Epigenetic Target Prediction with Accurate Machine Learning Models

Norberto Sánchez-Cruz* and José L. Medina-Franco*

DIFACQUIM Research Group, Department of Pharmacy, School of Chemistry, Universidad Nacional Autónoma de México, Mexico City 04510, Mexico

KEYWORDS: Cheminformatics; Chemogenomics; Drug Discovery; Epigenetics; Epi-Informatics; Machine learning; Structure-activity relationships.

ABSTRACT:

Epigenetic targets are a significant focus for drug discovery research, as demonstrated by the eight approved epigenetic drugs for treatment of cancer and the increasing availability of chemogenomic data related to epigenetics. This data represents a large amount of structure-activity relationships that has not been exploited thus far for the development of predictive models to support medicinal chemistry efforts. Herein, we report the first large-scale study of 26318 compounds with a quantitative measure of biological activity for 55 protein targets with epigenetic activity. Through a systematic comparison of machine learning models trained on molecular fingerprints of different design, we built predictive models with high accuracy for the epigenetic target profiling of small molecules. The models were thoroughly validated showing mean precisions up to 0.952 for the epigenetic target prediction task. Our results indicate that the herein reported models have considerable potential to identify small molecules with epigenetic activity. Therefore, our results were implemented as freely accessible and easy-to-use web application.

INTRODUCTION

Since the introduction of the term *epigenetics* by Conrad Waddington in 1942 to denote the mechanisms that relate genotype to phenotype,[1] the term has been used with multiple meanings, going from the classic definition that refers to epigenetics as the study of the alterations in the biological phenotype without underlying changes in the DNA sequence,[2] to one of the most recent and general definitions: "the structural adaptation of chromosomal regions to register, signal, or perpetuate altered activity states."[3] At the molecular level, this adaptation involves the reversible modification of nucleic acids and histones. These modifications are catalyzed by a plethora of proteins, which could be considered as the core epigenetic targets, and that are classified into three main groups: (a) writers - enzymes capable of adding chemical groups to nucleic acids and histones - such as DNA methyltransferases (DNMTs), histone methyltransferases (HMTs) and histone acetyltransferases (HATs), (b) erasers - enzymes capable of removing marks introduced by the writers - such as histone deacetylases (HDACs) and histone demethylases (HDMs), and (c) readers - proteins with specialized domains capable of recognizing these changes - such as the bromodomain and external terminal protein (BET) family.[4] In addition to these core epigenetic targets, a wide range of proteins also play important roles in epigenetic regulation; these proteins include histone chaperones[5] (critical for nucleosome assembly), chromatin remodelers[6,7] (CHRs - responsible for moving, ejecting, and restructuring the nucleosome), and even some classes of transcription factors.[8]

Epigenetics is an essential component in an organism's normal development and responsiveness, so its dysregulation has been associated with altered gene expression patterns related to multiple diseases.[9–12] This makes epigenetic targets a significant focus for drug discovery research. Successful examples can be found in cancer research, with the approval of eight epigenetic drugs (drugs targeting epigenetic proteins) for clinical use: azacytidine and decitabine targeting DNMT1, vorinostat, belinostat, panobinostat, romidepsin

and tucsibinostat targeting HDACs, and tazemostat targeting an HMT (EZH2).[3,13] The importance of epigenetics in drug discovery is also illustrated by the increasing availability of chemogenomic databases related to epigenetics over the past decade.[14–18] An example of this is EpiFactors,[16] to the best of our knowledge, the database with the largest number of annotated proteins related to epigenetics reported so far, with a total of 815 different targets. In a recent work,[19] we surveyed the status of the compounds tested against these and other epigenetic targets identified from ChEMBL,[20] Therapeutic Target Database,[21] and scientific literature. We found out that for 136 of these targets, there are more than ten reported inhibitors, which meant a considerable increase in comparison with the 52 targets fulfilling the same criteria in 2017.[18] The rich structure-activity relationships (SAR) contained in these large data sets represents an excellent source of information to develop predictive models that have not been developed thus far on a large-scale basis. In a previous work the authors explored the SAR of epigenetic target data sets using the concept of activity landscape. Although that work was a quantitative study, it was descriptive.[22]

The increase in the publicly available chemogenomic data for all target classes over the years opened up the opportunity for the construction of ligand-based machine learning models to assist target prediction of small molecules. Some of these methods are currently available as easy-to-access web applications, such as Similarity Ensemble Approach[23] (SEA), HitPick,[24,25] Polypharmacology Browser[26,27] (PPB), TargetHunter,[28] and SwissTargetPrediction,[29,30] to name a few examples. These methods usually assign the targets for a given small molecule from the known targets of the most similar ligands in their datasets, employing different descriptions and metrics for the similarity assessment, and often making use of additional statistical models to estimate the significance of the predictions.[23,25,27] Despite of the increasing number of chemogenomic databases related to epigenetics, this data still represents a minimal amount when compared to other protein families such as kinases (KINs), ion channels or G protein-coupled receptors.[31,32] This suggests that epigenetic targets

are commonly underrepresented in the current target prediction methods, and that unless the similarity of a known ligand is high enough, they are less likely to be predicted as potential targets of small molecules, which points out the need of developing predictive models focused on epigenetic targets to assist medicinal chemistry efforts in this area.

Machine learning methods have proven to be useful in multiple areas of drug discovery,[33–35] one such being target prediction of small molecules.[25,27,30] For instance, in a retrospective large-scale comparison of machine learning methods for target prediction on ChEMBL (in the context of biochemical assays),[36] deep neural networks were the best performing method for this task when trained on Extended Connectivity Fingerprints[37] (ECFP) of chemical compounds.[38] However, the application of machine learning models for large-scale epigenetic target prediction has been explored on a limited basis, with most works focused on single targets[39,40] or protein families such as HDACs[41] or the BET family.[42]

Herein, we aimed to develop accurate models for epigenetic target prediction based on state-of-the-art machine learning algorithms trained on different fingerprint representation of compounds. We describe the development of predictive models with high precision for 55 epigenetic targets. Derivation of such predictive models is relevant for medicinal chemistry to develop hypothesis for the discovery of new epigenetic probes and drugs. The best models herein generated are implemented in an easy-to-use web application freely available to support medicinal chemistry projects related to epigenetic drug and probe discovery. It is anticipated that this tool will assist epigenetic drug design and development projects in the design and selection of compounds with potential epigenetic activity.

RESULTS

This section is organized into three major parts. First, we described the results of the data sets of epigenetic targets used in this work. The second part, entitled “Epigenetic Target Prediction with Machine Learning,” presents the results of the development of the machine learning models and their validation using two main strategies. The third main section, “Retrospective Identification of Epigenetic Targets,” shows, as a case study, a practical application of the best machine learning model derived in the second part, to identify epigenetic targets for external and recently reported compounds. All the details of the methods used are described in the Experimental Section.

Chemogenomic Data for Epigenetic Targets. Quantitative compound-protein associations were extracted from ChEMBL 27[20] and PubChem[43] to build epigenetic target-associated compound datasets meeting the following criteria: (a) containing at least 30 compounds with a quantitative measure of biological activity (IC_{50}, EC_{50}, K_i or K_d) lower or equal to 10 μM (“active”) and at least 30 compounds with a quantitative measure of biological activity higher than 10 μM (“inactive”), and (b) modelability index (MODI)[44] higher than 0.7 for at least one of the three molecular fingerprints selected as compound representation (see Experimental Section for further details). As illustrated in Figure 1, a total of 55 epigenetic targets were included and distributed as follows: (a) 26 writers, including 16 KINs, six HMTs, three HATs, and DNMT1, (b) 21 erasers, consisting of 12 HDACs, six HDMs, two proteins with dual activity (HDAC/HDM) and one protein related to histone ubiquitination (USP7), (c) four readers, including three bromodomain (BRD) containing proteins and one histone methyl-lysine binding protein (L3MBTL1), and (d) other proteins, consisting of three CHRs and one cofactor involved in DNA demethylation (APEX1). Details on the 55 epigenetic targets and their corresponding target-associated compound datasets are included as Table S1 in the Supporting Information.

