Artificial Intelligence in the Life Sciences 1 (2021) 100008

Contents lists available at ScienceDirect

Artificial Intelligence in the Life Sciences

journal homepage: www.elsevier.com/locate/ailsci

Methods & Protocols

An *in silico* pipeline for the discovery of multitarget ligands: A case study for epi-polypharmacology based on DNMT1/HDAC2 inhibition

Fernando D. Prieto-Martínez[a], Eli Fernández-de Gortari[b,*], José L. Medina-Franco[c], L. Michel Espinoza-Fonseca[d,*]

[a] *Instituto de Química, Universidad Autónoma de México, 04510 Mexico City, Mexico*
[b] *Department of Nanosafety, International Iberian Nanotechnology Laboratory, Braga, Portugal*
[c] *DIFACQUIM Research Group, Department of Pharmacy, School of Chemistry, National Autonomous University of Mexico, 04510 Mexico City, Mexico*
[d] *Center for Arrhythmia Research, Department of Internal Medicine, Division of Cardiovascular Medicine, University of Michigan, Ann Arbor, MI 48109, United States*

ARTICLE INFO

Keywords:
Molecular deep generator network
DNA-methyltransferase
Dual inhibitors
Epigenetics
Histone deacetylase
Machine-learning
Molecular optimization
Open science
Polypharmacology

ABSTRACT

The search for novel therapeutic compounds remains an overwhelming task owing to the time-consuming and expensive nature of the drug development process and low success rates. Traditional methodologies that rely on the one drug-one target paradigm have proven insufficient for the treatment of multifactorial diseases, leading to a shift to multitarget approaches. In this emerging paradigm, molecules with off-target and promiscuous interactions may result in preferred therapies. In this study, we developed a general pipeline combining machine learning algorithms and a deep generator network to train a dual inhibitor classifier capable of identifying putative pharmacophoric traits. As a case study, we focused on dual inhibitors targeting DNA methyltransferase 1 (DNMT) and histone deacetylase 2 (HDAC2), two enzymes that play a central role in epigenetic regulation. We used this approach to identify dual inhibitors from a novel large natural product database in the public domain. We used docking and atomistic simulations as complementary approaches to establish the ligand-interaction profiles between the best hits and DNMT1/HDAC2. By using the combined ligand- and structure-based approaches, we discovered two promising novel scaffolds that can be used to simultaneously target both DNMT1 and HDAC2. We conclude that the flexibility and adaptability of the proposed pipeline has predictive capabilities of similar or derivative methods and is readily applicable to the discovery of small molecules targeting many other therapeutically relevant proteins.

Introduction

The epigenetic landscape encompasses a series of protein families associated with gene modulation. These mechanisms involve chemical modifications around histone side chains, thus changing its stability or favoring protein-protein interactions [1]. Currently, research efforts have turned to epigenetics as the missing link in chronic diseases such as cancer or Alzheimer's [2]. Histone deacetylases (HDAC) are a prominent example of clinically relevant epi-targets. The human genome includes 18 HDAC isoforms, classified in four classes based on their sequence homology to yeast protein [3]. To date, HDAC inhibitors constitute a novel alternative for the treatment of neoplasia. Other studies have shown increased therapeutic applications of HDAC inhibitors, including those as anti-inflammatory [4] and antiviral agents [5]. A significant concern for such therapies is drug safety because HDACs preserve a catalytic Zn^{2+} core and potent inhibitors often result in pan-HDAC inhibition and pleiotropic effects with no clear mechanism of action [6]. Recent studies have proposed the collective inhibition of other epi-targets, such as readers (e.g., bromodomains) or other writers (e.g., DNA-5-methyltransferases), to improve their biological activity and clinical efficacy [7]. Thus, research on epigenetic polypharmacology or epi-polypharmacology is on the rise [8–10].

DNA methylation is the addition of methyl groups to cytosine to produce 5-methyl cytosine [11]. This reaction is catalyzed by DNA-methyltransferases (DNMTs), which include DNMT1, which has a maintenance role as it preserves the methylation pattern, and DNMT3A/B, which is responsible for *de novo* DNA methylation [12]. Experimental evidence suggests that mitotic inheritance of this epi-mark has complex implications in aging and tumorigenic processes [13]. The pharmaceutical potential of DNMT inhibition is irrefutable, yet most clinically approved inhibitors include nucleosides, which have long-term safety concerns such as mitochondrial toxicity [14,15]. Therefore, it is urgently needed to search for novel non-nucleoside scaffolds or privileged structures as DNMT inhibitors [16].

* To whom correspondence should be addressed.
E-mail addresses: eli.fernandez@inl.int (E. Fernández-de Gortari), medinajl@unam.mx (J.L. Medina-Franco), lmef@umich.edu (L. Michel Espinoza-Fonseca).

https://doi.org/10.1016/j.ailsci.2021.100008
Received 24 August 2021; Received in revised form 8 September 2021; Accepted 9 September 2021
Available online 12 September 2021
2667-3185/© 2021 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)

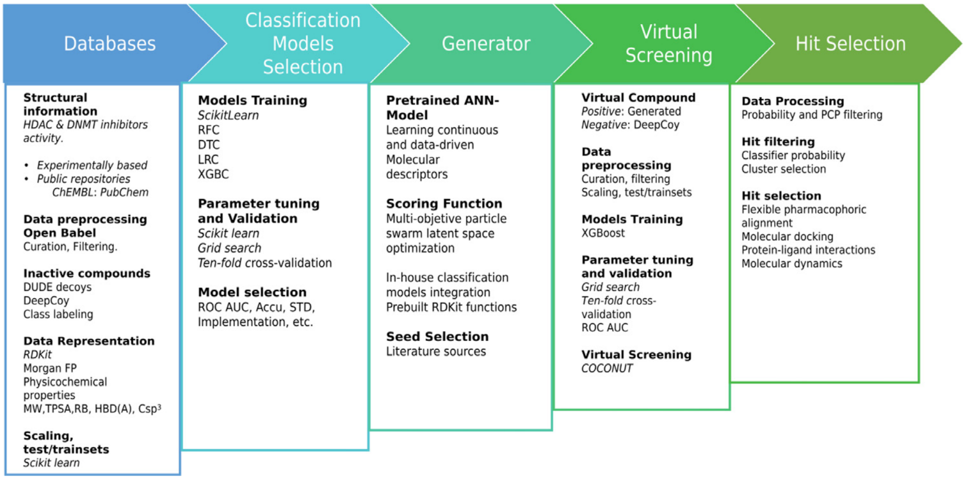

Fig. 1. The CADD pipeline proposed in this work integrates ligand-based and structure-based methods for virtual screening of DNMT1/HDAC2 dual inhibitors.

Computer-aided drug design (CADD) has revolutionized drug discovery and development. Although CADD still has several areas of improvement [17], it has contributed directly to the development of 70 drugs in clinical use [18]. CADD uses two main approaches: Structure-based drug design (SBDD) and ligand-based design drug design (LBDD). SBDD is a common approach that primarily relies on molecular modeling and simulation, while LBDD focuses on small molecule featurization and characterization of high-dimensional data. Computational methods have shown notable success for single-target molecules, but a major challenge in multitarget drug design is the selection of target combinations for the desired therapeutic effect [19]. In the SBDD approach, the pocket similarity between proteins is expected for better results and hit enrichment [20]. From the LBDD perspective, the hunting ground for multitarget molecules is the putative intercept of small molecule's chemical spaces. Recently, machine-learning approaches combined with fragment or scaffold searches have yielded promising results [21].

Despite these advances, a predictive model capable of informing the structure-activity relationships (SAR) of multitarget drugs (i.e., structure-multiple activity relationships) is highly dependent on the diversity, size, and quality of the available data. Over the last few decades, many computational tools derived from machine learning and natural language processing are now able to transform molecular structural information, from small amounts of experimental data, into a helpful representation suited for the generation of novel compounds generated *in silico*. Applying these new algorithms into the drug discovery pipeline represents a big step in reducing resource expenditure, thus providing a wide range of growth opportunities for non-traditional players in the market, such as academia and developing countries [22–24].

The development of multitarget pharmacological therapies, capable of regulating different biological targets related to a specific disease [25,26], represents a challenging pharmacological problem because of the high complexity of the protein signaling networks involved in the physiopathology of the disease [27], prompting for efficient tools for designing master key compounds. Examples of privileged scaffolds have been described in the literature, including benzodiazepines as sedative agents [28], CCK antagonists [29], and bromodomain inhibitors [30]. These structures may serve as starting points for lead development [31] or chemical library design [32], thus pressing the need for further study and characterization [33]. To overcome these challenges, in this study, we introduce a CADD pipeline that integrates SDBB and LBDD methods for the virtual screening of dual inhibitors (Fig. 1). As a case study, we applied this virtual screening protocol to identify DNMT1/HDAC2 dual inhibitors. The pipeline is based on open-access data and tools, using records from public molecular repositories and medicinal chemistry groups. This methodology takes advantage of conventional machine learning classifiers, a multi-objective particle swarm optimization algorithm, and a deep generator network latent space to train a dual inhibitor classifier for virtual screening [34,35]. We successfully generated new molecular hypotheses derived from large natural products public repositories that meet a required profile.

