble, FooDB, and UNPD-A present higher means (Figure 7). This result could be associated with the complexity penalty of SAscore. In the case of ChemDiv soluble, another hypothesis is that the fragments could be more difficult to obtain because solubility issues are usually part of the optimization process. Of note, only 19 molecules of the 5575 (0.34%) have a SAscore greater than six.

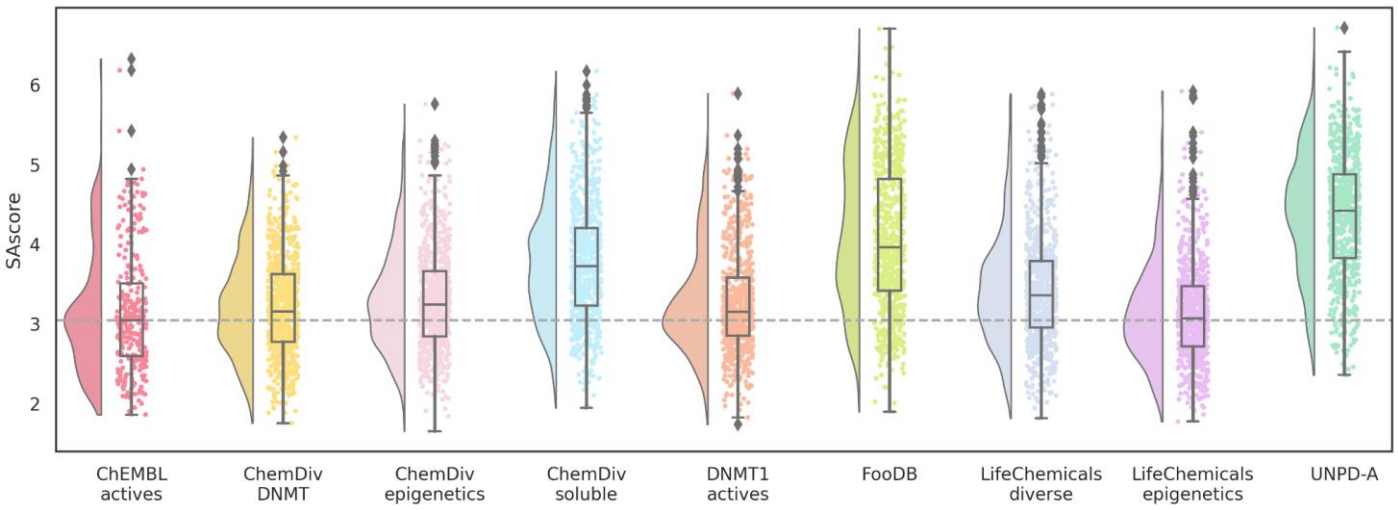

Figure 7. Distribution of SAscore values of *de novo* compounds designed with alvaBuilder. Each source of fragments is represented in a different color. Active compounds from ChEMBL are used as a reference (pastel red), and the mean value of these compounds is represented by a dotted gray line.

To compare the structural diversity of the generated databases from different fragment sources, the Tanimoto coefficient was used along with MACCS Keys (166-bits) and ECFP4 (1024 bits) fingerprints. According to MACCS Keys, molecules from UNPD-A are the most diverse, followed by FooDB and Life Chemicals diverse, which can be seen in the cumulative distribution functions (Figure 8) and values for mean and median similarity (Table S5 in the Supporting Information).

UNPD-A is also the most diverse according to the ECFP4 fingerprint, followed closely by the ChEMBL's DNMT1 inhibitors and molecules designed from FooDB (Table S6). ECFP4 similarity is considerably lower and has very close values among the nine databases, evidenced by the proximity of the curves (Figure 8). This is expected because ECFP4 codifies connectivity features along the selected radius, while MACCS Keys has a pre-defined list of features, giving the first more resolution [86].

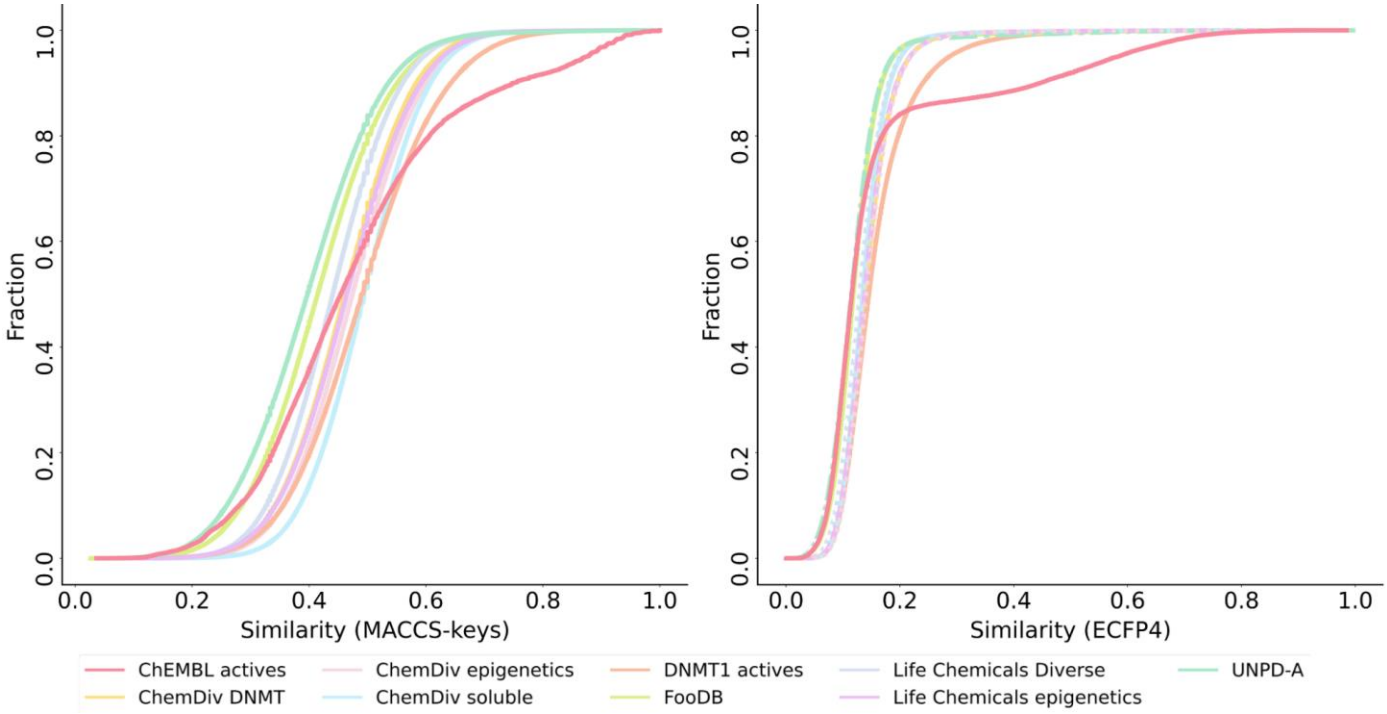

Figure 8. Cumulative distribution functions of the pairwise similarity values are computed with the Tanimoto coefficient and MACCS Keys 166-bits (left) and ECFP4 fingerprints (right) for the *de novo* libraries designed with alvaBuilder.

In the case of molecular docking scores, the nine alvaBuilder databases have median values around the reference database for LeDock, as well as for Vina (Figure 9). Compounds designed from FooDB and UNPD-A present less auspicious medians; this could be related to the molecular size. We have observed that ligand efficiency (LE), computed as the docking scores divided by the number of heavy atoms, improved the correlation between docking scores and biological activity. For that, LE values were calculated, and the results can be found in Tables S9 and S10. While means of FooDB and UNPD-A tend to the value of ChEMBL inhibitors with Vina LE, LeDock LE keeps the same trend as the scores.