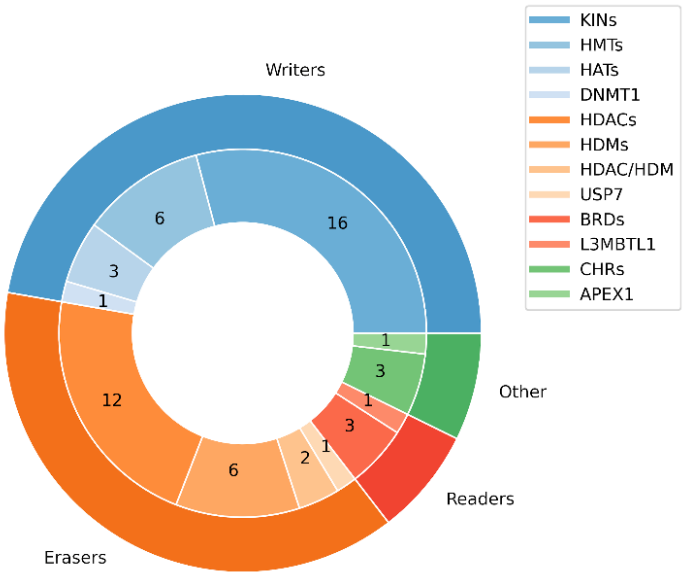

Figure 1. Distribution of Epigenetic Targets included in this work.

The compiled chemogenomic dataset contained 26318 unique compounds and 38129 compound-protein associations, with 28750 of them being labeled as active and 9379 labeled as inactive (due to the natural, although not the best practice of reporting mostly active compounds and not negative -inactive- data in ChEMBL). Consistently with the compound/compound-protein associations ratio, 20318 compounds (77.2%) in the dataset had known associations to a single target, and only 196 compounds (0.7%) had known associations to at least 10 targets, with a maximum of 15 targets for four compounds (Table 1).

Table 1. Distribution of known associations per compound.

Number of known associations	Number of compounds
1	20318
2	3853
3	1004
4	531
5	122
6	127
7	83
8	22
9	62
10	31
11	88
12	15
13	48
14	10
15	4
Total	26318

Epigenetic target-associated compound datasets consisted of 693 compounds on average, with a minimum of 73 for an HDM (KDM4E) and a maximum of 4901 for a KIN (JAK2). In agreement with the class imbalance in the entire dataset, all 55 compound datasets had different class imbalance levels, showing an average proportion of active compounds of 59.3%, with a minimum of 23.2% for a CHR (TOP2A) and a maximum of 92.4% for JAK2 (Figure 2).

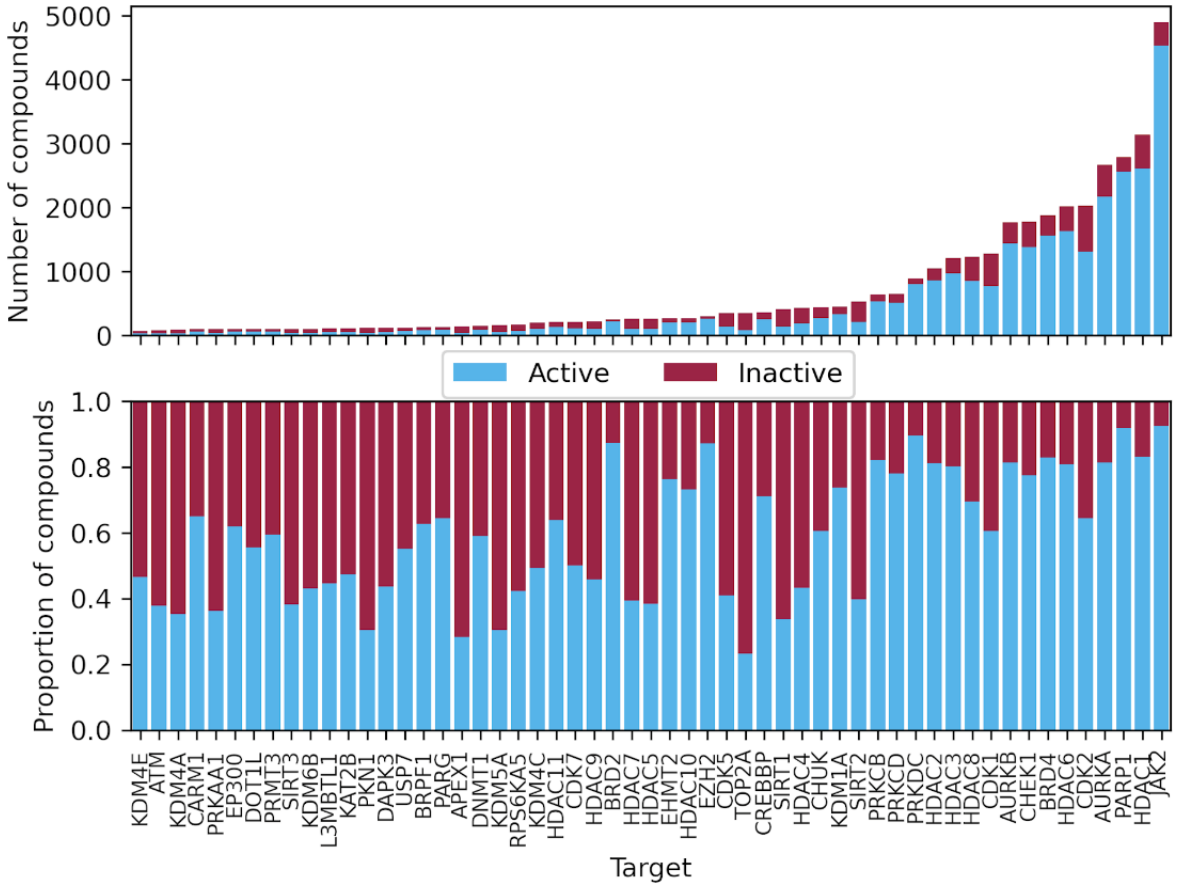

Figure 2. Size and composition of target-associated compound datasets.

Epigenetic Target Prediction with Machine Learning. Predictive models for epigenetic target prediction were built using two validation strategies summarized in Figure 3. The first strategy (Single Target Validation) involved the performance comparison of 15 different models on a stratified 10-fold cross-validation basis in the context of 55 single-target binary classification tasks. The two best performing models were combined to generate a consensus model, and the performance of these three models was assessed on a distance-to-model (DM) basis. The second strategy (Multi-Target Validation) focused on the global performance comparison of the best models identified in the first strategy when evaluated on 10 compound samples with the same number of known active associations for each epigenetic target. The results of each strategy are described in the next two sections.

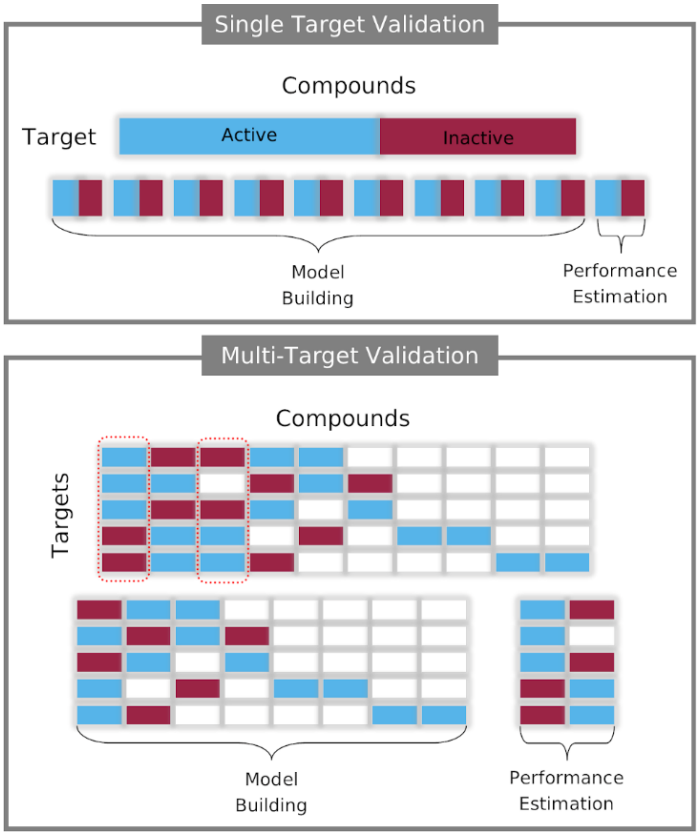

Figure 3. Two validation strategies employed for Epigenetic Target Prediction.