Table 1
Datasets used for classification training and compound generation.

Dataset	No. of compounds	Source
DNMT1	405	ChEMBL [42]
HDAC2	1490	PubChem [43]
Decoys DNMT1	411	DeepCoy [44]
Decoys HDAC2	1490	DeepCoy [44]

Methodology

Dataset acquisition and preprocessing

We trained two classifiers as future parameters of a swarm optimizer objective function to guide the optimization algorithm to an enriched segment of the chemical space. We obtained molecular bioactivity public information of DNMT1 inhibitors using a curated dataset previously published [36]. The dataset was enriched with activity data available in ChEMBL25 [37] selecting compounds with IC_{50} values lower than 100 μM. Also, we retrieved HDAC2 inhibitors from a dataset reported by our group [38]. We filtered and curated the effectors using the parameters reported by Fourches et al. [39]. We used Open Babel [40] and DataWarrior (v. 5.2.1) [41] to perform the canonical linear notation line-entry system cSMILES, protonation states, molecular properties values (between the Rule of Five and the Rule of Three range), as well as metal salts, small fragments, and removal of duplicated entries. The total number of compounds considered in this study is summarized in Table 1.

The negative class of HDAC2 (non-inhibitors) was obtained by calculating a diverse subset with RDKit (v.2018.09.3) and Mayachemtools [45] of the HDAC decoys in the Oxford Protein Informatics Group site [46]. For DNMT1 inhibitors, counterparts were generated on-demand with DeepCoy [44] using the active compounds dataset as reference. The compounds of interest were generated to match DUD-E descriptors in a 150:1 decoy-to-reference ratio following selection of top decoys using the provided DeepCoy scripts. Finally, the following descriptors for both sets were calculated in DataWarrior: molecular weight (MW), topological polar surface area (TPSA), cLogP, number of rotatable bonds, the fraction of sp^3 carbon atoms, number of H-bond acceptors, and donors.

We conducted principal component analysis (PCA) of the seven autoscaled properties for analysis and visualization.

Independent DNMT1 and HDAC2 classifiers

Data preprocessing. Each dataset point was represented with the help of the RDKit Python module [34] as a concatenated vector of Morgan Binary Fingerprints (n = 2) [47] and seven previously described molecular descriptors. Given the numerical features differences, each representation was standardized, subtracting the mean and scaling to unit variance using the Scikit-learn module [48] in Python.

Model training and validation. We compared the performance of four different classification algorithms for each dataset (i.e., DNMT1 and HDAC2): logistic regression (LRC), decision trees (DTC), random forest (RFC), and extreme gradient boosting classifiers (xgboost classifier; XGBC). The comparison was performed with Scikit-learn and xgboost [49] Python modules. Each model was trained and validated by ten-fold cross-validation and grid search optimization to tune their respective hyperparameters. Finally, we selected the logistic regression model for both classifiers, given its compatibility with the swarm optimizer architecture and performance comparing with the best performance model, xgboost.

Compound generation

The small number of dual inhibitors publicly available makes it almost impossible to apply traditional machine learning methodologies to determine the general molecular pattern responsible for dual DNMT1/HDAC2 inhibition. To overcome this limitation, we applied a multi-objective molecular particle swarm optimizer coupled to a continuous latent space obtained from a deep generator network pre-trained by the translation process from SMILES to canonical SMILES of 75 million molecules [34]. This algorithm uses RDKit functions like drug-likeness [50], synthetic accessibility [51], and partition coefficient to construct an objective function for guiding the compound generation.

The objective function allows the inclusion of in-house models as part of their parameters, enabling the pre-trained classifiers inclusion to constrain the chemical space into a region enriched by hypothetical multitarget compounds related to dual inhibition. In this work, the scoring function includes approximations of general desirable parameters for drug design, including drug-likeness, synthetic accessibility, heavy atoms count, substructure matching; also including DNMT1 inhibitors and HDAC2 inhibitors pre-trained classifiers. These parameters were heuristically weighted as follows: Objective Function = 22.5 (DNMT inhibitor classifier) + 22.5 (HDAC inhibitor classifier) + 15 (Drug-likeness function) + 40 (Synthetic Accessibility function + Macrocycles penalizing function + Heavy Atoms Count function + ChEMBL_Structure + Substructure_Match). We chose these specific weights to balance the structural features of the generated molecules, considering pharmacophoric elements previously suggested for DNMT1 [52] and HDAC2 [53] during the curation and filtering process. The substructure matching query was N-(*p*-methylphenethyl)-benzamide. This compound was selected based on selective HDAC2 inhibition, which may be related to other mechanisms beyond zinc interactions, such as hydrogen bonding networks [54].

To further restrict the chemical space of generated compounds, we selected four structurally diverse queries as the optimization starting point (molecular seeds, Chart 1): gliburide (DNMT inhibitor), panobinostat (HDAC inhibitor and low DNMT inhibitor; [16]), **15a** (nanomolar dual inhibitor) [55], and SAHA (HDAC inhibitor, also the structural basis for both **15a** and panobinostat).

Generation of putative dual inhibitors

We employed the previously described scoring function to generate between 1000 and 2500 molecules starting from the selected seeds (within an average of 30 optimization steps and 45 runs). The resulting structures were subjected to a semi-automated selection process to alleviate the tendency of these molecular generators to produce uncommon substructures, intricate chemical patterns, duplicate molecules, small diversity sets, or structurally incorrect compounds [56]. Given the pharmacophoric traits of the generated molecules, our hypothesis holds that many of these molecules contain enough structural information to elucidate the general pattern necessary to produce a biological response in one or two of the selected biological targets. To further increment the molecular diversity of putative dual inhibitors, we added ten decoys for each generated structure using DeepCoy generator. Finally, we randomly selected one of the ten decoys obtained per structure, resulting in a balanced proportion of approximately 1:1 positive and negative examples. The final dataset was preprocessed and represented as described for the independent DNMT1 and HDAC2 datasets.

Dual inhibitors classifier

Xgboost [49] was developed to control over-fitting, improve performance, and increase computing speed compared with other traditional non-ensemble algorithms. Therefore, we selected this classification method, applied grid search for hyperparameter tuning followed by ten-fold cross-validation. Training of the final dual inhibitors classifier was done using the dataset of generated putative dual inhibitors, described in the last section.

Virtual screening

Natural products continue to have paramount importance in drug discovery [57]. Often curated by a trial-and-error process searching for therapeutic agents, it is not surprising that almost one-third of FDA-approved new chemical entities were derived from natural sources [58]. For this reason, we selected a large natural product collection available in the public domain as the chemical space for virtual screening: the natural products database COCONUT with more than 406,076 unique flat natural products collected from 53 data sources [59]. Of note, the virtual screening protocol proposed in this work can be used with any small-molecule database.

Using the model trained with the generated molecules, we selected the compounds with class probability values between 0.5 and 0.9. We also removed highly halogenated molecules (i.e., those with three or more halogen substitutions) and structures with MW higher than 650 Da; this value was selected as a cutoff based on the MW distribution (see Figures S1 and S2 in the Supporting Information). A dimensionality reduction and visualization for the resulting molecules were applied using T-distributed stochastic neighbor embedding (t-SNE) [60] on molecules PathFp descriptors from DataWarrior (binary hashed vector that encodes up to seven atoms linear strands). The resulting color-coded classifier probability 2D space was used for selecting hits, with each node representing a low dimensionality molecular embedding and each edge representing the similarity between adjacent nodes. In this way, a pair of connected compounds with a substantial difference in probability could be considered an activity cliff [61], where slight probability differences may be interpreted as a continuous SAR.

Molecular modeling of natural product hits

After hit selection and inspection, compounds were submitted to a series of analyses to better characterize their putative interaction as dual DNMT1/HDAC2 inhibitors. The first step involves a flexible pharmacophoric alignment between hits and compounds **12a** (Figure S8 in the Supporting Information) and **15a** (Fig. 2) synthesized by Yuan et al. [62]. This was done with pharmACOphore, as this utility bases its scoring of alignment on the ant colony optimization metaheuristic [63]. The hits selected from the previous analysis were docked to DNMT1 (PDB ID: 3SWR) and HDAC2 (PDB ID: 6WBW) to test their putative interaction

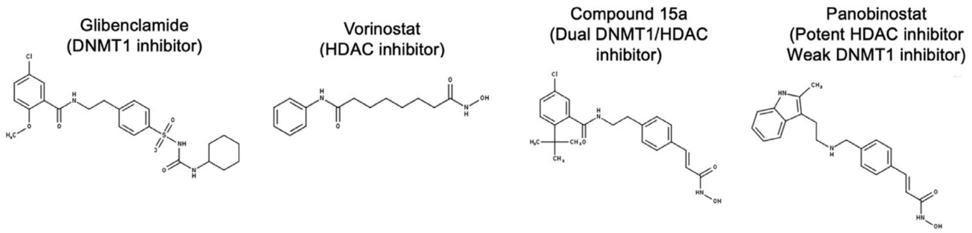

Chart 1. Chemical structures of the seed compounds used for molecular swarm optimization.