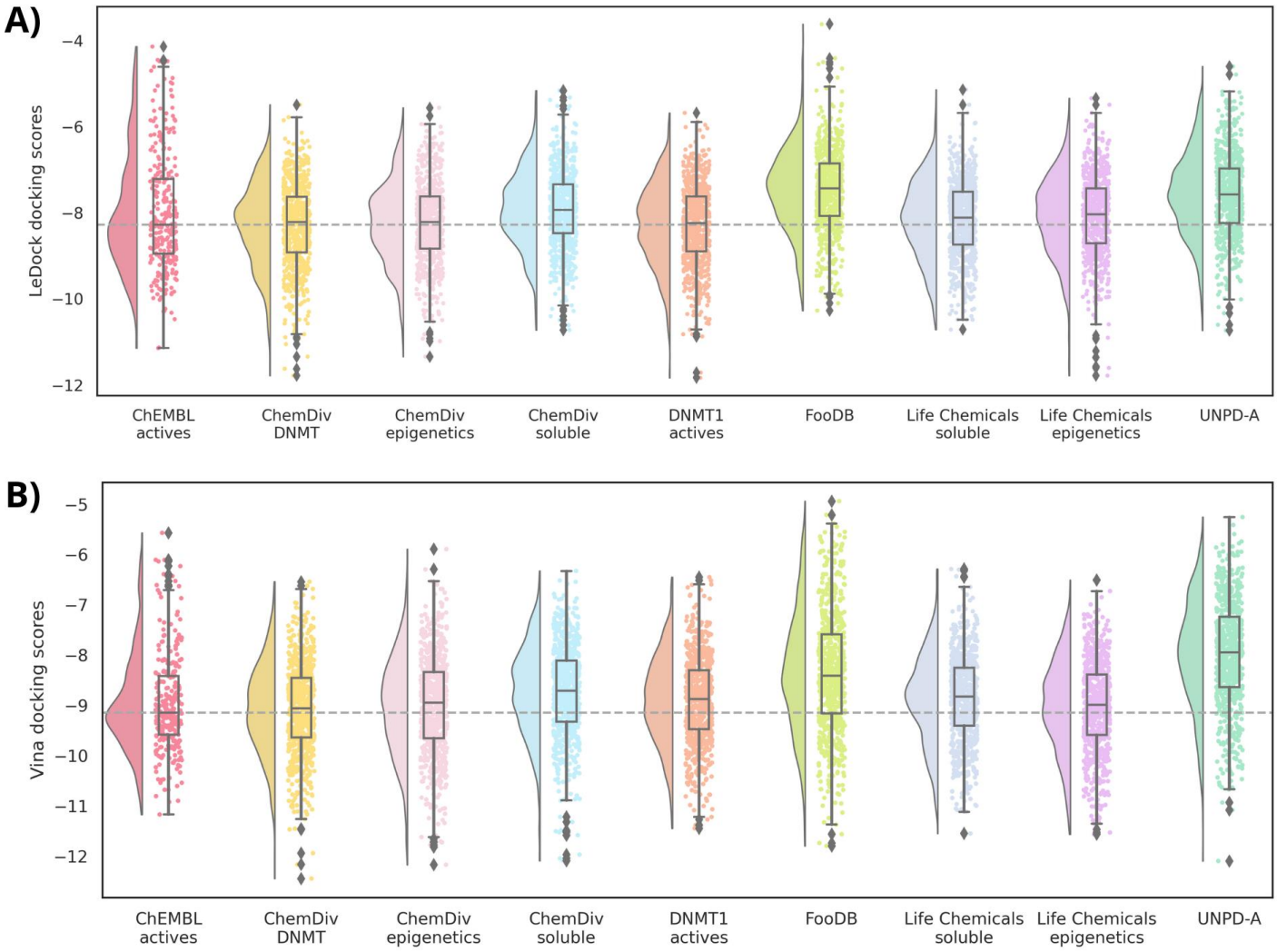

Figure 9. Raincloud plots summarizing the docking scores profile computed with **A**) LeDock and **B**) Vina for all compound data sets designed with AlvaBuilder. Each source of fragments is represented in a different color. Active compounds from ChEMBL are used as a reference (wine red). The mean value of these compounds is also marked with a dotted gray line.

3.2. Structure-based *de novo* design with LigBuilder

As detailed in the Methods, the molecular construction in LigBuilder was restricted by the characteristics of the binding pocket found with CAVITY. MW, logP, HBA, and HBD values from ChEMBL's DNMT1 inhibitors were also considered in the design. Both constraints were necessary to control the molecular size of the compounds designed. As a result, the chemical space visualizations based on drug-like properties (Figures 10 and 11A) show that the nine databases share the property-based chemical space, similar to the alvaBuilder compounds. Compounds from LigBuilder appear more diverse than the compounds designed with alvaBuilder. These results could be due to the limits of the molecular descriptors because the scoring function of alvaBuilder was more restricted with the median and standard deviation, while ranges for LigBuilder

included minimum and maximum descriptor values.

Likewise, designed compounds from ChEMBL´s DNMT1 inhibitors share the same areas of the reference database with the representation of both fingerprints, although exhibiting some distinct clusters (Figure 11). In general, molecules from the five commercial libraries exhibit distinguishable areas between all of them, both with MACCS Keys (Figure 11B) and ECFP4 (Figure 11 C). Also, for this library, FooDB compounds are populating similar areas to those designed from DNMT1 inhibitors, unlike alvaBuilder molecules that share more space with UNPD-A compounds. In contrast, molecules designed from UNPD-A fragments are less spread than the alvaBuilder ones. This could be due to a reduction in the number of molecules for this data set. Notably, Selleckchem compounds have a tighter cluster than the rest of the databases, especially for the visualization with ECFP4 (Figure 11C).

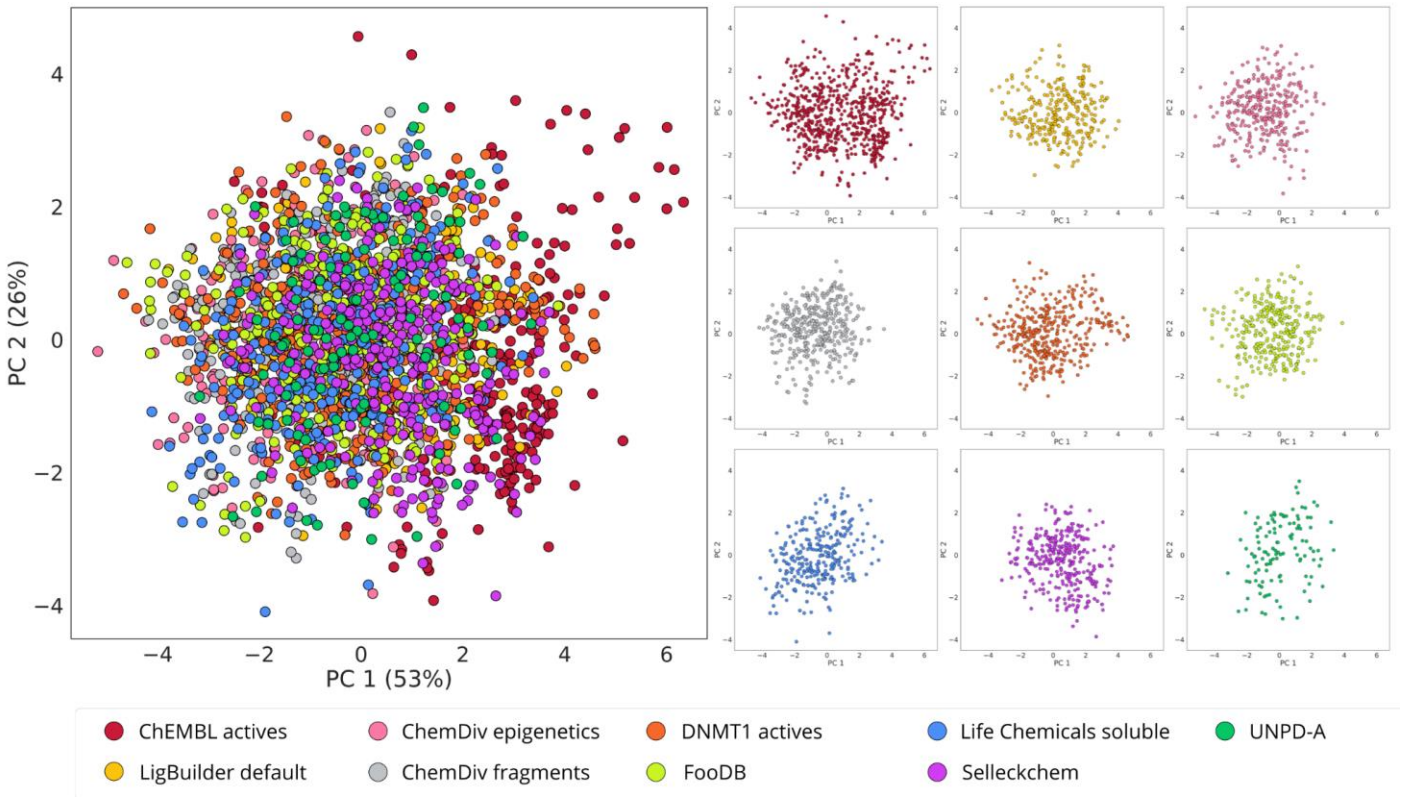

Figure 10. Chemical space visualization of the *de novo* designed libraries with LigBuilder. The visual representation was made with PCA as dimensionality reduction and six physicochemical properties as molecular representation. On the left are the nine superimposed databases, and on the right are the individual data sets using the same coordinates as the corresponding representations on the left.

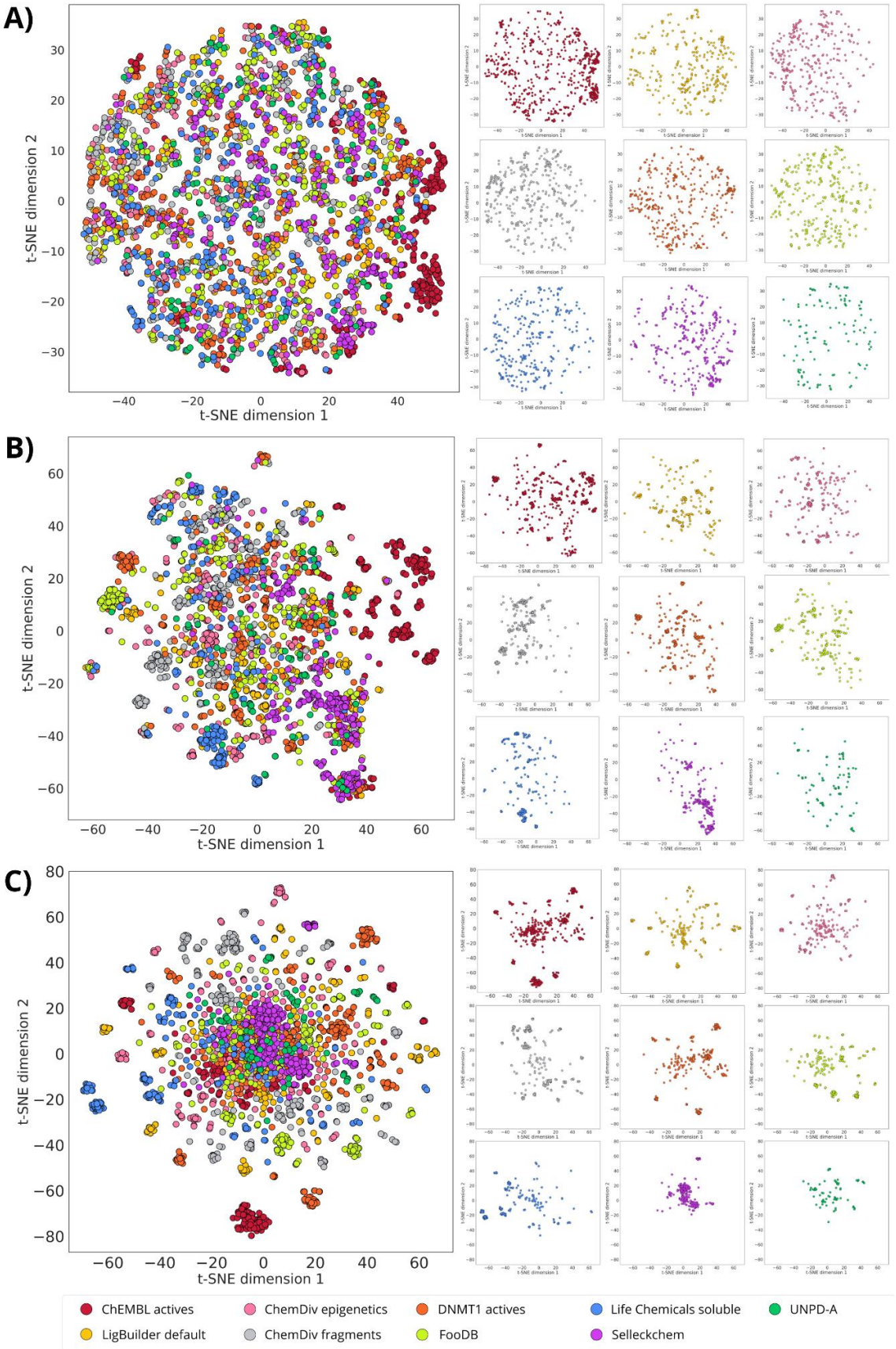

Figure 11. Chemical multiverse visualization of the *de novo* designed libraries with LigBuilder. The visual representations were generated using t-SNE and **A**) drug-like properties, **B**) MACCS Keys (166-bits), and **C**) ECFP4 (1024-bits) as molecular representations. On the left, the nine databases are superimposed, and on the right are the individual data sets using the same coordinates as their corresponding representations on the left.