Single Target Validation. Fifteen different binary classification models with optimized hyperparameters were built for each of the 55 target-associated compound datasets. Models were derived from the combination between five state-of-the-art machine learning algorithms: *k*-nearest neighbors (*k*-NN)[45], Random Forest (RF)[46], Gradient Boosting Trees (GBT)[47], Support Vector Machines (SVM)[48], and Feed-Forward Neural Networks (FFNN)[49], and three molecular fingerprints of different design used as compound representations: Molecular ACCess System (MACCS) Keys (166-bit),[50] Morgan fingerprint with radius 2 (2048-bit),[37] and RDK fingerprint (2048-bit). Each model is denoted as a combination of fingerprint and algorithm (fingerprint::algorithm). For each algorithm and target, hyperparameters were optimized from an exhaustive search detailed in the Experimental Section, using the mean balanced accuracy (BA) over a 10-fold cross-validation as the performance metric to select the best set of hyperparameters.

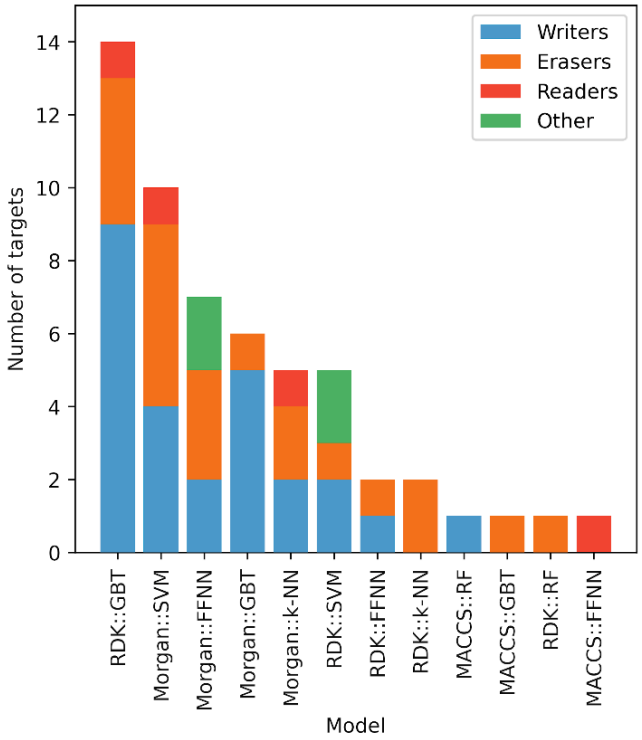

Figure 4. Distribution of best performing model per target class, considering balance accuracy as the evaluation metric.

Figure 4 shows the number of targets for which each model was identified as the best performing, considering the mean BA over the ten folds as a point metric. Under this approach, there is no model, fingerprint, nor machine learning algorithm that could be identified as the best performing for all 55 target datasets considered in this work. Figure 4 shows that RDK::GBT had the highest mean BA for 14 out of the 55 targets, making them the most frequent choice. However, in terms of compound representations only, Morgan fingerprint was the best choice for 28 targets, followed by 24 for RDK fingerprint and three for MACCS. Nevertheless, t-tests comparing the sets of BA scores calculated from the ten validation folds revealed that for all the targets, there is at least another model with no significant difference of performance to the one with the highest mean BA (Table S2 in the Supporting Information). Moreover, the t-test comparison revealed that for 35 out of the 55 targets, there are at least 9 other models with no significant difference of performance to the

one with the highest BA (at 95% confidence level), a surprising quantity considering the number of algorithms and compound representations included.

To compare the models herein generated in a more global context, the cross-validated predictions for each optimized model were stored and used to compute single point performance metrics in the context of each target, being BA, F1 score, and Mathews correlation coefficient (MMC). Summary results of the fifteen models' performance are summarized in Table 2, and their distribution across the 55 epigenetic targets is shown in Figure 5.

Table 2. Single Target Validation performance.

Model	BA	F1	MCC
Consensus	0.835 ± 0.067	0.851 ± 0.110	0.676 ± 0.123
Morgan::SVM	0.830 ± 0.065	0.862 ± 0.101	0.680 ± 0.123
RDK::SVM	0.827 ± 0.061	0.862 ± 0.096	0.670 ± 0.116
RDK::GBT	0.824 ± 0.067	0.859 ± 0.107	0.669 ± 0.123
RDK::FFNN	0.822 ± 0.057	0.859 ± 0.092	0.659 ± 0.108
Morgan::FFNN	0.819 ± 0.067	0.856 ± 0.100	0.651 ± 0.132
Morgan::*k*-NN	0.817 ± 0.068	0.859 ± 0.102	0.655 ± 0.134
RDK::RF	0.816 ± 0.067	0.856 ± 0.111	0.666 ± 0.115
Morgan::GBT	0.815 ± 0.073	0.855 ± 0.112	0.659 ± 0.136
RDK::*k*-NN	0.814 ± 0.063	0.855 ± 0.095	0.641 ± 0.124
Morgan::RF	0.811 ± 0.075	0.855 ± 0.118	0.663 ± 0.131
MACCS::SVM	0.807 ± 0.073	0.847 ± 0.118	0.632 ± 0.145
MACCS::GBT	0.806 ± 0.074	0.845 ± 0.117	0.629 ± 0.142
MACCS::RF	0.800 ± 0.072	0.846 ± 0.114	0.626 ± 0.134
MACCS::*k*-NN	0.791 ± 0.066	0.839 ± 0.109	0.600 ± 0.132
MACCS::FFNN	0.785 ± 0.069	0.829 ± 0.115	0.580 ± 0.137

Mean and standard deviation (mean ± SD) of BA, F1 and MCC for 55 single target binary classifiers built on 15 fingerprint::algorithm combinations and a consensus model. Results are sorted by decreasing BA.

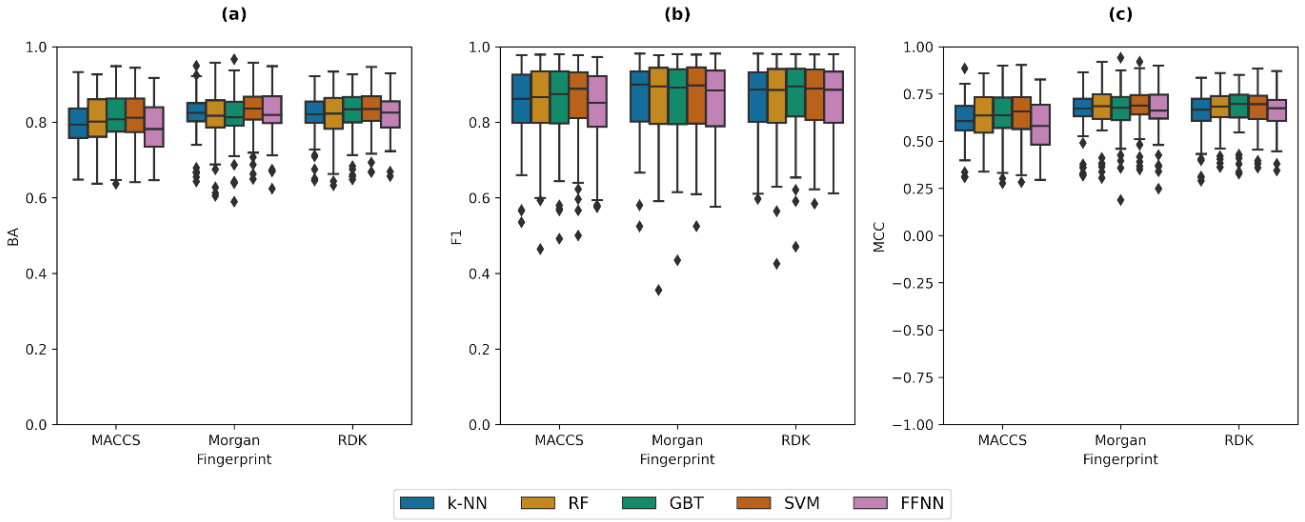

Figure 5. Performance comparison of single target binary classifiers. (a) balanced accuracy (BA), (b) F1 score, (c) Mathews correlation coefficient (MCC). Each boxplot contains the performance metrics for 55 different target-associated compound datasets.