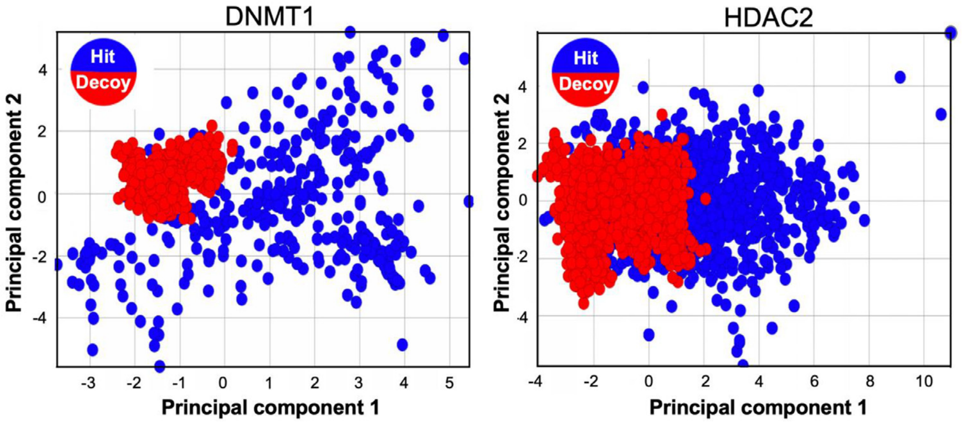

Fig. 2. Two-dimensional molecular representation of the properties space by principal component analysis: DNMT1 (left) and HDAC2 (right). The positive hits obtained from the curated datasets are shown as blue circles, whereas the inactive hits (decoys) are shown as red circles. Principal components covariance recovery was about 51% (PC1) and 30% (PC2) for DNMT1, and 53% (PC1) and 21% (PC2) for HDAC2. For additional details, see Table S1 in the Supporting Information.

Table 2
Accuracy and confusion matrices for DNMT1 classifier.

DNMT1	LRC	DTC	RFC	XGBC
True Negatives	98	96	95	99
False Negatives	0	0	0	0
False Positives	1	3	4	0
True Positives	106	106	106	106
Mean Accuracy	98.87	97.72%	99.03%	98.54%
Mean SD	1.21%	1.49%	1.07%	1.35%
Best Accuracy	99.03%	99.19%	100%	98.54%

Best parameters: XGBC: booster = dart, DTC: criterion = entropy, max_depth = 7, splitter = random, LRC: C = 100, class_ weight = balanced, penalty = l2, max_iter: 400, solver: lbfgs, RFC: criterion = gini, max_features = auto, min_samples_leaf = 1, min_samples_split = 2, n_estimators = 500.

Table 3
Accuracy and confusion matrices for the HDAC2 classifier.

HDAC2	LRC	DTC	RFC	XGBC
True Negatives	343	389	392	389
False Negatives	2	11	2	0
False Positives	7	13	10	13
True Positives	341	332	341	343
Mean Accuracy	98.52%	95.39%	98.17%	97.72%
Mean SD	0.85%	0.70%	0.91%	0. 90%
Best Accuracy	98.97%	96.69%	96.69%	97.85%

with the epi-enzymes. We performed molecular docking with PLANTS (v.1.2) [64]. The binding site was defined within a 10 Å sphere around the co-crystalized ligand of each protein; 25 orientations for each hit were further inspected and selected based on their 3D similarity and interactions using the orientation found in the crystal structures as reference. The best protein-ligand models were relaxed for 200-ns using all-atom molecular dynamics simulations (MD) in a solution containing 150 mM NaCl using AMBER on Tesla V100 GPUs [65]. In all cases, we used the AMBER ff19SB [66] and the General AMBER [67] force fields to model protein, water, ions, and small molecules. The trajectories were analyzed using MDTraj (v.1.9.4) [68] and Contact Map (v.0.7.0) [69].

Results and discussion

DNMT1 and HDAC 2 datasets

Visual representation of chemical space with PCA (for covariance values see Table S1 in Supporting Information) showed significant differences between DNMT1 actives and decoys (Fig. 2). This was evidenced from property distribution among the calculated descriptors (Figures S1-S3 in the Supporting Information), particularly those related to solubility and polarity. In contrast, for HDAC2, a significant overlap can be observed, as decoys cover similar regions on the visualized space (Fig. 2). The distribution of compounds in the chemical space is the result of DeepCoy using more than 25 descriptors and a high number of iterations (around 1000–2000 decoys per active molecule) to construct decoys. [44]. We also postulate that better coverage for DNMT1 decoys may be observed in this chemical space. To balance the computational cost and accuracy of the model, we selected 150 decoys per active molecule that have shown better performance than the original DUD-E [44].

Classification and model selection

We chose logistic regression as the best performance model by grid search model comparison. The results of cross-validation, confusion matrix, and overall mean accuracy for DNMT1 and HDAC2 inhibitors are reported in Tables 2 and 3, respectively. Also, we calculated the best accuracy values in all cases, representing the overall accuracy obtained after fine-tuning the hyperparameters (Table 2).

Best parameters: XGBC: booster = gblinear, DTC: criterion = entropy, max_depth = 14, splitter = random, LRC: C = 10, class_ weight = balanced, penalty = l2, max_iter: 400, solver: lbfgs, RFC: criterion = gini, max_features = auto, min_samples_leaf = 1, min_samples_split = 2, n_estimators = 1000.

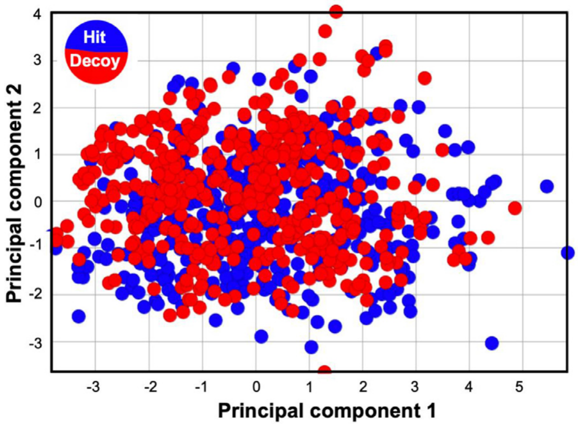

Fig. 3. Two-dimensional visual representation of the properties space for generated dual inhibitors by principal component analysis. The positive hits obtained from the curated datasets are shown as blue circles, whereas the inactive hits (decoys) are shown as red circles. The visual representation was obtained by principal component analysis. The covariance recovery for PC1 and PC2 is about 49% and 27%, respectively. For additional details, see Table S1 in the Supporting Information.

It has been reported that RFC often outperforms LRC [70,71]. In this study, we found that LRC produces similar results to those by RFC with modest computational resources (e.g., a typical laptop or desktop computer with four CPU cores). This feature is key for the performance of the algorithm, as virtual screening campaigns are often conducted on databases containing 10^5–10^7 molecules [72]. While it can be argued that the improved performance of LRC methods is achieved at the expense of accuracy, a recent study showed that there are no significant differences between LRC and other algorithms in predicting outcomes of similar complexity (i.e., high dimensionality) [73]. In addition, XGBC yielded similar values to LRC, and hyperparameter selection showed that XGBC used linear booster for best results (Tables 2-3). Nevertheless, the LRC model simplifies its implementation given module dependencies present in further steps along the pipeline.

Table 4
Accuracy and confusion matrices for DNMT1/HDAC2, dual classifier.

VS_Model_One run	XGBC_VS
True Negatives	108
False Negatives	6
False Positives	9
True Positives	115
Mean Accuracy	93.13%
Mean SD	2.71%
Best Accuracy	93.13%

Property space of generated dual inhibitors

We found that the decoy coverage improved significantly for the generated dual compounds (Fig. 3), which may be attributed to a higher control over property space delimitation, as molecular seeds share common traits and features. The impact this had on results is clear; for instance, classes are not easily separable, therefore giving further support to the hypothesis proposed above. Chemical space coverage of generated compounds shows the overall preservation of the calculated properties among decoys.

Model training and validation for dual inhibitor classification

With the generated compounds and decoys, an independent classifier was trained. We train an XGBC for the virtual screening task of using the Python Xgboost library. Model validation and optimization followed the same steps as implemented for DNMT1 and HDAC2 classifiers. Results are summarized in Table 4.