The distribution of the QED values indicates that the nine databases have lower QED medians than the reference database and, consequently, less drug-like properties (Figure 12). Furthermore, the point distribution (one point represents one molecule) and the half-violin from the raincloud plot tend to have even lower QED values than the median. Since the visualization of chemical space based on drug-like properties indicated that these data sets populated similar areas, the decrease could be associated with the complexity penalty of QED. The increase in molecular complexity could be due to the characteristics of the binding cavity of DNMT1 since LigBuilder constructs the compounds in the binding site.

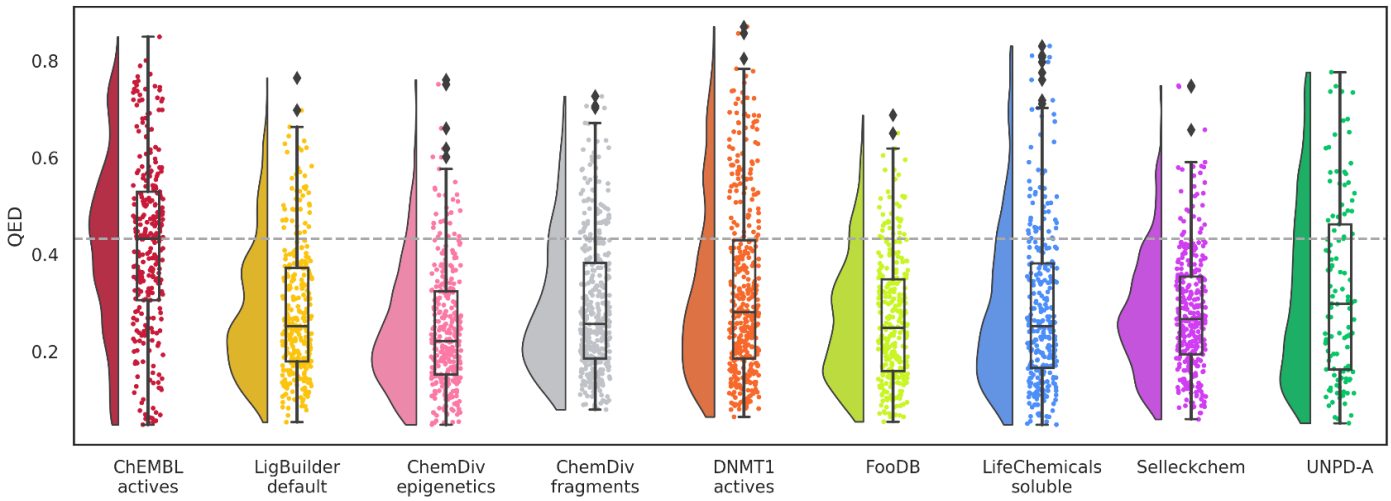

Figure 12. Distribution of QED values of *de novo* compounds designed with LigBuilder. Each source of fragments is represented in a different color. Active compounds from ChEMBL are used as a reference (wine red); the mean value of these compounds is also marked with a dotted gray line.

Equally, the nine databases have less favorable scores regarding the distribution of SAscore values (Figure 13). However, unlike the previous QED scores, most point distributions tend to cluster around the mean values. An increase in SAscore is expected and has been addressed as one of the issues with *de novo* design. Although the synthetic feasibility could be challenging, all the molecules are below the suggested threshold of six.

For the cumulative distribution functions computed with the Tanimoto coefficient and two fingerprints (MACCS Keys and ECFP4) (Figure 14), we found that the reference database of ChEMBL actives was the most diverse in both instances as confirmed by the median similarity values (Tables S7 and S8 in the Supplementary Material). In the case of MACCS Keys, the following most diverse data sets were molecules from UNPD-A and Selleckchem. In contrast, molecules from the Selleckchem library were the least diverse,

with ECFP4 as molecular representation. The most diverse database with this last fingerprint, after ChEMBL actives, were compounds constructed from this reference database (DNMT1 actives) and ChemDiv epigenetics.

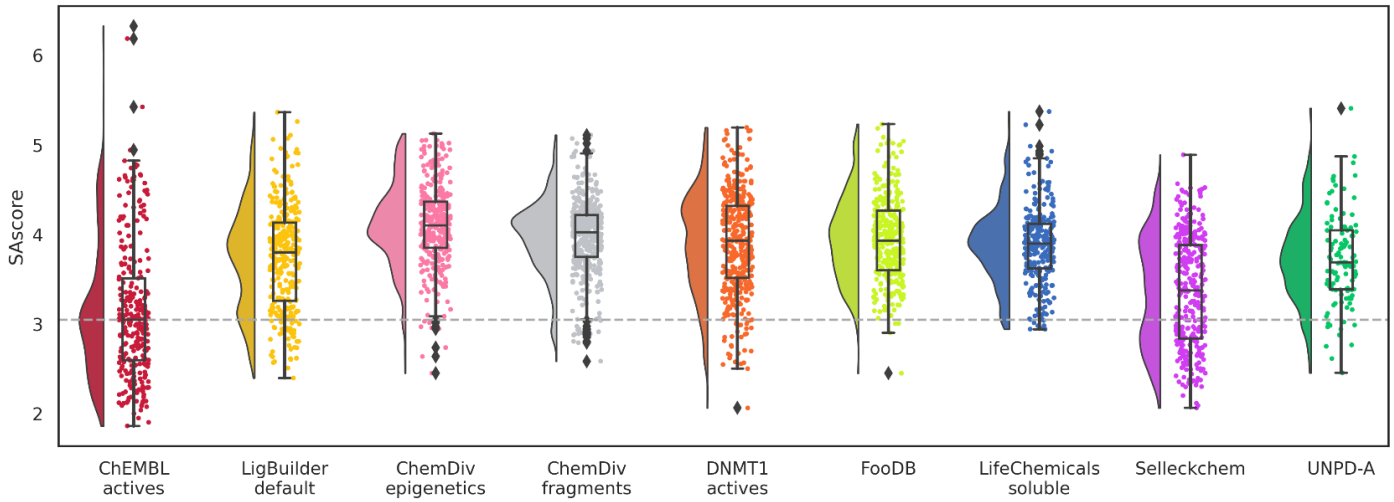

Figure 13. Distribution of SAscore values of *de novo* compounds designed with LigBuilder. Each source of fragments is represented in a different color. Active compounds from ChEMBL are used as a reference (wine red). The mean value of these compounds is represented with a dotted gray line.

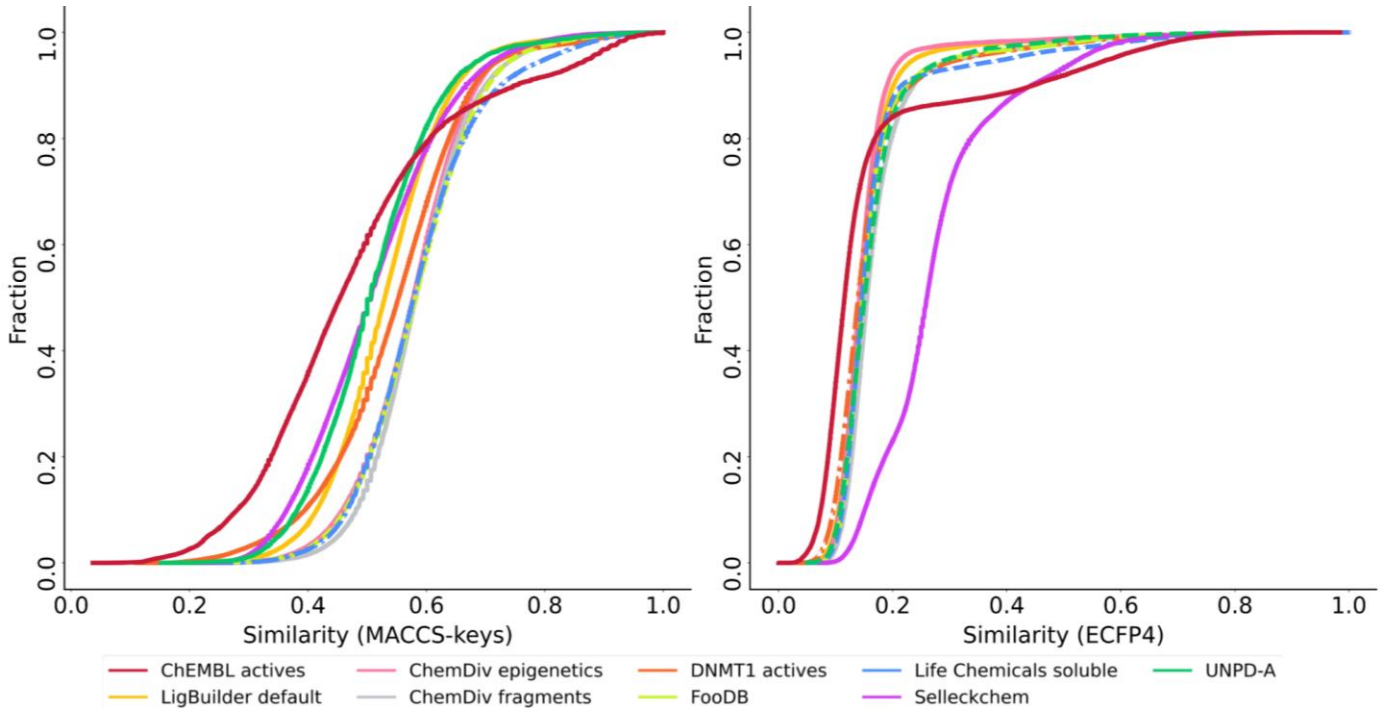

Figure 14. Cumulative distribution functions of the pairwise similarity values computed with the Tanimoto coefficient and MACCS keys 166-bits (left) and ECFP4 fingerprints (right) for the *de novo* libraries designed with LigBuilder.