Overall, most of the models performed well in the single-target prediction task, having a mean BA and F1 score higher than 0.5 and mean MCC higher than zero. To identify the global best performing model, we applied Wilcoxon signed-rank tests between all pairs of models for the three metrics of performance. Each test involves a comparison between sets of 55 values. The Morgan::SVM model showed the highest mean values for the three performance metrics and significantly higher values of BA and MCC when compared to all but the RDK::SVM model (at 95% confidence level). F1 score showed the lower differences between models, with the Morgan::SVM having the highest mean value and significantly higher values when compared to all but five models, being RDK::SVM, RDK::GBT, RDK::FNN, RDK::RF and Morgan::*k*-NN (at 95% confidence level). These results suggested Morgan and RDK fingerprints and the SVM algorithm as the best combinations to derive binary classifiers for the current sets of studied epigenetic targets.

Consensus Model. It has been pointed out that the combination of predictive models generally has a higher reliability than the individual models.[51,52] In other to identify the best models combination to construct a consensus model, we performed a hierarchical clustering

of the models relying on Morgan and RDK fingerprints by comparing their 38129 cross-validated predictions obtained in the single target validation strategy (vide supra). Jaccard distance was employed as the metric between models and an average linkage was used for the hierarchical clustering calculation as detailed in the Experimental Section. Figure 6 depicts a dendrogram of the hierarchical clustering. Predictions for all models are closely related, with all average distances between groups being lower than 0.1. It should be noted that models relying on the same fingerprint are clustered together before being grouped with models built on a different fingerprint. In the context of each fingerprint, the clustering follows the same order, with models relying on GBT and RF being grouped at first, followed by those built on SVM, FFNN, and *k*-NN.

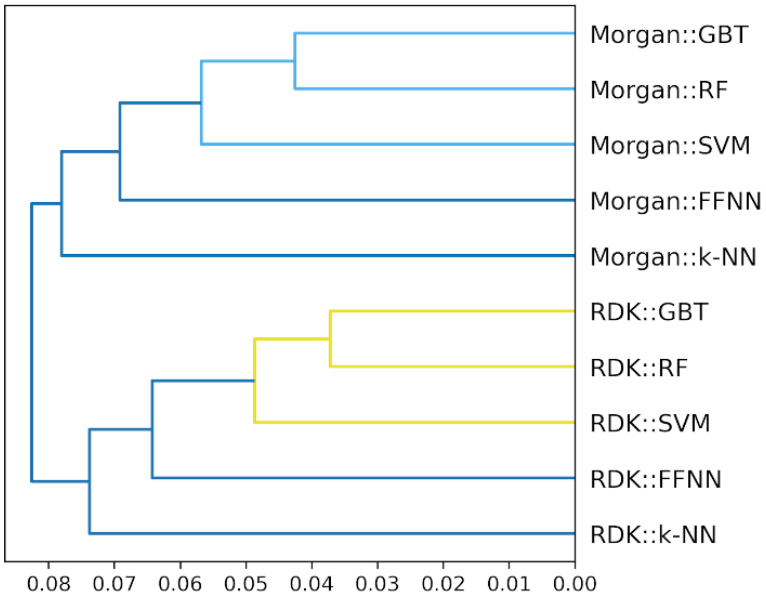

Figure 6. Hierarchical clustering of Morgan and RDK models. Average linkage and Jaccard distance between the models' predictions were used for the calculation.

Based on these findings, the best performing model built on each fingerprint, Morgan::SVM and RDK::SVM, were combined to derive a consensus model. To prioritize the correct identification of active compounds, the consensus model was constructed by combining the predictions of both models so that a compound was predicted as "active" for a given target

only if both models agreed in the prediction and “inactive” otherwise. This consensus model showed a mean BA, F1 score, and MCC of 0.835, 0.851, and 0.676, respectively. Wilcoxon signed-rank tests indicated significantly lower values for F1 score than those obtained by the individual models, and no significant difference for BA and MCC values (at 95% confidence level). Since F1 score is defined as the harmonic mean of precision (PPV) and recall (TPR) for the active class, and the consensus model was *a priori* built to have high precision, the significantly lower values obtained for F1 score are explained by a decrease in the TPR of the model (Table S3 and Figure S1 in the Supporting Information), which is related to the decrease in the number of “active” outcomes for the consensus model.

Distance-to-Model. Although BA, F1 score, and MCC are well-suited metrics for model performance estimation on imbalanced datasets, in a practical medicinal chemistry application, the correct identification of active compounds is often more important than the correct identification of inactive ones. To this end, the performance of the individual models and the derived consensus model were studied in terms of PPV, TPR, negative predictive value (NPV), and true negative rate (TNR). To estimate the models’ applicability domain, these metrics were computed on a distance-to-model (DM) basis as detailed in the Experimental Section. All cross-validated predictions were categorized into four quartiles (Q1-Q4) according to their mean Jaccard distances to the training set in the context of each target (Table S4 in the Supporting Information). Summary results of the three models’ performance are presented in Table 3, and their distribution across the 55 epigenetic targets is shown in Figure 7.

Table 3. Single Target Performance (Strategy I) in a distance-to-model basis.

Model	Quartile	PPV	TPR	NPV	TNR
Consensus	Q1	0.923 ± 0.081	0.894 ± 0.121	0.762 ± 0.197	0.777 ± 0.195
	Q2	0.914 ± 0.121	0.872 ± 0.143	0.790 ± 0.184	0.803 ± 0.197
	Q3	0.883 ± 0.114	0.826 ± 0.149	0.805 ± 0.114	0.855 ± 0.134
	Q4	0.810 ± 0.153	0.653 ± 0.242	0.764 ± 0.141	0.869 ± 0.152
Morgan::SVM	Q1	0.912 ± 0.086	0.915 ± 0.113	0.831 ± 0.175	0.741 ± 0.229
	Q2	0.893 ± 0.128	0.897 ± 0.131	0.827 ± 0.161	0.753 ± 0.222
	Q3	0.847 ± 0.141	0.864 ± 0.132	0.834 ± 0.099	0.800 ± 0.149
	Q4	0.781 ± 0.147	0.714 ± 0.226	0.791 ± 0.148	0.820 ± 0.176
RDK::SVM	Q1	0.891 ± 0.131	0.914 ± 0.123	0.792 ± 0.186	0.739 ± 0.203
	Q2	0.878 ± 0.137	0.901 ± 0.127	0.811 ± 0.189	0.721 ± 0.238
	Q3	0.843 ± 0.133	0.866 ± 0.134	0.840 ± 0.111	0.793 ± 0.181
	Q4	0.780 ± 0.137	0.730 ± 0.217	0.825 ± 0.095	0.822 ± 0.146

Mean and standard deviation (mean ± SD) of PPV, TPR, NPV and TNR for 55 single target binary classifiers built on two fingerprint::algorithm combinations and a consensus model.