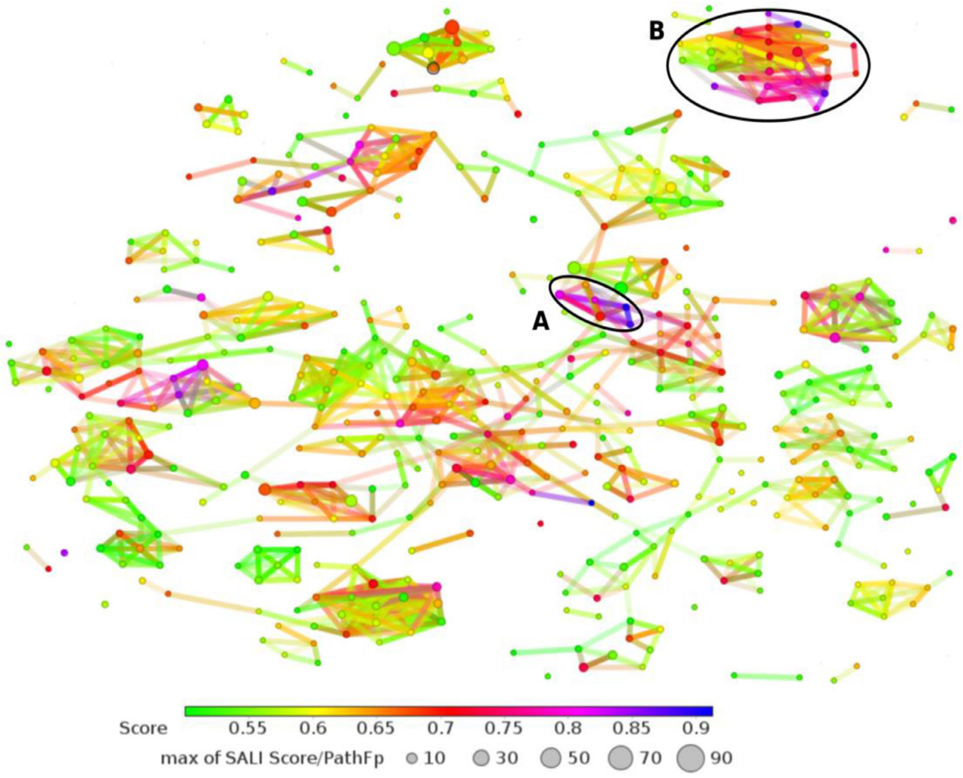

Fig. 4. A t-SNE dimensionality reduction and network representation of the screened molecules represented by the PathFp descriptors included in DataWarrior. In the network, compounds are represented as nodes linked by their structural similarity. Each node is color-coded by the output probability calculated by the dual inhibitor's classifier.

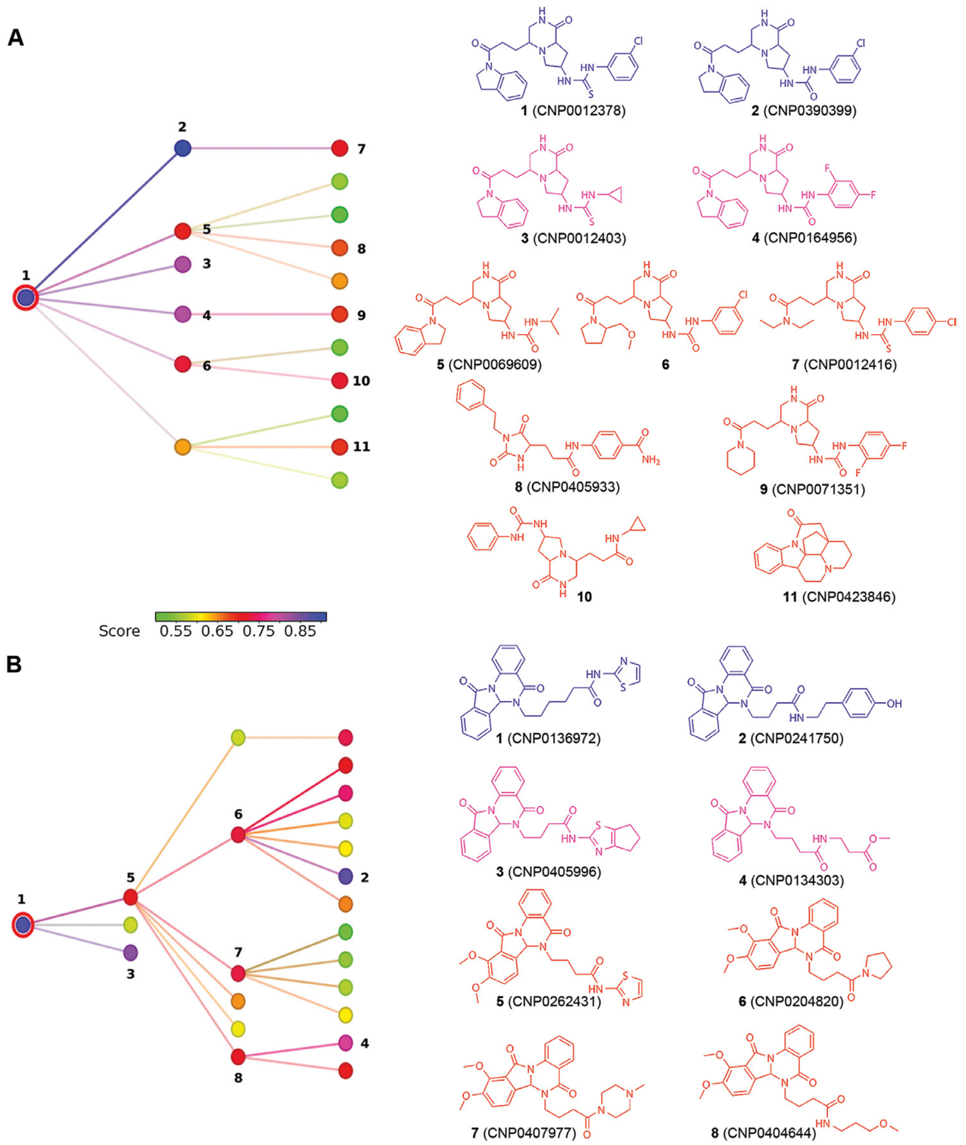

Fig. 5. Node neighborhood of the two clusters with higher probability members. (A) and (B) recapitulate the molecular clusters and chemical structures identified in Fig. 4.

Best parameters: booster = dart

We found that hyperparameter optimization yielded better results based on dropouts for the adaptive regression trees (DART) method. This ensemble algorithm was conceived to prevent over-specialization often found in gradient boosted trees (GBT) [74]. We also selected this particular model because of its use in GBT, and because this algorithm has shown notable performance for protein-ligand interaction modeling [71]. Validation procedures are data-dependent, as class imbalances can lead to artifacts on classification or prediction. In this sense, ROC curves are used to assess the true predictive power of a model by using the area under the curve as the primary value for comparison. Herein, we included ten-fold cross-validation and ROC curves. Based on this the independent classifiers (DNMT1 and HDAC2) obtained with LRC showed good consistency and performance.

In addition to ROC curves, we calculated the Matthews correlation coefficient (MCC) to test the predictive power of developed classifiers (see Table S2 in the Supporting Information). This coefficient was initially developed to evaluate classification performance, as it is obtained directly from the confusion matrix. The possible values for MCC range from −1 to 1, with zero being equivalent to random chance [75]. MCC offers some advantages over traditional metrics such as accuracy and/or F1 score [76,77]. Furthermore, MCC is robust enough to evaluate the predictive power of models even on “adverse” situations such as class imbalance [78].All of the models used herein obtained a good correlation with values above 0.9 and 0.85 for independent and dual classifiers respectively.

Another common problem during machine learning campaigns is overfitting. Indeed, algorithms are prone to poor generalization if caution is not exercised during model training. To overcome this issue, best practices include the use of a validation set and early stopping [79]. Additional measures include hyperparameter tuning; for logistic regression penalties, such as L1 and L2, are commonly used to avoid overfitting. For XGBoost slow learning rates and early stopping are strongly recommended as best practices. For the case study presented hereunder, overfitting may also arise from data. Given the trend to follow specific pharmacophoric traits to design DNMT (nucleosides or charged amines) and HDAC (hydroxamates) inhibitors, we used a consensus approach for hit pruning using SBDD.

Virtual screening of natural products

The broad chemical space covered by the natural products contained in the COCONUT database [80] offers an excellent case study to identify

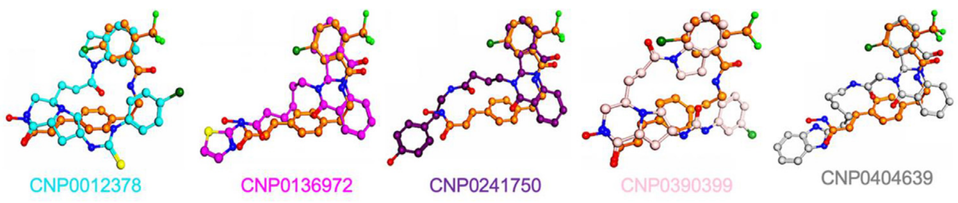

Fig. 6. Pharmacophoric alignment of virtual screening hits and reference compounds. For simplicity, only pairs of compounds are shown. **CNP0012378** (Carbon atoms shown in cyan); **CNP0136972** (Carbon atoms in magenta); **CNP0241750** (Carbon atoms shown in purple); **CNP0390399** (Carbon atoms shown in pink), and **CNP0404639** (Carbon atoms shown in gray). In all cases, we used compound **15a** (Carbon atoms shown in orange) as a reference.