Regarding the molecular docking scores, mean values are equal to or better than the computed scores for ChEMBL´s DNMT1 inhibitors for the case of LeDock (Figure 15A). In contrast, the eight designed databases have less favorable scores with Vina, except for Selleckchem compounds that equal the reference median. Although the numerical variations are around one point, these differences are not considered significant. LE values (Tables S11 and S12 in the Supplementary Material) also do not have marked changes in the numerical values.

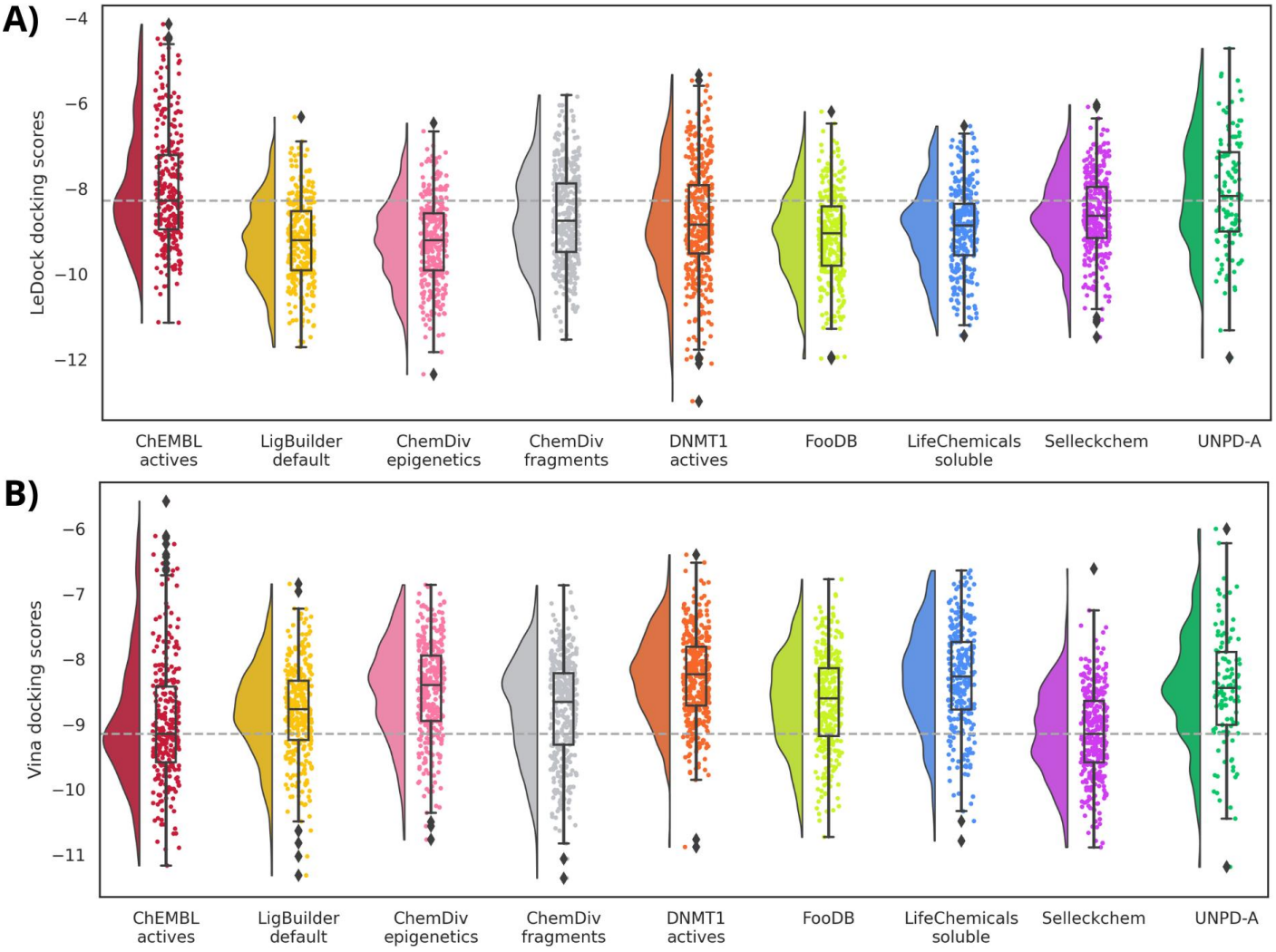

Figure 15. Raincloud plots summarizing the docking scores profile computed with **A**) LeDock and **B**) Vina for all compound data sets designed with LigBuilder. Each source of fragments is represented in a different color. Active compounds from ChEMBL are used as a reference (wine red). The mean value of these compounds is also marked with a dotted gray line.

3.3 Global diversity and scaffold analysis

As described in the Methods section, we compared the diversity of the *de novo-designed* libraries and the DNMT1 inhibitor dataset in terms of fingerprints, molecular scaffolds, and properties of pharmaceutical interest employing CDP [72]. Figure 16 shows a CDP comparing the overall (global) structural diversity of all three data sets, considering four parameters described in Table 3. Every point in this plot corresponds to a single dataset. The median corresponding to each data set, computed with MACCS Keys/Tanimoto, is plotted on the X-axis. The area under the curve (AUC) for the scaffold recovery curves is plotted on the Y-axis. The size of the data points represents the relative sizes of each data set, and the color of each data point represents the diversity of molecular properties.

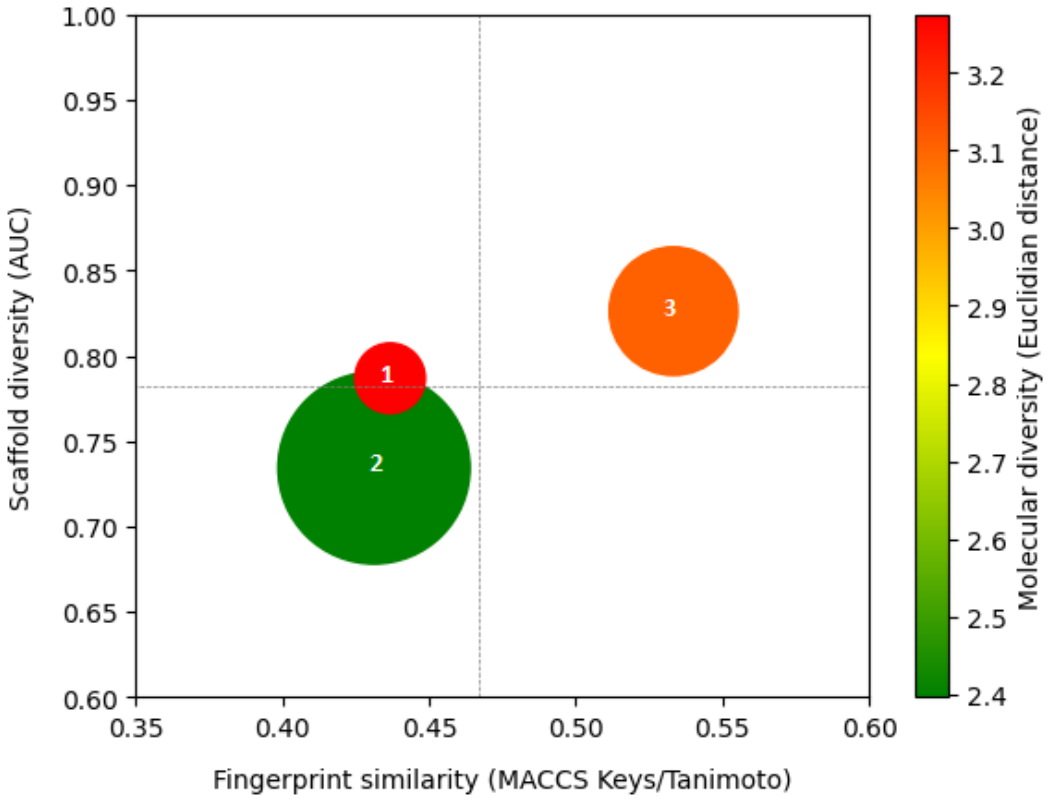

Figure 16. Consensus Diversity Plot comparing the diversity of DNMT1 inhibitors (**1**) and *de novo* compounds generated by alvaBuilder (**2**) and LigBuilder (**3**) software. The median similarity computed with MACCS Keys and the Tanimoto coefficient of the data set is plotted on the X axis and the AUC of the scaffold recovery curves on the Y axis. Data points are colored by the diversity of the physicochemical properties of the data set as measured by the Euclidean distance of six properties of pharmaceutical relevance. The distance is represented with a continuous color scale from red (more diverse) to orange/brown (intermediate diversity) to green (less diverse). The data point's size represents the data set's relative size: smaller data points indicate compound data sets with fewer molecules.