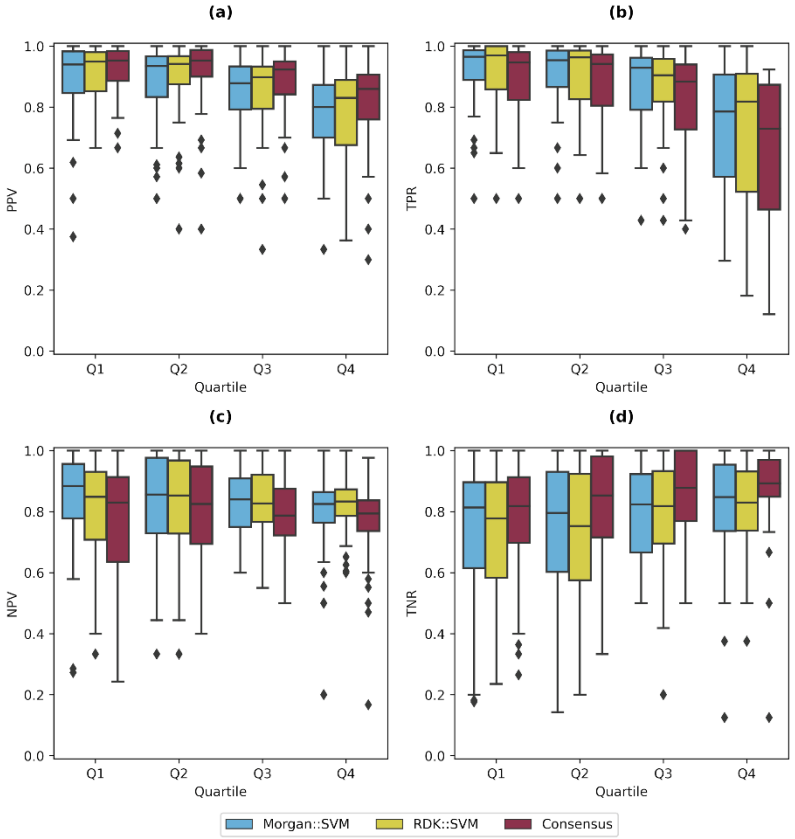

Figure 7. Performance comparison of single target binary classifiers in a distance-to-model basis. (a) positive predictive value (PPV), (b) true positive rate (TPR), (c) negative predictive value (NPV), (d) true negative rate (TNR). Each boxplot contains the performance metrics for up to 55 different target-associated compound datasets.

All performance metrics showed similar trends for the three models. As shown in Figure 7, PPV and TPR decreased as the distance from a compound to the training set increased, while, in the same scenario, NPV and TNR generally decreased. This suggests that predictions, particularly those for active compounds, are more reliable when the predicted compound is closer to the compounds in the training set. Wilcoxon signed-rank tests indicated significantly higher PPV and TNR values, and lower NPV and TPR values for all quartiles when comparing the consensus model to any of the two individual models (at 95% confidence level). These results agree with the lower probability of the consensus model of having an "active" outcome compared to the individual models (since both individual models must agree with the prediction). The lower number of compounds predicted active is associated with the lower recovery of the known active compounds (low TPR) and low precision in predicting inactive compounds (low NPV) compared to the individual models. However, this also implies that the known inactive compounds are well differentiated by the model (high TNR) and the precision in the prediction of active compounds is higher for the consensus model (high PPV), which is desirable in a typical medicinal chemistry project. It should be noted that despite the decrease in the PPV at high DM for the consensus model, the mean values of PPV for all quartiles were higher than 0.8, with a maximum 0.923 at Q1 and a minimum of 0.810 for Q4, suggesting high reliability on the predictions of active compounds, even when the predicted compounds are far from the compounds in the training set (Figure 7). Moreover, regardless of the performance difference in TPR and NPT between the consensus models and the individual models, these performance metrics for the consensus model are still high, showing mean values higher than 0.6 for all quartiles, where the lower mean values were 0.653 and 0.764 for TPR and NPT in Q4, respectively.

Multi-Target Validation. All results in the previous sections were analyzed in the cross-validated predictions of 55 individual binary classifiers. However, given that each classifier was trained and tested on compound datasets of different sizes, assessing the performance

of the combination of these 55 predictive models in the epigenetic target prediction task would lead to an incorrect performance estimation, with a bias over the targets with the most populated compound datasets associated. For this reason, 10 compound samples containing exactly six known active compounds for each target were assembled. For each sample, Morgan::SVM, RDK::SVM, and the consensus model were re-trained on the whole compounds datasets, excluding the compounds in the sample and evaluated on the sample initially excluded (as an external set). In this case, only metrics considering the correct identification of active compounds were calculated (PPV and TPR) on a DM basis following the same approach described in the previous section, considering only the predictions with a truly known label. Samples contained between 184 and 229 compounds (210 on average), and no more than 40 repeated compounds among them (Figure S2 in the Supporting Information). Summary results of the three models' performance are presented in Table 4, and their distribution across the 10 samples is shown in Figure 8.

Table 4. Multi-Target Performance (Strategy II) in a distance-to-model basis.

Model	Quartile	PPV	TPR
Consensus	Q1	0.952 ± 0.022	0.879 ± 0.033
	Q2	0.924 ± 0.036	0.833 ± 0.051
	Q3	0.822 ± 0.062	0.719 ± 0.065
	Q4	0.773 ± 0.056	0.558 ± 0.073
Morgan::SVM	Q1	0.948 ± 0.022	0.901 ± 0.027
	Q2	0.912 ± 0.030	0.871 ± 0.054
	Q3	0.744 ± 0.073	0.751 ± 0.058
	Q4	0.688 ± 0.063	0.624 ± 0.060
RDK::SVM	Q1	0.947 ± 0.019	0.918 ± 0.029
	Q2	0.899 ± 0.050	0.862 ± 0.059
	Q3	0.759 ± 0.086	0.772 ± 0.060
	Q4	0.707 ± 0.054	0.624 ± 0.056

Mean and standard deviation (mean ± SD) of PPV and TPR for 10 combinations of 55 single target binary classifiers built on two fingerprint::algorithm combinations and a consensus model.

Under this validation strategy, PPV and TPR showed the same trends as in the single target validation: both decreased as the DM increased for all models. Wilcoxon signed-rank tests indicated significantly lower TPR values and higher PPV values for all quartiles when comparing the consensus model to any of the two individual models (at 95% confidence level) (Table 4 and Figure 8).

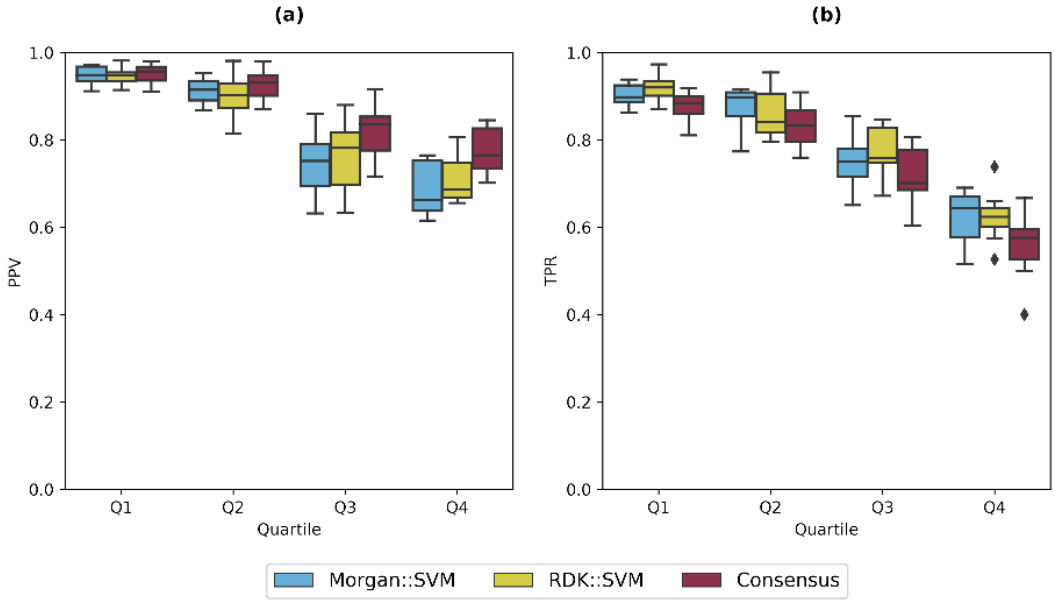

Figure 8. Performance comparison of the combination of 55 single target binary classifiers in a distance-to-model basis. (a) positive predictive value (PPV), (b) true positive rate (TPR). Each boxplot contains the performance metrics for up to 10 different combinations.