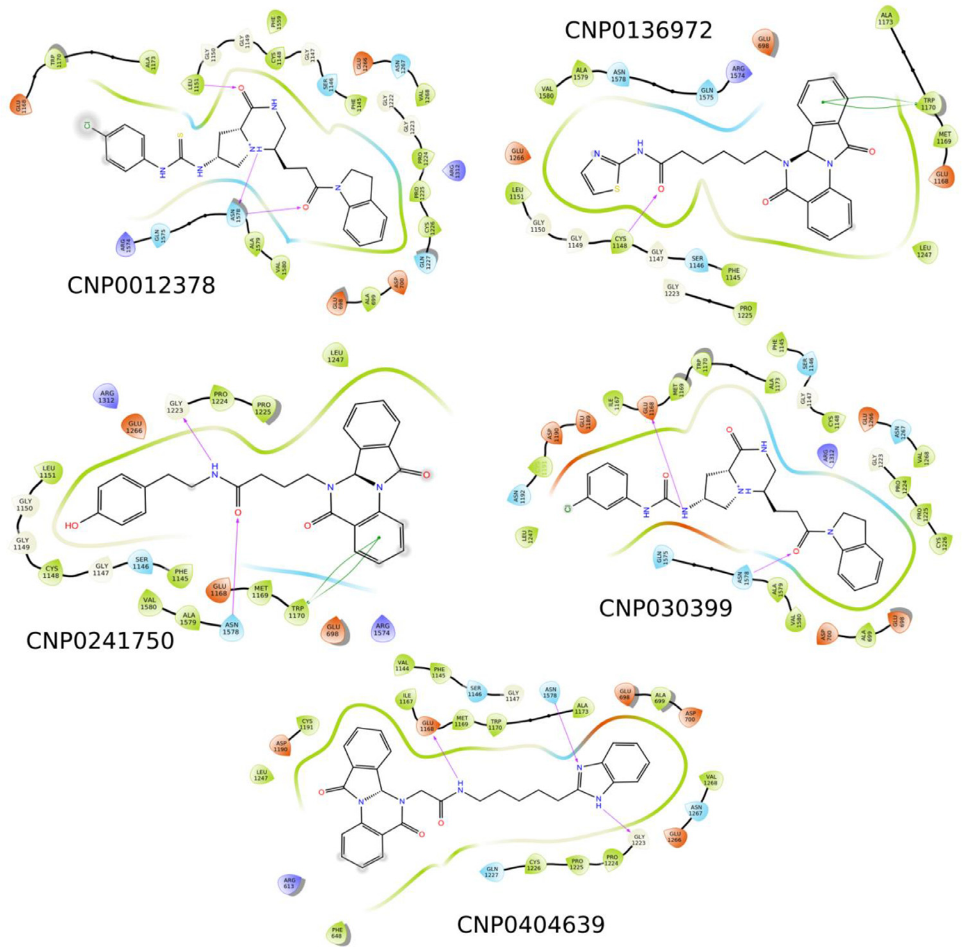

Fig. 7. Two-dimensional interaction diagrams for the docked complexes between DNMT1 and the dual ligands obtained in this study.

potential dual inhibitors. A total of 1239 compounds (0.3% of the entire database) were recovered by our protocol (Fig. 4). Based on scoring values and the t-SNE connectivity, we identified two main clusters (A and B) to select hit compounds. In Fig. 4, the color scale facilitates identifying compounds with a higher probability of the desired trait for dual inhibition. Connectivity, on the other hand, guides the identification of high structural similarities among compounds. We note, however, that caution must be exercised when interpreting the connectivity plots as a cluster with heterogeneous coloring may be indicative of activity cliffs [81].

Cluster A (6 compounds, Fig. 4) showed consistent probability values, mainly due to structure conservation. The source of diversity in this cluster comes from the fragments found in the main chain of the scaffold, 1-oxo-hexahydro-2H-pyrrolo[1,2-a]pyrazin-7-yl-urea (Fig. 5A). Therefore, this scaffold has an apparent continuous SAR, with the main difference being the aromatic substituent adjacent to the urea moiety. However, on closer inspection, there are notable changes on neighboring scaffolds when the classifier's probability is lower than 0.7. Conversely, cluster B (32 compounds, Fig. 4) showed a more heterogeneous distribution of dual activity scorings, i.e., the network representation suggests abrupt changes in dual activity. We found that most of the best-ranked compounds share a common scaffold: 2-{5,11-dioxo-6aH-isoindolo[2,1-a]quinazolin-6-y]}acetamide (Fig. 5B). In this cluster, the primary differences among neighboring compounds are the length of the substituent chain and the size of the capping group at the end of the chain. The structures of the scaffolds emerging from the virtual screening are shown in Chart 2.

Molecular modeling of virtual screening hits

Alignment

We further explored the pharmacophoric similarity using the hits discovered in the COCONUT database. We selected five hits based on the classification score used in this study; specifically, we selected two molecules from cluster A (CNP0136972 and CNP0241750, Fig. 5A), which represent the scaffolds with a continuous SAR, and three molecules from cluster B (CNP0012378, CNP0390399, and CNP0404639, Fig. 5B), which represent the scaffolds that have a more

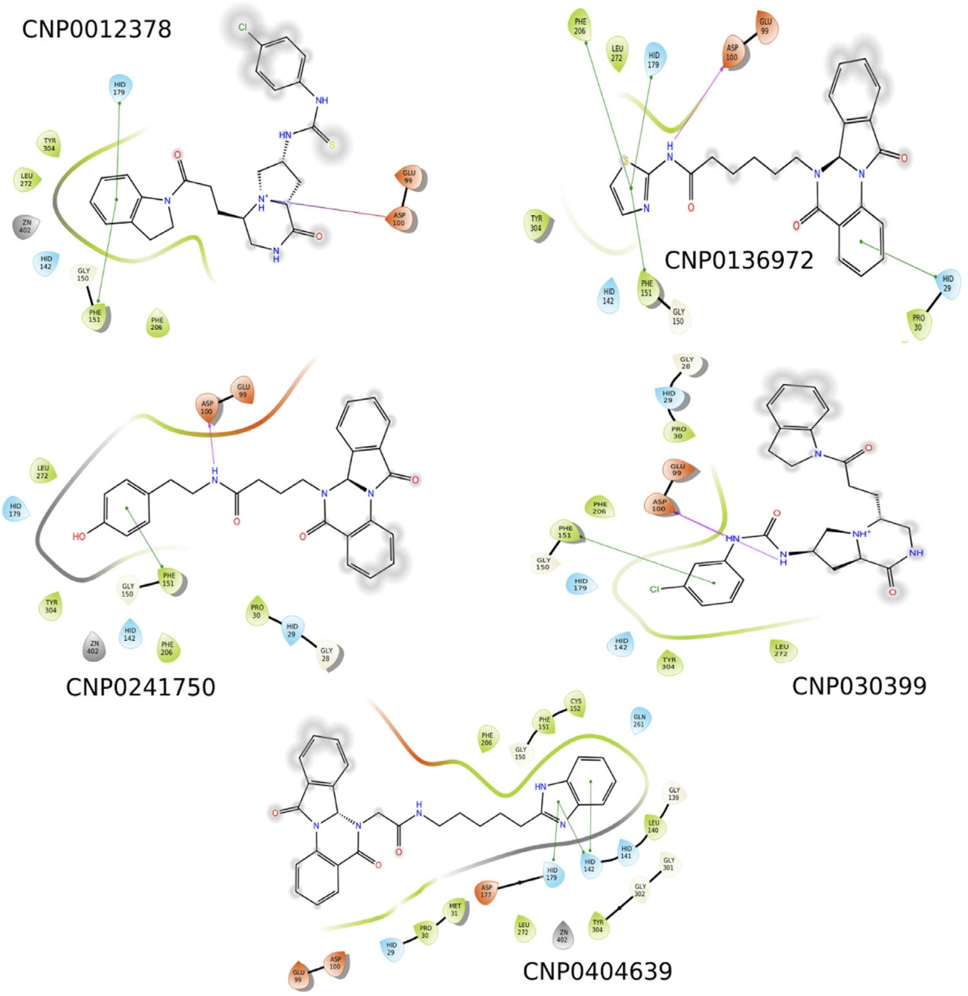

Fig. 8. Two-dimensional interaction diagrams for the docked complexes between HDAC2 and the dual ligands obtained in this study.

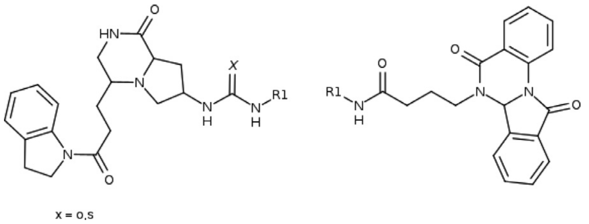

Chart 2. Murcko scaffolds present in the virtual screening hits that were discovered in the COCONUT database.

heterogeneous SAR. Our results suggest that the computational hits share pharmacophoric similarity with the molecular seeds used for compounds generation. Therefore, we conducted flexible alignments of pharmacophoric elements. This approach is based on the notion that aligned compounds share a similar structural arrangement when bound to their target, thus facilitating the analysis of conserved interaction patterns using spatial criteria [82]. Our analysis showed that the best hits give good matches to reference dual compounds (Fig. 6). Comparison of the hits with compound **15a** revealed that the common spatial distribution of interacting elements is shared among hits discovered with the virtual screening approach (Fig. 6). It is important to note that while alignment is an LBDD approach that cannot provide direct insight on the putative binding mode, it supports the hypothesis that a common pattern can lead to a pharmacophore hypothesis for dual DNMT1/HDAC2 activity.