According to Figure 16, the compounds designed by the alvaBuilder software are the most diverse as measured by fingerprints, which assess diversity based on complete structures. Regarding scaffold diversity, compounds designed by the alvaBuilder program are also the most diverse, followed by the DNMT1 inhibitors and compounds generated by LigBuilder. Furthermore, compounds designed with LigBuilder are also the most diverse in terms of molecular properties, only behind the DNMT1 inhibitors.

Table 3. Global diversity analysis of the newly designed compound libraries.

DATASET	Code CDP	Size	ECFP4 [a]	MACCS Keys [a]	N/M	AUC	F_{50}	SSE10	Molecular properties
DNMT1	1	743	0.140	0.437	0.634	0.787	0.045	0.927	3.275
alvaBuilder	2	5575	0.148	0.431	0.795	0.734	0.138	0.739	2.396
LigBuilder	3	2491	0.171	0.533	0.501	0.826	0.020	0.939	3.108

M: Number of molecules, N: number of scaffolds, AUC: area under the curve, F_{50}:Fraction of chemotypes that contain 50% of the dataset, SEE10: Scaled Shannon Entropy at the ten most frequent scaffolds. [a] Median of the pairwise fingerprint similarity distribution.

Table 3 further summarizes additional metrics of scaffold diversity, such as the ratio (N/M), $F_{50,}$ and Scaled Shannon Entropy (SSE). Unlike the CSR curves, which assess the diversity of entire datasets, SSE measures the scaffold diversity of the most populated scaffolds. The SSE value ranges from 0, indicating uniformity in chemotype distribution across all compounds (minimum diversity), to 1.0 when all the compounds are evenly distributed among the n acyclic and/or cyclic systems (maximum diversity). According to SSE values depicted in Table 3 for the ten most frequent scaffolds, compounds designed with LigBuilder are the most diverse, followed by DNMT1 inhibitors and compounds designed by alvaBuilder. Figure 17 depicts an overview of each database's ten most frequent scaffolds. Unique scaffolds for each database are highlighted in blue. The prevalence of acyclic compounds was notably higher in DNMT1 inhibitors (5.24%) and compounds designed by alvaBuilder (2.76%). LigBuilder, conversely, generated only 19 (0.76%) acyclic compounds. Among the 472 scaffolds identified in DNMT1 inhibitors, 22 are present in compounds generated by alvaBuilder, five in compounds generated by LigBuilder, and three are shared between both libraries.

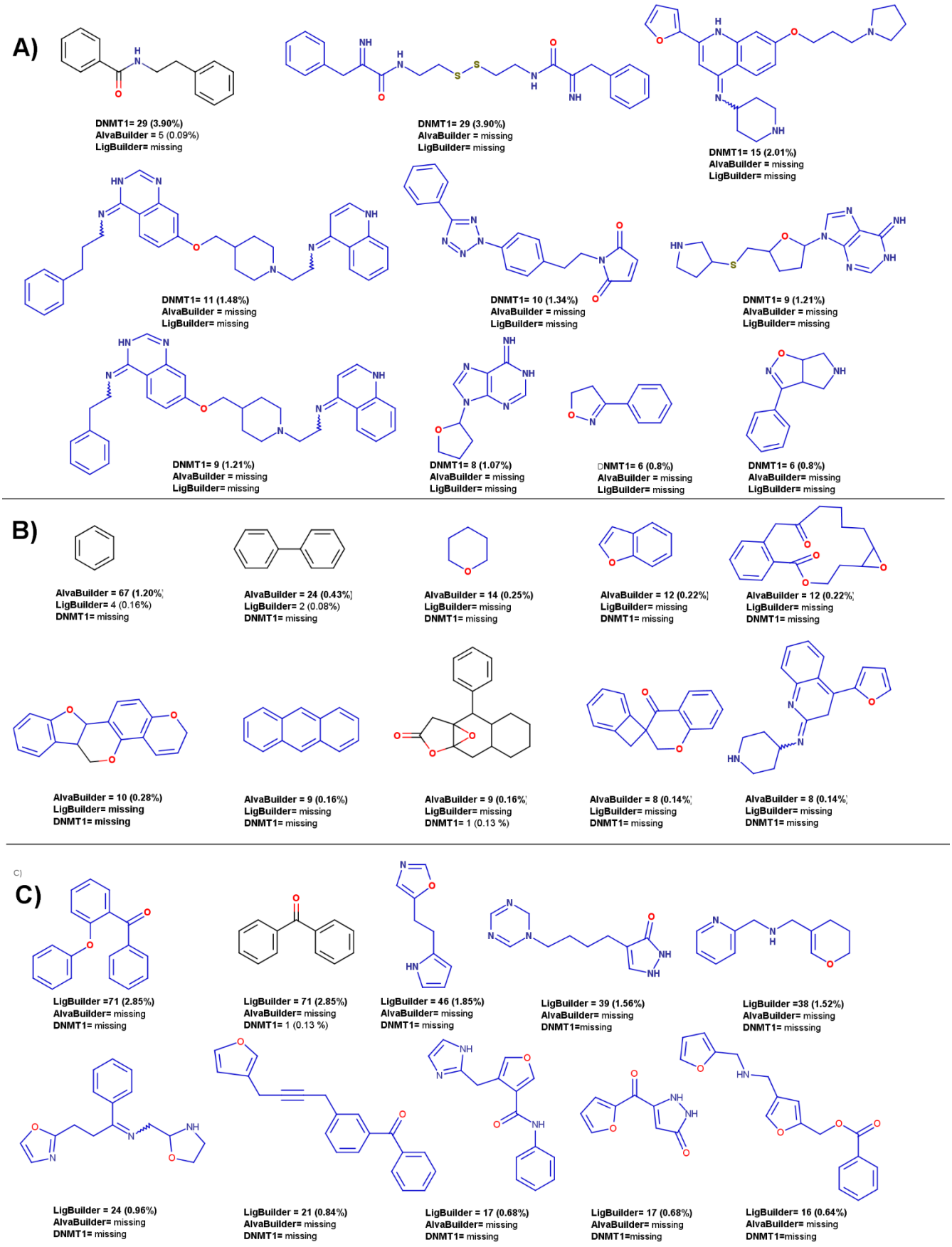

Figure 17. Ten most frequent scaffolds from **A**) DNMT1 inhibitors and, libraries designed with **B**) alvaBuilder and **C**) LigBuilder. The presence and frequency in each of the three data sets is indicated. Unique scaffolds for each set of compounds are indicated in blue.

3.4. Classification models

Fifteen optimized binary classification models were constructed to predict the activity class of compounds tested against DNMT1. These models resulted from training five machine learning algorithms using three different molecular fingerprints as descriptors as detailed in the Methods section. Each model is represented as a combination of a fingerprint and algorithm (e.g., algorithm + fingerprint). Hyperparameters for each model were optimized through an exhaustive search employing LOO-CV with BA as the performance metric for selecting the best set of hyperparameters.

Overall, most of the optimized models exhibited strong performance, with a mean BA score exceeding 0.6, with the models SVM + RDK and FFNN + Morgan having the best performance on the training set with BA values of 0.849 and 0.847 respectively, and consistent performances on the external test set with BA values of 0.775 and 0.793 respectively. Table S13 in the Supplementary Material includes summary statistics for all models.

Although BA is suitable for assessing model performance on imbalanced datasets, correctly identifying active compounds is prioritized in practical medicinal chemistry applications. Therefore, individual models were evaluated in terms of their precision and recall, which represents the model's ability to correctly classify active compounds, computed on a DM basis as detailed in the Methods Section. These results are summarized in Table 4.

Table 4. Distance-to-model performance of classification models on the training set.

Quartile	Number of active compounds		SVM + RDK		FFNN + Morgan	
	Active	Inactive	Precision	Recall	Precision	Recall
Q1	44	1	0.978	0.978	0.978	1.000
Q2	40	5	0.974	0.950	0.950	0.950
Q3	20	25	0.667	0.300	0.667	0.600
Q4	9	36	0.667	0.222	0.600	0.333

For both models, precision and recall values decrease as long as the predicted compounds become more distant from those in the training set. However, even for distant compounds (Q3 and Q4), precision remains

above 0.6, which is better than a random guess considering the data imbalance on each quartile (20 / 45 for Q3 and 9 / 45 for Q4). In terms of recall, the FFNN + Morgan model shows the best performance in Q3 and Q4, making it the best choice for selecting active compounds distant from those in the training set. Distance-to-model performance for the test set is included in Table S14, which shows a similar behavior. However, considering the limited number of compounds on each quartile, they tended to be over-optimistic (e.g., perfect precision at Q4) and were not considered for the final classification of compounds.

Both models were retrained in the entire dataset and applied to predict the activity class of the compounds generated by *de novo* design, and their DM was calculated. The number of compounds predicted as active for each quartile and dataset are summarized in Table 5 and Table 6 for SVM + RDK and FFNN + Morgan, respectively. Overall, most of the compounds generated by *de novo* design were distant to the training set, with just two of them falling in Q2, 63 of them falling in Q3, and 7842 falling in Q4.

The two models predicted Eight compounds as active; their chemical structures are shown in Figure 18. Significantly, all have “long or extended scaffolds,” a feature highlighted for active molecules against DNMT1 [15,87,88]. Furthermore, the eight molecules are quinolines with similar substitution patterns, and this family of compounds has been reported before by several research groups focused on DNMT inhibitors [89–92].

Moreover, 159 compounds obtained mean Jaccard distances higher than those observed for Q4, which involves an unknown confidence in their predictions. The SVM + RDK model predicted only 15 compounds as active against DNMT1, all generated by alvaBuilder, particularly from the ChemDiv DNMT dataset**.** On the other hand, the FFNN + Morgan model predicted active compounds for all groups of compounds generated by *de novo* design, with marked differences.

Table 5. Compounds predicted as active by the SVM + RDK model.

Database	Quartile			Proportion
	Q2	Q3	Q4	
alvaBuilder - ChemDiv DNMT	1	6	8	0.024

Table 6. Compounds predicted as active by the FFNN + Morgan model.