Retrospective Identification of Epigenetic Targets. As a proof of concept on the practical applicability of the herein developed consensus model, we employed it in the retrospective identification of the epigenetic targets for two external and recently reported compounds (Figure 9): (1) compound **17**, an inhibitor of EP300 and CREBBP, as targets representative of the less populated compound datasets, and (2) compound **43a**, an inhibitor of HDACs 1, 3, 6, 8 and BRD4, representing targets with the most populated compound datasets. These results were compared to those obtained by four general target prediction tools freely available online when performing this study: HitPickV2, PPB2, SEA, and SwissTargetPrediction. Figure 9

summarizes the results obtained by the consensus model and the four target prediction tools. The full list of predictions is available as Tables S5-S15 in the Supporting Information. The number of targets predicted by each of the web tools is fixed, being 10 for HitPickV2, 20 for PPB2 and SEA, and 100 for SwissTargetPrediction, while the herein reported consensus model was re-fitted using the entire datasets and set up to perform the predictions for the 55 epigenetic targets, with only those involving an “active” outcome considered as the predicted targets.

For compound **17**, our consensus model was the only one able to correctly identify CREBBP and EP300 as its targets (from 12 predicted targets). For compound **43a**, our model (from 18 predicted targets) and SEA identified correctly its five known targets. HitPickV2 and PPB2 predicted correctly the four HDACs but not BRD4, while SwissTargetPrediction predicted correctly only HDACs 1, 6 and 8. Although a more exhaustive external validation is needed, these results suggests that epigenetic targets with large amounts of chemogenomic data associated (such as HDACs) are generally well represented in current target prediction tools, while those with fewer data are not well covered. Moreover, it should be noted that the known epigenetic targets were not always among the top predictions for the available tools. For instance, HDACs 3, 6 and 8 were ranked 12, 13 and 15 by SEA, and HDACs 1, 6 and 8 from SwissTargetPrediction were ranked in positions 42, 43 and 44, so in a practical application, these targets would be hardly prioritized. Although the experimental validation of the predictions from all models would be needed to provide better means of comparison, these findings reinforce the potential usefulness of a tool focused on epigenetic targets for medicinal chemistry applications in drug discovery.

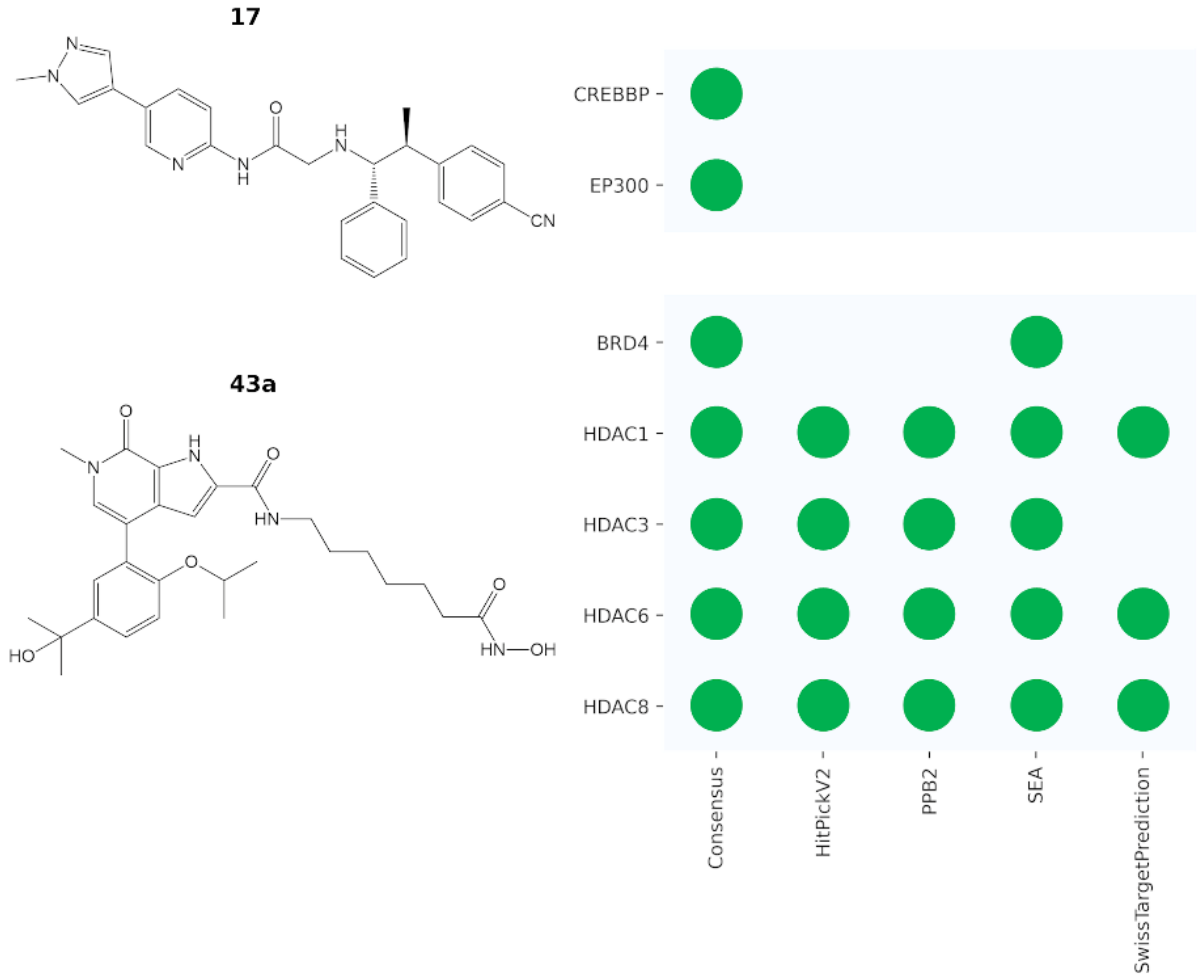

Figure 9. Comparison of target prediction tools for the retrospective identification of epigenetic targets.

Availability and Implementation. All row data to reproduce the results presented in this work is available free of charge at figshare repository (10.6084/m9.figshare.13519580). To encourage the medicinal chemistry community to apply the predictive consensus model developed in this work, the model was re-fitted using the entire datasets and has been implemented as a freely accessible and easy-to-use web application described in a separate work and available at http://www.epigenetictargetprofiler.com/.

DISCUSSION AND CONCLUSIONS

Epigenetic drug discovery is increasingly important across different therapeutic areas. Despite the large amount of SAR data stored in public data sets, that information has not been used on a large scale to develop predictive models that support the medicinal chemistry community's efforts working on these cutting-edge targets. To fill this gap, in this study, we developed and evaluated the performance of five state-of-the-art machine learning algorithms built on three molecular fingerprints of different designs to predict 55 epigenetic targets of small molecules. To the best of our knowledge, this is the first study covering epigenetic targets on a large-scale basis. The performance of the herein reported models was validated using two different approaches, involving their performance estimation for binary classifications in 10-fold cross-validations in the context of each target, as well as the performance of their combination in the epigenetic target prediction task evaluated over 10 balanced samples of compounds containing an equal number of known active compounds for each target.

Although none of the herein reported models was identified as the best performing one for all the 55 targets, our results suggested Morgan and RDK fingerprints as the best representations for the derivation of binary classifiers for the studied targets, particularly when derived using SVM, where no significative difference was found for their performance. This cannot be generalized for other, or even for these targets, since it could be associated with the hyperparameter space employed to optimize the models. Moreover, a model's performance is also dependent on the dataset composition, so the trends herein presented could change as more bioactivity data is published and different sets of hyperparameters are studied.

A consensus model was built by combining the predictions of the best models derived from Morgan and RDK fingerprints (Morgan::SVM and RDK::SVM), also supported on the fact that predictions between models relying on the same fingerprint are more closely related than

those relying on different representations as demonstrated by the hierarchical clustering analysis of their cross-validated predictions. The consensus models' performance and the two source models were analyzed on a DM basis, categorizing the predictions according to the Jaccard distance of the compounds in the test set to those in the training set. For the single target binary classification, the consensus model showed a significantly higher precision for identifying active compounds than those obtained by the individual models regardless of the DM. This trend was preserved when the models were evaluated to predict epigenetic targets.

The consensus model showed a mean BA of 0.835 considering the cross-validated predictions of the 55 target-associated binary classifiers, with mean precisions for identifying active compounds ranging from 0.923 for those compounds closer to the training set, to 0.810 for those farther from the training set. For the epigenetic target prediction task, mean precisions ranged from 0.952 to 0.773 under the same scheme.