Molecular docking of the hits onto DNMT1 and HDAC2

For context, we summarize known interaction patterns and significant residues for each target. Further reference can be found in Figures S9-S10 in the Supporting Information, which illustrates the interaction of DNMT1 and HDAC2 with the inhibitors used as a control in this study. For DNMT1, the main protein-ligand interactions associated with inhibition are glutamic acids, evidenced by the recurring trait of charged amines in potent inhibitors [83]. Other important residues for the interaction with nucleosides include a hydrophobic dyad formed by residues F1145 and W1170. It has recently been suggested that residue N1578 may serve as a bridge between the enthalpic polar glutamic acids subpocket and hydrophobic/aromatic regions of the pocket, providing an unexplored trait for DNMT1 inhibition [16].

The design of HDAC inhibitors often relies on zinc-binding groups [84] that result in potent but non-specific HDAC compounds. Accordingly, several hypotheses have been proposed to guide the design of selective inhibitors towards HDAC2. Recent observations point to hydrogen bonding for selectivity [85]. Other compounds, such as benzamides, have shown weak interactions towards the zinc center. However, benzamides also show selectivity for class I HDACs [86]. Molecular modeling of computational hits showed adequate interaction profiles, illustrated by the docking poses of screened compounds showing good agreement with co-crystal references, with higher deviations present for HDAC2 (Figs. 7 and 8). We hypothesize that this difference arises primarily from the absence of a zinc anchor and the size of the compounds. Nevertheless, consistency among poses was observed across shared scaffolds. Therefore, the binding mode for HDAC2 may not be the most representative one because of the pre-existing conformation of the binding pocket, which is a well-known bias found in SBDD approaches [87,88].

Similar to the pharmacophoric alignment, molecular superposition of docking poses to co-crystallized ligands was conducted to identify putative binding configurations (Figures S11-S12, in the Supporting Information). To this aim, we used complementary LBDD and SBDD methods to compare the pharmacophoric alignments and gain further insight into the molecular traits needed for dual DNMT1/HDAC2 inhibition.

Molecular dynamics simulations of the ligand-protein complexes

We performed MD simulations to determine the stability of the ligand-protein interactions predicted by molecular docking and to examine potential variations in the orientation and conformation of the molecules in the binding pockets. The selection of docking poses is a significant challenge in SBDD [89], so we selected both consistent and divergent poses to analyze the interaction profile of computational hits further. In this way, it is possible to better distinguish true binders from decoys [90]. Results showed that computational hits and protein complexes are relatively stable during production runs (Figures S13-S32 in the Supporting Information).

Histograms in Figures S32 and S33 show contact concurrences in MD simulations, as obtained with Contact Map. These are computed based on distance thresholds between protein and ligand. Concurrences may be used to identify metastable states in binding events. All hits showed reasonable proximity with F1145 and W1170; other notable contacts include S1146 which forms contacts with SAM and sinefungin, and with N1578, a prominent contact proposed for novel DNMT1 inhibitors [16]. For HDAC2, the predicted ligands are more mobile than in DNMT1, although this is probably an inherent behavior given their size and orientation in the binding pocket. A notable trait is the hydrophobic contact of the molecules with phenylalanine. Similar results have been obtained in other studies [91]. Molecular modeling results (docking and dynamics simulations) support the hypothesis of dual inhibition.

Conclusions

Drug discovery endeavors are on the hunt for novel methodologies or optimization strategies to diminish the high attrition rate and cost involved. Polypharmacology become a promising venue for better therapies, however, this increases the complexity of the rational design. This work introduces a general CADD pipeline that integrates LBDD and SBDD to support dual-target compound discovery and design. As a case study, we successfully developed a classifier for DNMT1/HDAC2 inhibitors using public datasets and open-access software. The classifier uncovered two promising scaffolds in the COCONUT natural product database that was used as an example of a large screening compound database. Both scaffolds showed high scores and good interaction with both proteins at the molecular level. We conclude that the flexibility and adaptability of the proposed pipeline has predictive capabilities of similar or derivative methods and is readily applicable to the discovery of small molecules targeting many other therapeutically relevant proteins.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the National Institutes of Health grant R01HL148068 (to L.M.E.-F.), EU ERA Chair Sinfonia Project, grant agreement 857253 (to E.F.-de G.), by DGAPA, UNAM, Programa de Apoyo a Proyectos de Investigación e Innovación Tecnológica (PAPIIT) grant IN201321 (to J.L.M.-F.), and by a CONACyT Postdoctoral Fellowship grant 31146 (to F.D.P.-M.). We thank Dr. Diego Prada-Gracia of the Unit of Computational Biology and Drug Design of Children's Hospital of Mexico Federico Gómez for granting access to the high-computing cluster Ixtlilton used for model training, decoy generation, and molecule optimization described in this work. This research was supported in part through computational resources and services provided by Advanced Research Computing at the University of Michigan, Ann Arbor, Michigan.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.ailsci.2021.100008.

References

[1] Zaware N, Zhou MM. Chemical modulators for epigenome reader domains as emerging epigenetic therapies for cancer and inflammation. Curr Opin Chem Biol 2017;39:116–25. doi:10.1016/j.cbpa.2017.06.012.

[2] Adhikari N, Jha T, Ghosh B. Dissecting histone deacetylase 3 in multiple disease conditions: selective inhibition as a promising therapeutic strategy. J Med Chem 2021;64:8827–69. 10.1021/acs.jmedchem.0c01676.

[3] Robert C, Rassool FV. HDAC inhibitors: roles of DNA damage and repair. Adv Cancer Res 2012;116:87–129. doi:10.1016/B978-0-12-394387-3.00003-3.

[4] Leus NG, Zwinderman MR, Dekker FJ. Histone deacetylase 3 (HDAC 3) as emerging drug target in NF-kappaB-mediated inflammation. Curr Opin Chem Biol 2016;33:160–8. 10.1016/j.cbpa.2016.06.019.

[5] Julg B, Barouch DH. Novel immunological strategies for HIV-1 eradication. J Virus Erad 2015;1:232–6.

[6] Jain AK, Barton MC. Bromodomain histone readers and cancer. J Mol Biol 2017;429:2003–10. doi:10.1016/j.jmb.2016.11.020.

[7] Duan YC, Zhang SJ, Shi XJ, Jin LF, Yu T, Song Y, Guan YY. Research progress of dual inhibitors targeting crosstalk between histone epigenetic modulators for cancer therapy. Eur J Med Chem 2021;222:113588. doi:10.1016/j.ejmech.2021.113588.

[8] Naveja JJ, Medina-Franco JL. Insights from pharmacological similarity of epigenetic targets in epipolypharmacology. Drug Discov Today 2018;23:141–50. doi:10.1016/j.drudis.2017.10.006.

[9] Tomaselli D, Lucidi A, Rotili D, Mai A. Epigenetic polypharmacology: a new frontier for epi-drug discovery. Med Res Rev 2020;40:190–244. doi:10.1002/med.21600.

[10] de Lera AR, Ganesan A. Two-hit wonders: the expanding universe of multitargeting epigenetic agents. Curr Opin Chem Biol 2020;57:135–54. doi:10.1016/j.cbpa.2020.05.009.

[11] Esteller M. Epigenetics in cancer. N Engl J Med 2008;358:1148–59. doi:10.1056/NEJMra072067.

[12] Nishiyama A, Nakanishi M. Navigating the DNA methylation landscape of cancer. Trends Genet 2021. doi:10.1016/j.tig.2021.05.002.

[13] Ming X, Zhu B, Li Y. Mitotic inheritance of DNA methylation: more than just copy and paste. J Genet Genomics 2021;48:1–13. doi:10.1016/j.jgg.2021.01.006.

[14] Lafeuillade A, Hittinger G, Chadapaud S. Increased mitochondrial toxicity with ribavirin in HIV/HCV coinfection. Lancet 2001;357:280–1. doi:10.1016/S0140-6736(00)03618-7.

[15] Mckenzie R, Fried MW, Sallie R, Conjeevaram H, Dibisceglie AM, Park Y, Savarese B, Kleiner D, Tsokos M, Luciano C, Pruett T, Stotka JL, Straus SE, Hoofnagle JH. Hepatic-failure and lactic-acidosis due to fialuridine (Fiau), an investigational nucleoside analog for Chronic Hepatitis-B. New Engl J Med 1995;333:1099–105. doi:10.1056/Nejm199510263331702.

[16] Juarez-Mercado KE, Prieto-Martinez FD, Sanchez-Cruz N, Pena-Castillo A, Prada-Gracia D, Medina-Franco JL. Expanding the structural diversity of DNA methyltransferase inhibitors. Pharmaceuticals (Basel) 2020;14. doi:10.3390/ph14010017.

[17] Medina-Franco JL. Grand challenges of computer-aided drug design: the road ahead. Front Drug Discov 2021;1:728551. doi:10.3389/fddsv.2021.728551.

[18] Sabe VT, Ntombela T, Jhamba LA, Maguire GEM, Govender T, Naicker T, Kruger HG. Current trends in computer aided drug design and a highlight of drugs discovered via computational techniques: a review. Eur J Med Chem 2021;224:113705. doi:10.1016/j.ejmech.2021.113705.