Database	Quartile			Proportion
	Q2	Q3	Q4	
alvaBuilder				
ChemDiv DNMT	0	0	130	0.186
ChemDiv epigenetics	0	0	110	0.157
ChemDiv soluble	0	0	97	0.139
DNMT1 actives	1	12	141	0.220
FooDB	0	0	97	0.139
Life Chemicals diverse	0	0	132	0.189
Life Chemicals epigenetics	0	0	111	0.159
UNPD-A	0	0	137	0.202
LigBuilder				
ChemDiv epigenetics	0	0	16	0.049
ChemDiv fragments	0	0	3	0.007
ChemDiv soluble	0	0	7	0.024
LigBuilder default	0	0	13	0.042
DNMT1 actives	0	0	37	0.094
FooDB	0	0	7	0.022
Selleckchem	0	0	9	0.027
UNPD-A	0	0	6	0.051

For all the groups generated by alvaBuilder, the proportion of compounds predicted as active was higher than 0.13, with DNMT1 actives and UNPD-A being the datasets with higher proportions, while for compounds generated by LigBuilder, this proportion was always below 0.10. This observation could be associated with the differences in fragment acquisition for each software. Namely, alvaBuilder incorporates an internal

fragmentation algorithm based on Bemis-Murcko scaffolds, while LigBuilder directly takes fragments from the library.

As a result of the alvaBuilder fragmentation, the information about the position of the linkers is kept so the complete molecules can be reconstructed from their corresponding fragments. This preserves chemical knowledge inherent to the scaffold and the substitution pattern in the fragments. In contrast, information about the original substitution of a fragment in LigBuilder has to be user-defined as a growing site. That information was not disclosed for the selected fragment libraries and was only available for the RECAP fragments.

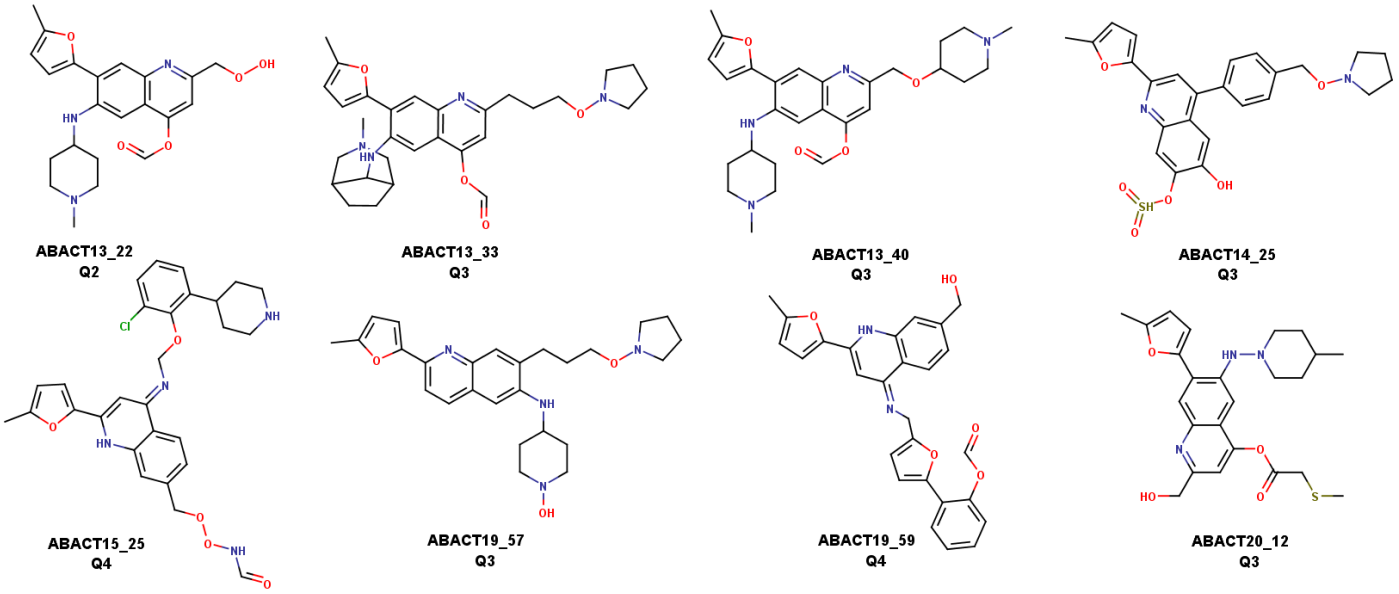

Figure 18. Chemical structures of the eight compounds predicted as active by the two classification models.

4. Conclusions and perspectives

In this study, two *de novo* libraries focused on the epigenetic target DNMT1 were designed with 5575 and 2491 compounds, respectively. The larger library was created with AlvaBuilder, a ligand-based program, while the other was constructed with LigBuilder, software based on protein structure. The number of compounds was smaller since the structure-based strategy took more time and computational resources.

The visualization of the chemical spaces showed that *de novo* compounds kept the physicochemical properties of the already known DNMT1 active inhibitors reported in ChEMBL 31. This could be associated with the secondary constraints used in both strategies for the design, in the scoring function of alvaBuilder and the parameters for the BUILD module in LigBuilder. These results are encouraging, since the visualizations with the two fingerprints used as molecular representations suggest that newly designed

compounds populate different areas of the chemical space. Therefore, the new focused libraries are relevant for expanding the chemical space of DNMT1 inhibitors while keeping drug-like properties.

In general, QED and SAscore values for alvaBuilder molecules were alike or better than the reference database (ChEMBL´s DNMT1 inhibitors). Designed molecules from FooDB and UNPD-A exhibit decreased values of both scores, possibly because of the higher molecular complexity of food chemicals and natural products compared to small molecules of synthetic origin. Means from LigBuilder molecules were below the reference database for both QED and SAscore. These relative values could be associated with the penalty of molecular complexity and inherent synthetic feasibility problems related to *de novo* design [3].

Compounds designed with alvaBuilder were the most diverse based on MACCS Keys, ECFP4 fingerprints, and scaffolds.

Results from the classification evidenced that predicted active molecules with both models are quinolines. Moreover, compounds designed from UNPD-A with both programs had the second-highest proportion of active compounds after those designed from ChEMBL´s DNMT1 inhibitors. This suggests that, even though the QED and SAscore are less favorable for UNPD-A compounds, the activity could be the one searched, making this data set suitable for the beginning of an optimization project. Finally, there is evidence that the biological activity information from the ChEMBL´s DNMT1 actives is kept, since compounds constructed from them had the highest proportion of actives for both *de novo* software. Medians of the molecular docking scores of the total of designed molecules had the same trend.

The newly focused libraries developed in this study contain a complete description of the methodology used to construct them and represent a notable addition to the commercial epigenetic-focused libraries currently available. In agreement with open science and its symbiosis with artificial intelligence and other computational applications [93], all fragment libraries generated in this work are freely available and can be used for further virtual and experimental screening for epigenetic drug discovery, emphasizing DNMT1. The ligand- and structure-based *de novo* design strategies implemented in this work are general and could be used to build screening libraries focused on other epigenetic targets.

Overall, both focused libraries have different profiles and represent valuable starting points for expanding the chemical space of DNMT1 inhibitors. AlvaBuilder and LigBuilder compounds generally preserve molecular properties calculated from known DNMT1 inhibitors. Compounds from the ligand-based strategy are more

diverse than those from the LigBuilder ones and exhibit better profiles of QED and SAscore. However, both libraries had similar results with docking scores and had predicted actives with the classification model. Since LigBuilder compounds are restricted to the binding pocket, the diversity could diminish. Meanwhile, alvaBuilder compounds also had higher proportions of actives with the classification models, probably due to the fragment characteristics. This could also translate into more novelty for the LigBuilder molecules. One library or another could be chosen depending on the objective of the discovery project.

Perspectives of this study include the chemical synthesis and testing of the entire compound libraries or selected compounds, and further virtual screening of the *de novo* designed libraries to select additional individual compounds for synthesis and testing. Similarity searches in commercial libraries for later acquisition and experimental screening with DNMT1 in enzymatic inhibition and other assays are underway in our research group.

Acknowledgments

DL P-R and FI S-G thank *Consejo Nacional de Humanidades, Ciencias y Tecnologías* (CONAHCyT), Mexico, for the postgraduate scholarships 888207 and 848061. The invaluable discussions, code-sharing, and technical support of Juan F. Avellaneda-Tamayo, Ana L. Chávez-Hernández, Raziel Cedillo-González, and Alejandro Gómez-García are greatly acknowledged. We also thank Ana L. Chávez-Hernández and Juan F. Avellaneda-Tamayo for providing the curated FooDB and UNPD-A chemical library datasets and the resulting fragments of both. Thank you to Dr. Matteo Bertola and Dr. Yaxia Yuan for their recommendations regarding the use of alvaBuilder and LigBuilder. We also acknowledge alvaScience for the support in providing the alvaBuilder license and ChemAxon for providing MarvinSketch that was used for drawing and displaying chemical structures, MarvinSketch 22.18, ChemAxon (https://www.chemaxon.com).

Funding

This project was funded by DGAPA, UNAM, *Programa de Apoyo a Proyectos de Investigación e Innovación Tecnológica* (PAPIIT), grants No. IN201321 and IG200124. Miztli supercomputer at UNAM, project number LANCAD-UNAM-DGTIC-335.