We showed the consensus model's practical applicability by the retrospective identification of the epigenetic targets of two external and recently reported compounds. These results showed the consensus model as a robust and accurate method for epigenetic target prediction of small molecules, which led us to implement it as an easy-to-use web application available for free. It is hoped that this model will be helpful in practical medicinal chemistry applications for epigenetic drug discovery.

EXPERIMENTAL SECTION

Data Sets. Our primary source of SAR data was ChEMBL 27,[20] we collected all the quantitative compound-protein associations from single protein assays, related to the 136 epigenetic targets identified in our previous work[19] (biological activity reported as IC_{50}, EC_{50}, K_i or K_d). In the context of each target, compounds were labeled as "active" when they had unequivocally assigned activities lower than or equal to 10 μM, and as "inactive" in the opposite case. Compounds whose label could not be unequivocally assigned (e.g., activity <

100 μM or activity > 1 μM) were removed from the data set. The remaining compounds were curated using the open-source cheminformatics toolkit RDKit, version 2020.03.1 and the functions Standardizer, LargestFragmentChoser, Uncharger, Reionizer and TautomerCanonicalizer implemented in the molecule validation and standardization tool MolVS, as described in previous works.[54,55] In short, the Simplified Molecular Input Line Entry System[56] (SMILES) of each compound was standardized, those compounds consisting of multiple components were split and the largest component was retained. Compounds containing any element other than H, B, C, N, O, F, Si, P, S, Cl, Se, Br and I, as well as compounds with valence errors, were removed from the data set. The remaining compounds were neutralized and reionized to subsequently generate a canonical tautomer without preserved stereochemistry. Once all compounds were standardized, those with molecular weight higher than 800 Da as well as duplicated compounds with contradictory labels were removed. We preserved compound-protein associations only for those targets with at least 30 compounds labeled as "active," corresponding to 72 different targets. Since chemogenomic data for these epigenetic targets include a higher proportion of associations for "active" compounds (64% on average), we extended our initial data with "inactive" compounds from PubChem.[43] We included only compounds with annotated quantitative data (IC_{50}), all these compounds were curated using the same procedure described above and added only if they were not already included in the datasets. Finally, we kept 58 target-associated datasets containing at least 30 compounds labeled as "inactive."

Molecular Representations. To develop the machine learning models, we selected three molecular fingerprints of different design: (a) Molecular ACCess System (MACCS) Keys (166-bit)[50] as a dictionary based fingerprint where each position indicates presence or absence of a predefined structure, (b) Morgan fingerprint with radius 2 (2048-bit)[37] as a circular fingerprint where each position represents an atom environment including all atoms connected up to a radius of 2 bonds, and (c) RDK fingerprint (2048-bit) as a topological fingerprint where each

position represents a linear substructure including all atoms connected up to a length of 7 bonds. All fingerprints were generated using the open-source cheminformatics toolkit RDKit, version 2020.03.1 for Python.

Data Modelability. To *a priori* estimate the feasibility to obtain predictive binary classification models for each target, we calculated the modelability index (MODI)[44] for each target-associated dataset. MODI is defined as the proportion of compounds in a dataset for which its nearest neighbor belongs to the same class in a given feature space. For its calculation, we selected as compound representation the three different fingerprints described above and as metric to identify the nearest neighbors the Jaccard distance, defined as:

$$J(A,B) = 1 - \frac{c}{a+b-c}$$

where J (A, B) is the Jaccard distance between compounds A and B in a given fingerprint representation, with a and b being the number of "on" bits for compound A and B, respectively, and c being the number of "on" bits for both compounds. Further modeling was performed only for 55 datasets with a MODI higher or equal than 0.7 for at least one molecular representation.

Machine Learning Methods. Binary classification models for each target were generated using five different machine learning algorithms: *k*-nearest neighbors(*k*-NN)[45], Random Forest (RF)[46], Gradient Boosting Trees(GBT)[47], Support Vector Machines(SVM)[48], and Feed-Forward Neural Networks (FFNN)[49]. All machine learning methods were implemented using the Scikit-learn Python library (0.22.1).[57] For model building, training instances were represented by a feature vector (fingerprint) and associated to a class label ("active" / "inactive"). To avoid hyperparameter bias when comparing different models, the hyperparameters for each model were optimized using stratified 10-fold cross-validation in an exhaustive search over a limited hyperparameter space. To keep the search space small, only selected hyperparameters on each algorithm were optimized. Hereunder, we provide

brief explanations on each algorithm and the hyperparameters considered for its optimization; all hyperparameters not explicitly indicated in the text were set as default.

In *k*-NN classification, the predicted label of a sample is assigned according to the most common label among its *k* nearest neighbors in the training dataset for a given feature space. For this algorithm, we selected the Jaccard distance as the metric to identify the nearest neighbors using a brute-force search. The optimal number of nearest neighbors was optimized using candidate values of 1, 3, 5, 7, and 9.

RF is one of the so-called ensemble methods relying on decision trees. In RF classification, a fixed number of decision tree classifiers are fitted on various bootstrapped subsamples of the training dataset. For a given sample, each decision tree predicts a label, and the final prediction of the sample is the label predicted by most of the trees. For this algorithm, the number of decision trees was fixed to 1000 and the number of features to consider when searching for the best splits in the individual trees was optimized in a representation-dependent manner using candidate values of 1, 2, 3, 4 and 5 times the square root of the number of features in the fingerprint representation.

GBT is another ensemble method relying on decision trees. In this case, the decision tree classifiers are fitted in stages for the whole training dataset, where each subsequent tree is intended to “correct” the errors made by the previous one in terms of a loss function, usually the deviance of the fitted model with respect to a perfect model. For this algorithm, the number of decision trees was fixed to 1000, the number of features to consider when looking for the best splits in the individual trees was optimized in a representation-dependent manner using candidate values of 1, 2, 3, 4 and 5 times the square root of the number of features in the fingerprint representation, for the maximum depth of the individual trees we used candidate values of 4, 6, 8 and 10, and for the minimum number of samples to split an internal node in the individual trees we used candidate values of 2, 3, 4, and 5.

In SVM classification, the hyper-plane that best separate the two classes in the training dataset is constructed by maximizing the distance between training instances belonging to different classes (margin). As this hyper-plane does not always exist, a limited number of errors is allowed using a “cost” hyperparameter to control the relation between the training errors and the margin size. If linear separation of training classes is not possible in a given feature space, kernel functions are applied to project the data into a higher dimensional space where linear separation is possible. For this algorithm, “cost” was optimized using candidate values of 0.01, 0.1, 1.0, 10.0 and 100.0, and the kernel type to be used was selected from three options being non-kernel (“linear”), radial basis functions (“rbf”), and hyperbolic tangent (“sigmoid”).

A FFNN is composed by different layers of computational neurons: an input layer, one or more hidden layers, and an output layer. Neurons in the input layer are associated to the features describing the data, each neuron in the hidden layer accepts the inputs of all neurons in the input layer and transform them to a weighted sum of the original inputs, then a nonlinear activation function is applied to this weighted sum and the result is passed to the neurons in the output layer, where the prediction is performed. The weights from the network are iteratively adjusted during the training stage on the basis of a cost function to minimize, typically cross entropy. For this algorithm, the solver for weight optimization was set as “lbgfs”, the maximum number of iterations (how many times a training data point is passed to the network) was set to 1000, and the number of hidden layers was fixed to 1. The number of neurons in the hidden layer was optimized in a representation-dependent manner using candidate values of 0.1, 0.2, 0.3, 0.4, 0.5, 0.6 and 0.7 times the number of features in the fingerprint representation, and the activation function was selected from three different options, being logistic sigmoid function (“logistic”), hyperbolic tangent function (“tanh”) and rectified linear unit function (“relu”).

Training and Test Sets. For model building, two different validation strategies were implemented (Figure 3), the first comparing different combinations of fingerprints and machine learning algorithms in single-target binary classification tasks (Single Target Validation), and the second evaluating the best performing models from the first strategy in the epigenetic target prediction task (Multi-Target Validation).