[19] Hopkins AL. Network pharmacology: the next paradigm in drug discovery. Nat Chem Biol 2008;4:682–90. doi:10.1038/nchembio.118.

[20] Zhang W, Pei J, Lai L. Computational multitarget drug design. J Chem Inf Model 2017;57:403–12. doi:10.1021/acs.jcim.6b00491.

[21] Achenbach J, Klingler FM, Blocher R, Moser D, Hafner AK, Rodl CB, Kretschmer S, Kruger B, Lohr F, Stark H, Hofmann B, Steinhilber D, Proschak E. Exploring the chemical space of multitarget ligands using aligned self-organizing maps. ACS Med Chem Lett 2013;4:1169–72. doi:10.1021/ml4002562.

[22] Paul SM, Mytelka DS, Dunwiddie CT, Persinger CC, Munos BH, Lindborg SR, Schacht AL. How to improve R&D productivity: the pharmaceutical industry's grand challenge. Nat Rev Drug Discov 2010;9:203–14. doi:10.1038/nrd3078.

[23] Nyigo VA, Malebo HM. Drug discovery and developments in developing countries: bottlenecks and way forward. Tanzan Health Res Bull 2005;7:154–8. doi:10.4314/thrb.v7i3.14253.

[24] Willems H, De Cesco S, Svensson F. Computational chemistry on a budget: supporting drug discovery with limited resources. J Med Chem 2020;63:10158–69. doi:10.1021/acs.jmedchem.9b02126.

[25] Anighoro A, Bajorath J, Rastelli G. Polypharmacology: challenges and opportunities in drug discovery. J Med Chem 2014;57:7874–87. doi:10.1021/jm5006463.

[26] Espinoza-Fonseca LM. The benefits of the multi-target approach in drug design and discovery. Bioorg Med Chem 2006;14:896–7. doi:10.1016/j.bmc.2005.09.011.

[27] Csermely P, Agoston V, Pongor S. The efficiency of multi-target drugs: the network approach might help drug design. Trends Pharmacol Sci 2005;26:178–82. doi:10.1016/j.tips.2005.02.007.

[28] Clayton T, Chen JL, Ernst M, Richter L, Cromer BA, Morton CJ, Ng H, Kaczorowski CC, Helmstetter FJ, Furtmuller R, Ecker G, Parker MW, Sieghart W, Cook JM. An updated unified pharmacophore model of the benzodiazepine binding site on gamma-aminobutyric acid(a) receptors: correlation with comparative models. Curr Med Chem 2007;14:2755–75. doi:10.2174/092986707782360097.

[29] Evans BE, Rittle KE, Bock MG, DiPardo RM, Freidinger RM, Whitter WL, Lundell GF, Veber DF, Anderson PS, Chang RS, et al. Methods for drug discovery: development of potent, selective, orally effective cholecystokinin antagonists. J Med Chem 1988;31:2235–46. doi:10.1021/jm00120a002.

[30] Filippakopoulos P, Picaud S, Fedorov O, Keller M, Wrobel M, Morgenstern O, Bracher F, Knapp S. Benzodiazepines and benzotriazepines as protein interaction inhibitors targeting bromodomains of the BET family. Bioorg Med Chem 2012;20:1878–86. doi:10.1016/j.bmc.2011.10.080.

[31] Costantino L, Barlocco D. Privileged structures as leads in medicinal chemistry. Curr Med Chem 2006;13:65–85. doi:10.2174/092986706775197999.

[32] Welsch ME, Snyder SA, Stockwell BR. Privileged scaffolds for library design and drug discovery. Curr Opin Chem Biol 2010;14:347–61. doi:10.1016/j.cbpa.2010.02.018.

[33] Meyers J, Chessum NEA, Ali S, Mok NY, Wilding B, Pasqua AE, Rowlands M, Tucker MJ, Evans LE, Rye CS, O'Fee L, Le Bihan YV, Burke R, Carter M, Workman P, Blagg J, Brown N, van Montfort RLM, Jones K, Cheeseman MD. Privileged Structures and Polypharmacology within and between Protein Families. ACS Med Chem Lett 2018;9:1199–204. 10.1021/acsmedchemlett.8b00364.

[34] Winter R, Montanari F, Steffen A, Briem H, Noe F, Clevert DA. Efficient multi-objective molecular optimization in a continuous latent space. Chem Sci 2019;10:8016–24. doi:10.1039/c9sc01928f.

[35] Winter R, Montanari F, Noe F, Clevert DA. Learning continuous and data-driven molecular descriptors by translating equivalent chemical representations. Chem Sci 2019;10:1692–701. doi:10.1039/c8sc04175j.

[36] Fernandez-de Gortari E, Medina-Franco JL. Epigenetic relevant chemical space: a chemoinformatic characterization of inhibitors of DNA methyltransferases. RSC Adv 2015;5:87465–76. doi:10.1039/c5ra19611f.

[37] Mendez D, Gaulton A, Bento AP, Chambers J, De Veij M, Felix E, Magarinos MP, Mosquera JF, Mutowo P, Nowotka M, Gordillo-Maranon M, Hunter F, Junco L, Mugumbate G, Rodriguez-Lopez M, Atkinson F, Bosc N, Radoux CJ, Segura-Cabrera A, Hersey A, Leach AR. ChEMBL: towards direct deposition of bioassay data. Nucleic Acids Res 2019;47:D930–40. doi:10.1093/nar/gky1075.

[38] Prieto-Martinez FD, Fernandez-de Gortari E, Mendez-Lucio O, Medina-Franco JL. A chemical space odyssey of inhibitors of histone deacetylases and bromodomains. RSC Adv 2016;6:56225–39. doi:10.1039/c6ra07224k.

[39] Fourches D, Muratov E, Tropsha A. Trust, but verify: on the importance of chemical structure curation in cheminformatics and QSAR modeling research. J Chem Inf Model 2010;50:1189–204. doi:10.1021/ci100176x.

[40] O'Boyle NM, Banck M, James CA, Morley C, Vandermeersch T, Hutchison GR. Open Babel: an open chemical toolbox. J Cheminform 2011;3:33. doi:10.1186/1758-2946-3-33.

[41] 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;55:460–73. doi:10.1021/ci500588j.

[42] Gaulton A, Bellis LJ, Bento AP, Chambers J, Davies M, Hersey A, Light Y, McGlinchey S, Michalovich D, Al-Lazikani B, Overington JP. ChEMBL: a large-scale bioactivity database for drug discovery. Nucleic Acids Res 2012;40:D1100–7. 10.1093/nar/gkr777.

[43] Kim S, Chen J, Cheng T, Gindulyte A, He J, He S, Li Q, Shoemaker BA, Thiessen PA, Yu B, Zaslavsky L, Zhang J, Bolton EE. PubChem 2019 update: improved access to chemical data. Nucleic Acids Res 2019;47:D1102–9. doi:10.1093/nar/gky1033.

[44] Imrie F, Bradley AR, Deane CM. Generating property-matched decoy molecules using deep learning. Bioinformatics 2021. doi:10.1093/bioinformatics/btab080.

[45] Sud M. MayaChemTools: an open source package for computational drug discovery. J Chem Inf Model 2016;56:2292–7. doi:10.1021/acs.jcim.6b00505.

[46] Oxford Protein Informatics Group, http://opig.stats.ox.ac.uk/resources (accessed August 19 2021).

[47] Rogers D, Hahn M. Extended-connectivity fingerprints. J Chem Inf Model 2010;50:742–54. doi:10.1021/ci100050t.

[48] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E. Scikit-learn: machine Learning in Python. J Mach Learn Res 2011;12:2825–30.

[49] Chen TQ, Guestrin C. XGBoost: A Scalable Tree Boosting System, Kdd'16. In: Proceedings of the 22nd ACM Sigkdd international conference on knowledge discovery and data mining; 2016. p. 785–94. doi:10.1145/2939672.2939785.

[50] Bickerton GR, Paolini GV, Besnard J, Muresan S, Hopkins AL. Quantifying the chemical beauty of drugs. Nat Chem 2012;4:90–8. doi:10.1038/nchem.1243.

[51] 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. doi:10.1186/1758-2946-1-8.

[52] Hassanzadeh M, Kasymov R, Mahernia S, Adib M, Emperle M, Dukatz M, Bashtrykov P, Jeltsch A, Amanlou M. Discovery of novel and selective DNA Methyltransferase 1 inhibitors by pharmacophore and docking-based virtual screening. Chem Select 2017;2:8383–92. doi:10.1002/slct.201701734.

[53] Kalyaanamoorthy S, Chen YP. Energy based pharmacophore mapping of HDAC inhibitors against class I HDAC enzymes. Biochim Biophys Acta 2013;1834:317–28. doi:10.1016/j.bbapap.2012.08.009.

[54] Luo Y, Li H. Structure-based inhibitor discovery of class I histone deacetylases (HDACs). Int J Mol Sci 2020;21. doi:10.3390/ijms21228828.