References

1. Jiménez-Luna, J.; Grisoni, F.; Weskamp, N.; Schneider, G. Artificial Intelligence in Drug Discovery: Recent Advances and Future Perspectives. *Expert Opin. Drug Discov.* **2021**, *16*, 949–959.
2. Liu, X.; IJzerman, A.P.; van Westen, G.J.P. Computational Approaches for De Novo Drug Design: Past, Present, and Future. *Methods Mol. Biol.* **2021**, *2190*, 139–165.
3. Mouchlis, V.D.; Afantitis, A.; Serra, A.; Fratello, M.; Papadiamantis, A.G.; Aidinis, V.; Lynch, I.; Greco, D.; Melagraki, G. Advances in de Novo Drug Design: From Conventional to Machine Learning Methods. *Int. J. Mol. Sci.* **2021**, *22*, 1676.
4. Hartenfeller, M.; Schneider, G. Enabling Future Drug Discovery by *de Novo* Design. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2011**, *1*, 742–759.
5. Focused and Targeted Libraries Available online: https://www.chemdiv.com/catalog/focused-and-targeted-libraries/ (accessed on 2 November 2023).
6. Active Reference Sets Available online: https://www.chemdiv.com/catalog/sets/ (accessed on 2 November 2023).
7. Zhang, Z.; Wang, G.; Li, Y.; Lei, D.; Xiang, J.; Ouyang, L.; Wang, Y.; Yang, J. Recent Progress in DNA Methyltransferase Inhibitors as Anticancer Agents. *Front. Pharmacol.* **2022**, *13*, 1072651.
8. Yu, J.; Xie, T.; Wang, Z.; Wang, X.; Zeng, S.; Kang, Y.; Hou, T. DNA Methyltransferases: Emerging Targets for the Discovery of Inhibitors as Potent Anticancer Drugs. *Drug Discov. Today* **2019**, *24*, 2323–2331.
9. Zhu, J.; Yang, Y.; Li, L.; Tang, J.; Zhang, R. DNA Methylation Profiles in Cancer: Functions, Therapy, and beyond. *Cancer Biol Med* **2023**, *21*, 111–116.
10. Sullivan, M.; Hahn, K.; Kolesar, J.M. Azacitidine: A Novel Agent for Myelodysplastic Syndromes. *Am. J. Health. Syst. Pharm.* **2005**, *62*, 1567–1573.
11. Derissen, E.J.B.; Beijnen, J.H.; Schellens, J.H.M. Concise Drug Review: Azacitidine and Decitabine. *Oncologist* **2013**, *18*, 619–624.
12. Zwergel, C.; Valente, S.; Mai, A. DNA Methyltransferases Inhibitors from Natural Sources. *Curr. Top. Med. Chem.* **2016**, *16*, 680–696.
13. Saldívar-González, F.I.; Gómez-García, A.; Chávez-Ponce de León, D.E.; Sánchez-Cruz, N.; Ruiz-Rios, J.; Pilón-Jiménez, B.A.; Medina-Franco, J.L. Inhibitors of DNA Methyltransferases From Natural Sources: A Computational Perspective. *Front. Pharmacol.* **2018**, *9*, 1144.
14. 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*, e2100285.
15. Juárez-Mercado, K.E.; Prieto-Martínez, F.D.; Sánchez-Cruz, N.; Peña-Castillo, A.; Prada-Gracia, D.; Medina-Franco, J.L. Expanding the Structural Diversity of DNA Methyltransferase Inhibitors. *Pharmaceuticals* **2020**, *14*, 17.
16. Medina-Franco, J.L.; López-López, E.; Martínez-Fernández, L.P. 7-Aminoalkoxy-Quinazolines from Epigenetic Focused Libraries Are Potent and Selective Inhibitors of DNA Methyltransferase 1. *Molecules*

2022, *27*, 2892.

17. Wang, X.-S.; Zheng, Q.-C. Theoretical Research in Structure Characteristics of Different Inhibitors and Differences of Binding Modes with CBP Bromodomain. *Bioorg. Med. Chem.* **2018**, *26*, 712–720.
18. Prado-Romero, D.L.; Medina-Franco, J.L. Advances in the Exploration of the Epigenetic Relevant Chemical Space. *ACS Omega* **2021**, *6*, 22478–22486.
19. Prado-Romero, D.L.; Gómez-García, A.; Cedillo-González, R.; Villegas-Quintero, H.; Avellaneda-Tamayo, J.F.; López-López, E.; Saldívar-González, F.I.; Chávez-Hernández, A.L.; Medina-Franco, J.L. Consensus Docking Aid to Model the Activity of an Inhibitor of DNA Methyltransferase 1 Inspired by de Novo Design. *Front. Drug Des. Discovery* **2023**, *3*, doi:10.3389/fddsv.2023.1261094.
20. Lanka, G.; Banerjee, S.; Adhikari, N.; Ghosh, B. Fragment-Based Discovery of New Potential DNMT1 Inhibitors Integrating Multiple Pharmacophore Modeling, 3D-QSAR, Virtual Screening, Molecular Docking, ADME, and Molecular Dynamics Simulation Approaches. *Mol. Divers.* **2024**, doi:10.1007/s11030-024-10837-5.
21. Mauri, A.; Bertola, M. AlvaBuilder: A Software for De Novo Molecular Design. *J. Chem. Inf. Model.* **2023**, *64*, 2136–2142.
22. Yuan, Y.; Pei, J.; Lai, L. LigBuilder V3: A Multi-Target de Novo Drug Design Approach. *Front. Chem.* **2020**, *8*, 142.
23. Lewell, X.Q.; Judd, D.B.; Watson, S.P.; Hann, M.M. RECAP--Retrosynthetic Combinatorial Analysis Procedure: A Powerful New Technique for Identifying Privileged Molecular Fragments with Useful Applications in Combinatorial Chemistry. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 511–522.
24. Bickerton, G.R.; Paolini, G.V.; Besnard, J.; Muresan, S.; Hopkins, A.L. Quantifying the Chemical Beauty of Drugs. *Nat. Chem.* **2012**, *4*, 90–98.
25. Ertl, P.; Schuffenhauer, A. Estimation of Synthetic Accessibility Score of Drug-like Molecules Based on Molecular Complexity and Fragment Contributions. *J. Cheminform.* **2009**, *1*, 8.
26. Eberhardt, J.; Santos-Martins, D.; Tillack, A.F.; Forli, S. AutoDock Vina 1.2.0: New Docking Methods, Expanded Force Field, and Python Bindings. *J. Chem. Inf. Model.* **2021**, *61*, 3891–3898.
27. Trott, O.; Olson, A.J. AutoDock Vina: Improving the Speed and Accuracy of Docking with a New Scoring Function, Efficient Optimization, and Multithreading. *J. Comput. Chem.* **2010**, *31*, 455–461.
28. Lephar Research Software Available online: http://www.lephar.com/software.htm (accessed on 12 July 2023).
29. Sánchez-Cruz, N.; Medina-Franco, J.L. Epigenetic Target Fishing with Accurate Machine Learning Models. *J. Med. Chem.* **2021**, *64*, 8208–8220.
30. Sánchez-Cruz, N.; Pilón-Jiménez, B.A.; Medina-Franco, J.L. Functional Group and Diversity Analysis of BIOFACQUIM: A Mexican Natural Product Database. *F1000Res.* **2019**, *8*, doi:10.12688/f1000research.21540.2.
31. DIFACQUIM *IFG_General: Repository for the Work Functional Group and Diversity Analysis of BIOFACQUIM: A Mexican Natural Product Database*; Github, 2020.
32. Landrum, G. RDKit: Open-Source Cheminformatics Available online: https://www.rdkit.org (accessed on

5 December 2023).
33. MolVS: Molecule Validation and Standardization — MolVS 0.1.1 Documentation Available online: https://molvs.readthedocs.io/en/latest/ (accessed on 5 December 2023).
34. Sun, J.; Jeliazkova, N.; Chupakin, V.; Golib-Dzib, J.-F.; Engkvist, O.; Carlsson, L.; Wegner, J.; Ceulemans, H.; Georgiev, I.; Jeliazkov, V.; et al. ExCAPE-DB: An Integrated Large Scale Dataset Facilitating Big Data Analysis in Chemogenomics. *J. Cheminform.* **2017**, *9*, 17.
35. López-López, E.; Fernández-de Gortari, E.; Medina-Franco, J.L. Yes SIR! On the Structure-Inactivity Relationships in Drug Discovery. *Drug Discov. Today* **2022**, *27*, 2353–2362.
36. Mauri, A. alvaDesc: A Tool to Calculate and Analyze Molecular Descriptors and Fingerprints. In *Ecotoxicological QSARs*; Roy, K., Ed.; Springer US: New York, NY, 2020; pp. 801–820.
37. Congreve, M.; Carr, R.; Murray, C.; Jhoti, H. A "Rule of Three" for Fragment-Based Lead Discovery? *Drug Discov. Today* **2003**, *8*, 876–877.
38. Alvascience, alvaMolecule (software to View and Prepare Chemical Datasets) Version 1.0.4, 2020 Available online: https://www.alvascience.com/ (accessed on 15 November 2023).
39. Selecting Diverse Sets Of Compounds. In *An Introduction To Chemoinformatics*; Leach, A.R., Gillet, V.J., Eds.; Springer Netherlands: Dordrecht, 2007; pp. 119–139.
40. Jaccard, P. The Distribution of the Flora in the Alpine Zone. *New Phytol.* **1912**, *11*, 37–50.
41. Tanimoto, T.T. *An Elementary Mathematical Theory of Classification and Prediction*; International Business Machines Corporation, 1958; *3*,30.
42. DNMT-Targeted Library Available online: https://www.chemdiv.com/catalog/focused-and-targeted-libraries/dnmt-targeted-library/ (accessed on 8 November 2023).
43. Epigenetics Focused Set Available online: https://www.chemdiv.com/catalog/focused-and-targeted-libraries/epigenetics-focused-set/ (accessed on 8 November 2023).
44. Soluble Diversity Library Available online: https://www.chemdiv.com/catalog/diversity-libraries/soluble-diversity-library/ (accessed on 24 February 2023).
45. Davies, M.; Nowotka, M.; Papadatos, G.; Dedman, N.; Gaulton, A.; Atkinson, F.; Bellis, L.; Overington, J.P. ChEMBL Web Services: Streamlining Access to Drug Discovery Data and Utilities. *Nucleic Acids Res.* **2015**, *43*, W612–W620.
46. FooDB Version 1.0 Available online: https://foodb.ca/ (accessed on 1 December 2023).
47. Diversity Screening Libraries Available online: https://lifechemicals.com/screening-libraries/pre-plated-diversity-sets (accessed on 13 March 2023).
48. Epigenetic Screening Libraries Available online: https://lifechemicals.com/screening-libraries/targeted-and-focused-screening-libraries/epigenetic-screening-libraries (accessed on 8 November 2023).
49. Chávez-Hernández, A.L.; Medina-Franco, J.L. Natural Products Subsets: Generation and Characterization. *Artificial Intelligence in the Life Sciences* **2023**, *3*, 100066.
50. Gu, J.; Gui, Y.; Chen, L.; Yuan, G.; Lu, H.-Z.; Xu, X. Use of Natural Products as Chemical Library for Drug Discovery and Network Pharmacology. *PLoS One* **2013**, *8*, e62839.
51. Yuan, Y.; Pei, J.; Lai, L. LigBuilder 2: A Practical de Novo Drug Design Approach. *J. Chem. Inf. Model.*