[55] Yuan Z, Chen S, Gao C, Dai Q, Zhang C, Sun Q, Lin JS, Guo C, Chen Y, Jiang Y. Development of a versatile DNMT and HDAC inhibitor C02S modulating multiple cancer hallmarks for breast cancer therapy. Bioorg Chem 2019;87:200–8. 10.1016/j.bioorg.2019.03.027.

[56] Krenn M, Häse F, Nigam A, Friederich P, Aspuru-Guzik A. Self-referencing embedded strings (SELFIES): a 100% robust molecular string representation. Sci. Technol. 2020;1:045024. doi:10.1088/2632-2153/aba947.

[57] Atanasov AG, Zotchev SB, Dirsch VM, International Natural Product Sciences T, Supuran CT. Natural products in drug discovery: advances and opportunities. Nat Rev Drug Discov 2021;20:200–16. doi:10.1038/s41573-020-00114-z.

[58] Newman DJ, Cragg GM. Natural products as sources of new drugs from 1981 to 2014. J Nat Prod 2016;79:629–61. 10.1021/acs.jnatprod.5b01055.

[59] Sorokina M, Merseburger P, Rajan K, Yirik MA, Steinbeck C. Coconut online: Collection of Open Natural Products database. J Cheminform 2021;13:2. doi:10.1186/s13321-020-00478-9.

[60] van der Maaten L, Hinton G. Visualizing data using t-SNE. J Mach Learn Res 2008;9:2579–605.

[61] Maggiora GM. On outliers and activity cliffs–why QSAR often disappoints. J Chem Inf Model 2006;46:1535. doi:10.1021/ci060117s.

[62] Yuan Z, Sun Q, Li D, Miao S, Chen S, Song L, Gao C, Chen Y, Tan C, Jiang Y. Design, synthesis and anticancer potential of NSC-319745 hydroxamic acid derivatives as DNMT and HDAC inhibitors. Eur J Med Chem 2017;134:281–92. 10.1016/j.ejmech.2017.04.017.

[63] Korb O, Monecke P, Hessler G, Stutzle T, Exner TE. pharmACOphore: multiple flexible ligand alignment based on ant colony optimization. J Chem Inf Model 2010;50:1669–81. doi:10.1021/ci1000218.

[64] Korb O, Stutzle T, Exner TE. Empirical scoring functions for advanced protein-ligand docking with PLANTS. J Chem Inf Model 2009;49:84–96. doi:10.1021/ci800298z.

[65] Salomon-Ferrer R, Gotz AW, Poole D, Grand SLe, Walker RC. Routine microsecond molecular dynamics simulations with AMBER on GPUs. 2. Explicit solvent particle mesh ewald. J Chem Theory Comput 2013;9:3878–88. doi:10.1021/ct400314y.

[66] Tian C, Kasavajhala K, Belfon KAA, Raguette L, Huang H, Migues AN, Bickel J, Wang YZ, Pincay J, Wu Q, Simmerling C. ff19SB: amino-acid-specific protein backbone parameters trained against quantum mechanics energy surfaces in solution. J Chem Theory Comput 2020;16:528–52. 10.1021/acs.jctc.9b00591.

[67] Wang JM, Wolf RM, Caldwell JW, Kollman PA, Case DA. Development and testing of a general amber force field. J Comput Chem 2004;25:1157–74. doi:10.1002/jcc.20035.

[68] McGibbon RT, Beauchamp KA, Harrigan MP, Klein C, Swails JM, Hernandez CX, Schwantes CR, Wang LP, Lane TJ, Pande VS. MDTraj: a Modern open library for the analysis of molecular dynamics trajectories. Biophys J 2015;109:1528–32. doi:10.1016/j.bpj.2015.08.015.

[69] Swenson, D.W.H. and Roet, S. Contact Map Explorer https://github.com/dwhswenson/contact_map (accessed August 19 2021).

[70] Ehrman TM, Barlow DJ, Hylands PJ. Virtual screening of Chinese herbs with random forest. J Chem Inf Model 2007;47:264–78. doi:10.1021/ci600289v.

[71] Sanchez-Cruz N, Medina-Franco JL, Mestres J, Barril X. Extended connectivity interaction features: improving binding affinity prediction through chemical description. Bioinformatics 2021;37:1376–82. doi:10.1093/bioinformatics/btaa982.

[72] Manelfi C, Gemei M, Talarico C, Cerchia C, Fava A, Lunghini F, Beccari AR. Molecular Anatomy": a new multi-dimensional hierarchical scaffold analysis tool. J Cheminform 2021;13:54. doi:10.1186/s13321-021-00526-y.

[73] Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. J Clin Epidemiol 2019;110:12–22. doi:10.1016/j.jclinepi.2019.02.004.

[74] Rashmi KV, Gilad-Bachrach R. DART: dropouts meet multiple additive regression trees. JMLR Worksh Conf Pro 2015;38:489–97.

[75] Shin SJ, Kim H, Ham S-t. Comparison of the performance evaluations in classification. Int J Adv Res Comput Commun Eng 2016;5:441–4. doi:10.17148/IJARCCE.2016.5890.

[76] Chicco D, Jurman G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics 2020;21:6. doi:10.1186/s12864-019-6413-7.

[77] Chicco D, Totsch N, Jurman G. The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation. BioData Min 2021;14:13. doi:10.1186/s13040-021-00244-z.

[78] Boughorbel S, Jarray F, El-Anbari M. Optimal classifier for imbalanced data using matthews correlation coefficient metric. PLoS ONE 2017;12:e0177678. doi:10.1371/journal.pone.0177678.

[79] M.A. Lones, How to avoid machine learning pitfalls: a guide for academic researchers, 2021, pp. arXiv:2108.02497.

[80] Chavez-Hernandez AL, Sanchez-Cruz N, Medina-Franco JL. Fragment library of natural products and compound databases for drug discovery. Biomolecules 2020;10. doi:10.3390/biom10111518.

[81] Cruz-Monteagudo M, Medina-Francos JL, Perez-Castillo Y, Nicolotti O, Cordeiro MNDS, Borges F. Activity cliffs in drug discovery: Dr Jekyll or Mr Hyde? Drug Discov. Today 2014;19:1069–80. doi:10.1016/j.drudis.2014.02.003.

[82] Klenner A, Hartenfeller M, Schneider P, Schneider G. Fuzziness' in pharmacophore-based virtual screening and de novo design. Drug Discov Today Technol 2010;7:e203–70. doi:10.1016/j.ddtec.2010.10.004.

[83] San Jose-Eneriz E, Agirre X, Rabal O, Vilas-Zornoza A, Sanchez-Arias JA, Miranda E, Ugarte A, Roa S, Paiva B, Estella-Hermoso de Mendoza A, Alvarez RM, Casares N, Segura V, Martin-Subero JI, Ogi FX, Soule P, Santiveri CM, Campos-Olivas R, Castellano G, de Barrena MGF, Rodriguez-Madoz JR, Garcia-Barchino MJ, Lasarte JJ, Avila MA, Martinez-Climent JA, 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. doi:10.1038/ncomms15424.

[84] Li Y, Wang F, Chen X, Wang J, Zhao Y, Li Y, He B. Zinc-dependent deacetylase (HDAC) inhibitors with different zinc binding groups. Curr Top Med Chem 2019;19:223–41. doi:10.2174/1568026619666190122144949.

[85] Bresciani A, Ontoria JM, Biancofiore I, Cellucci A, Ciammaichella A, Di Marco A, Ferrigno F, Francone A, Malancona S, Monteagudo E, Nizi E, Pace P, Ponzi S, Rossetti I, Veneziano M, Summa V, Harper S. Improved Selective Class I HDAC and Novel Selective HDAC3 Inhibitors: beyond Hydroxamic Acids and Benzamides. ACS Med Chem Lett 2019;10:481–6. 10.1021/acsmedchemlett.8b00517.

[86] Zhang L, Zhang J, Jiang Q, Zhang L, Song W. Zinc binding groups for histone deacetylase inhibitors. J Enzyme Inhib Med Chem 2018;33:714–21. doi:10.1080/14756366.2017.1417274.

[87] Ross GA, Morris GM, Biggin PC. One size does not fit all: the limits of structure-based models in drug discovery. J Chem Theory Comput 2013;9:4266–74. doi:10.1021/ct4004228.

[88] Anighoro A, Bajorath J. Three-dimensional similarity in molecular docking: prioritizing ligand poses on the basis of experimental binding modes. J Chem Inf Model 2016;56:580–7. 10.1021/acs.jcim.5b00745.

[89] Fischer A, Smiesko M, Sellner M, Lill MA. Decision making in structure-based drug discovery: visual inspection of docking results. J Med Chem 2021;64:2489–500. doi:10.1021/acs.jmedchem.0c02227.

[90] Liu K, Kokubo H. Exploring the stability of ligand binding modes to proteins by molecular dynamics simulations: a cross-docking study. J Chem Inf Model 2017;57:2514–22. doi:10.1021/acs.jcim.7b00412.

[91] Hassanzadeh M, Bagherzadeh K, Amanlou M. A comparative study based on docking and molecular dynamics simulations over HDAC-tubulin dual inhibitors. J Mol Graph Model 2016;70:170–80. doi:10.1016/j.jmgm.2016.10.007.