2011, *51*, 1083–1091.
52. Schneider, G.; Fechner, U. Computer-Based de Novo Design of Drug-like Molecules. *Nat. Rev. Drug Discov.* **2005**, *4*, 649–663.
53. Yuan, Y.; Pei, J.; Lai, L. Binding Site Detection and Druggability Prediction of Protein Targets for Structure-Based Drug Design. *Curr. Pharm. Des.* **2013**, *19*, 2326–2333.
54. Berman, H.M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T.N.; Weissig, H.; Shindyalov, I.N.; Bourne, P.E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235–242.
55. Berthold, M.R.; Cebron, N.; Dill, F.; Gabriel, T.R.; Kötter, T.; Meinl, T.; Ohl, P.; Sieb, C.; Thiel, K.; Wiswedel, B. KNIME: The Konstanz Information Miner. In *Data Analysis, Machine Learning and Applications*; Studies in classification, data analysis, and knowledge organization; Springer Berlin Heidelberg: Berlin, Heidelberg, 2008; pp. 319–326.
56. Xiong, G.; Wu, Z.; Yi, J.; Fu, L.; Yang, Z.; Hsieh, C.; Yin, M.; Zeng, X.; Wu, C.; Lu, A.; et al. ADMETlab 2.0: An Integrated Online Platform for Accurate and Comprehensive Predictions of ADMET Properties. *Nucleic Acids Res.* **2021**, *49*, W5–W14.
57. Sander, T.; Freyss, J.; von Korff, M.; Rufener, C. DataWarrior: An Open-Source Program for Chemistry Aware Data Visualization and Analysis. *J. Chem. Inf. Model.* **2015**, *55*, 460–473.
58. Systèmes, D. BIOVIA Discovery Studio Visualizer Available online: https://discover.3ds.com/discovery-studio-visualizer-download (accessed on 2024).
59. The PyMOL Molecular Graphics System, Version 2.5 Schrödinger, LLC Available online: https://pymol.org/.
60. Fragments Library Available online: https://www.chemdiv.com/catalog/focused-and-targeted-libraries/fragments-library/ (accessed on 8 November 2023).
61. Privileged Fragments Annotated Library Available online: https://www.chemdiv.com/catalog/focused-and-targeted-libraries/privileged-fragments-annotated-library/ (accessed on 9 February 2024).
62. Diversity Screening Subsets of Soluble Fragments Available online: https://lifechemicals.com/fragment-libraries/soluble-fragment-diversity-subsets (accessed on 21 February 2024).
63. Compound Libraries for High Throughput/Content Screening Available online: https://www.selleckchem.com/screening/fragment-library.html (accessed on 25 August 2023).
64. *Molecular Operating Environment (MOE), 2022.02 Chemical Computing Group ULC, 910-1010 Sherbrooke St. W., Montreal, QC H3A 2R7, Canada, 2024*.
65. Van der Maaten, L.; Hinton, G. Visualizing Data Using T-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579-2605.
66. Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742–754.
67. Maier, J.A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K.E.; Simmerling, C. ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J. Chem. Theory Comput.* **2015**, *11*, 3696–3713.
68. Gerber, P.R.; Müller, K. MAB, a Generally Applicable Molecular Force Field for Structure Modelling in Medicinal Chemistry. *J. Comput.-Aided Mol. Des.* **1995**, *9*, 251–268.
69. Jakalian, A.; Jack, D.B.; Bayly, C.I. Fast, Efficient Generation of High-Quality Atomic Charges. AM1-BCC

Model: II. Parameterization and Validation. *J. Comput. Chem.* **2002**, *23*, 1623–1641.
70. O'Boyle, N.M.; Banck, M.; James, C.A.; Morley, C.; Vandermeersch, T.; Hutchison, G.R. Open Babel: An Open Chemical Toolbox. *J. Cheminform.* **2011**, *3*, 33.
71. Kirkpatrick, S.; Gelatt, C.D., Jr; Vecchi, M.P. Optimization by Simulated Annealing. *Science* **1983**, *220*, 671–680.
72. González-Medina, M.; Prieto-Martínez, F.D.; Owen, J.R.; Medina-Franco, J.L. Consensus Diversity Plots: A Global Diversity Analysis of Chemical Libraries. *J. Cheminform.* **2016**, *8*, 63.
73. Willett, P.; Barnard, J.M.; Downs, G.M. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983–996.
74. Bemis, G.W.; Murcko, M.A. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996**, *39*, 2887–2893.
75. Medina-Franco, J.; MartÃ-nez-Mayorga, K.; Bender, A.; Scior, T. Scaffold Diversity Analysis of Compound Data Sets Using an Entropy-Based Measure. *QSAR Comb. Sci.* **2009**, *28*, 1551–1560.
76. Zdrazil, B.; Felix, E.; Hunter, F.; Manners, E.J.; Blackshaw, J.; Corbett, S.; de Veij, M.; Ioannidis, H.; Lopez, D.M.; Mosquera, J.F.; et al. The ChEMBL Database in 2023: A Drug Discovery Platform Spanning Multiple Bioactivity Data Types and Time Periods. *Nucleic Acids Res.* **2024**, *52*, D1180–D1192.
77. Durant, J.L.; Leland, B.A.; Henry, D.R.; Nourse, J.G. Reoptimization of MDL Keys for Use in Drug Discovery. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1273–1280.
78. Altman, N.S. An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression. *Am. Stat.* **1992**, *46*, 175.
79. Ho, T.K. The Random Subspace Method for Constructing Decision Forests. *IEEE Trans. Pattern Anal. Mach. Intell.* **1998**, *20*, 832–844.
80. Friedman, J.H. Greedy Function Approximation: A Gradient Boosting Machine. *Ann. Statist.* **2001**, *29*, 1189–1232.
81. Cortes, C.; Vapnik, V. Support-Vector Networks. *Mach. Learn.* **1995**, *20*, 273–297.
82. Hopfield, J.J. Neural Networks and Physical Systems with Emergent Collective Computational Abilities. *Proc. Natl. Acad. Sci. U. S. A.* **1982**, *79*, 2554–2558.
83. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Müller, A.; Nothman, J.; Louppe, G.; et al. Scikit-Learn: Machine Learning in Python. J. *Mach. Learn. Res.* **2011***, 12*, 2825-2830.
84. Tetko, I.V.; Sushko, I.; Pandey, A.K.; Zhu, H.; Tropsha, A.; Papa, E.; Oberg, T.; Todeschini, R.; Fourches, D.; Varnek, A. Critical Assessment of QSAR Models of Environmental Toxicity against Tetrahymena Pyriformis: Focusing on Applicability Domain and Overfitting by Variable Selection. *J. Chem. Inf. Model.* **2008**, *48*, 1733–1746.
85. Sushko, I.; Novotarskyi, S.; Körner, R.; Pandey, A.K.; Cherkasov, A.; Li, J.; Gramatica, P.; Hansen, K.; Schroeter, T.; Müller, K.-R.; et al. Applicability Domains for Classification Problems: Benchmarking of Distance to Models for Ames Mutagenicity Set. *J. Chem. Inf. Model.* **2010**, *50*, 2094–2111.
86. Capecchi, A.; Probst, D.; Reymond, J.-L. One Molecular Fingerprint to Rule Them All: Drugs,

Biomolecules, and the Metabolome. *J. Cheminform.* **2020**, *12*, 43.

87. Datta, J.; Ghoshal, K.; Denny, W.A.; Gamage, S.A.; Brooke, D.G.; Phiasivongsa, P.; Redkar, S.; Jacob, S.T. A New Class of Quinoline-Based DNA Hypomethylating Agents Reactivates Tumor Suppressor Genes by Blocking DNA Methyltransferase 1 Activity and Inducing Its Degradation. *Cancer Res.* **2009**, *69*, 4277–4285.
88. Gros, C.; Fleury, L.; Nahoum, V.; Faux, C.; Valente, S.; Labella, D.; Cantagrel, F.; Rilova, E.; Bouhlel, M.A.; David-Cordonnier, M.-H.; et al. New Insights on the Mechanism of Quinoline-Based DNA Methyltransferase Inhibitors. *J. Biol. Chem.* **2015**, *290*, 6293–6302.
89. Gamage, S.A.; Brooke, D.G.; Redkar, S.; Datta, J.; Jacob, S.T.; Denny, W.A. Structure-Activity Relationships for 4-Anilinoquinoline Derivatives as Inhibitors of the DNA Methyltransferase Enzyme DNMT1. *Bioorg. Med. Chem.* **2013**, *21*, 3147–3153.
90. 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.; et al. Detailed Exploration around 4-Aminoquinolines Chemical Space to Navigate the Lysine Methyltransferase G9a and DNA Methyltransferase Biological Spaces. *J. Med. Chem.* **2018**, *61*, 6546–6573.
91. López-López, E.; Prieto-Martínez, F.D.; 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*, 3282.
92. Rabal, O.; San José-Enériz, E.; Agirre, X.; Sánchez-Arias, J.A.; Vilas-Zornoza, A.; Ugarte, A.; de Miguel, I.; Miranda, E.; Garate, L.; Fraga, M.; et al. Discovery of Reversible DNA Methyltransferase and Lysine Methyltransferase G9a Inhibitors with Antitumoral in Vivo Efficacy. *J. Med. Chem.* **2018**, *61*, 6518–6545.
93. Miljković, F.; Medina-Franco, J.L. Artificial Intelligence-Open Science Symbiosis in Chemoinformatics. *Artif. Intell. Life Sci.* **2024**, *5*, 100096.