Single Target Validation. Considering that the compound-target bioactivity matrix for the studied targets is sparse, the first strategy involved the construction of target-specific classification models and comparison of their performance across the different combinations of fingerprints and machine learning algorithms. Fifteen different binary classification models were built for each target, resulting from the combinations of the three fingerprints used as molecular representations and the five machine learning algorithms used for model fitting. Hyperparameters for each model were optimized using a stratified 10-fold cross-validation, with balanced accuracy (BA) employed as metric for selection of the best performing set of hyperparameters. The cross-validated predictions of the best model were used for the calculation of different performance metrics and comparison of the models. Each model performance was assessed using three metrics unbiased to the class imbalance in the data, BA, F1 score, and Mathews correlation coefficient (MMC), defined as:

$$BA = \frac{0.5\,TP}{TP + FN} + \frac{0.5\,TN}{TN + FP}$$

$$F1 = 2\,x\,\frac{TP}{2TP + FP + FN}$$

$$MCC = \frac{TP\,x\,TN - FP\,x\,FN}{\sqrt{(TP + FP)\,(TP + FN)(TN + FP)(TN + FN)}}$$

where TP means “true positives”, TN “true negatives”, FP “false positives”, and FN “false negatives”, with “positive” and “negative” refereeing to “active” and “inactive” compound labels, respectively.

We built a consensus model by combining the predictions of the best performing models showing the lower relation among their predictions. For that, we performed a hierarchical clustering with average linkage of the models relying on Morgan and RDK fingerprints (the best performing fingerprints), being described by their cross-validated predictions across all targets. As the distance metric for the construction of the hierarchical clustering, we selected de Jaccard distance defined in the Data Modelability section, where in this case J (A, B) represents the distance between two models, with a and b being the number of "active" predictions for model A and B, respectively, and c being the number of "active" predictions for both models.

We compared the consensus model and the single models of which it is composed using precision (positive predictive value - PPV), sensitivity (true positive rate - TPR), negative predictive value (NPV), and specificity (true negative rate - TNR), defined as:

$$PPV = \frac{TP}{TP + FP}$$

$$TPR = \frac{TP}{TP + FN}$$

$$NPV = \frac{TN}{TN + FN}$$

$$TNR = \frac{TN}{TN + FP}$$

In order to estimate the applicability domain of the models, these metrics were computed on a distance-to-model (DM) basis.[58,59] For that, the mean Jaccard distance from each compound in the test sets to all compounds in the training sets was calculated as the DM metric, using the three different fingerprints employed as molecular representation. These average distances were categorized in four quartiles considering all the cross-validated predictions, and all four metrics were calculated for each target and quartile, when predictions on the corresponding quartile were available.

Multi-Target Validation. The consensus model and the corresponding individual models were compared for the epigenetic target prediction problem. To assess the global performance of the combination of the single-target binary classifiers in the epigenetic target prediction task, ten samples of compounds containing the same number of active compounds for each target were assembled. To reduce the target-bias associated to the different sizes on the target-associated compound datasets, each of the compound samples was constructed by iteratively sampling one compound labeled as "active" from the less populated dataset in the sample (or in alphabetical order according to its gene code when there was more than one less populated sample). This process was performed until the sample contained exactly 6 active compounds (20% of the active compounds for the smaller dataset) for each target. If the addition of a compound yields a target containing more than 6 active compounds, the compound was discarded, and if the equal number of active compounds for each target was not satisfied after 1000 iterations, 10% of the sample was randomly discarded, and the iterative sampling continued. These ten samples were used as validation sets so that compounds in the sample were removed from the original target-associated datasets. The single target binary classifiers were refitted using the hyperparameters selected in the Single Target Validation strategy. The performance for the combination of the single-target binary classifiers was assessed by its capability of identifying the known active compounds among the known compound target associations, using PPV and TPR as metrics in the same DM basis described in the Single Target Validation strategy.

Retrospective Identification of Epigenetic Targets. Two external and recently reported compounds with more than one associated epigenetic target were selected to show the practical applicability of the herein reported models in a retrospective identification of its targets: (1) compound **17**,[60] a dual inhibitor of the HATs CREBBP (IC_{50} = 3.2 nM) and EP300 (IC_{50} = 2.5 nM), and (2) compound **43a**,[61] a *pan*-HDAC/BRD inhibitor with reported activities over BRD4 (IC_{50} = 29.5 nM), HDAC1 (IC_{50} = 19.4 nM), HDAC3 (IC_{50} = 36.4 nM), HDAC6 (IC_{50}

= 5.4 nM) and HDAC8 (IC_{50} = 99.6 nM). The consensus model was re-trained on the whole datasets using the hyperparameters identified in the single target validation strategy and 55 predictions were made for each of the external compounds. The epigenetic targets predicted for a compound were those for which the compound was predicted as “active” for the consensus model. These results were compared to those obtained from four currently freely available ligand-based tools for target prediction, being HitPickV2, SEA, PPB2 and SwissTargetPrediction.

ASSOCIATED CONTENT

Supporting Information

Table S1. Target-associated compound datasets included in this work.

Table S2. Models with no significant difference of performance.

Table S3. Single Target Performance.

Table S4. Distance-to-model quartiles.

Figure S1. Performance comparison of single target binary classifiers.

Figure S2. Compounds overlap between samples employed in the Multi-Target Validation.

Table S5. External compounds employed for retrospective target prediction.

Table S6. Targets predicted by the consensus model for compound 17.

Table S7. Targets predicted by HitPickV2 for compound 17.

Table S8. Targets predicted by PPB for compound 17.

Table S9. Targets predicted by SEA for compound 17.

Table S10. Top 45 targets predicted by SwissTargetPrediction for compound 17.

Table S11. Targets predicted by the consensus model for compound 43a.

Table S12. Targets predicted by HitPickV2 for compound 43a.

Table S13. Targets predicted by PPB for compound 43a.

Table S14. Targets predicted by SEA for compound 43a.

Table S15. Top 45 targets predicted by SwissTargetPrediction for compound 43a.

AUTHOR INFORMATION

Corresponding Authors

José L. Medina-Franco; orcid.org/0000-0003-4940-1107; E-mail: medinajl@unam.mx

Norberto Sánchez-Cruz; orcid.org/0000-0003-2707-3966; E-mail: norberto.sc90@gmail.com

ACKNOWLEDGEMENTS

NS-C is thankful to *Consejo Nacional de Ciencia y Tecnología* (CONACyT), Mexico, for the granted scholarship number 335997. We thank *Dirección General de Cómputo y de Tecnologías de Información y Comunicación* (DGTIC), UNAM that provided computational resources to use Miztli supercomputer with the project LANCAD-UNAM-DGTIC-335. We also thank *Consejo Nacional de Ciencia y Tecnología* (CONACyT), Mexico, grant 282785.

ABBREVIATIONS USED

BA, balanced accuracy; BET, bromodomain and external terminal protein; BRD, bromodomain; CHR, chromatin remodeler; Da, Dalton; Da, Dalton; DM, distance-to-model; DNA, deoxyribonucleic acid; DNMT, DNA methyltransferase; EC50, half maximal effective concentration; ECFP, extended connectivity fingerprints; FFNN, feed-forward neural network; FN, false negatives; FP, false positives; GBT, gradient boosting trees; HAT, histone acetyltransferase; HDAC, histone deacetylase; HDM, histone demethylase; HMT, histone methyltransferase; IC50, half maximal inhibitory concentration; Kd, dissociation constant; Ki, inhibition constant; KIN, kinase; k-NN, k-nearest neighbors; MACCS, molecular access system; MCC, Mathews correlation coefficient; MODI, modelability index; nM, nanomolar; NPV, negative predictive value; PPB, polypharmacology browser; PPV, positive predictive value; RF, random forest; SAR, structure-activity relationships; SD, standard deviation; SEA, similarity ensemble approach; SMILES, simplified molecular input line entry system; SVM, support vector machines; TN, true negatives; TNR, true negative rate; TP, true positives; TPR, true positive rate; µM, micromolar.

REFERENCES

(1) Waddington, C. H. The Epigenotype. *Int. J. Epidemiol.* **2012**, *41* (1), 10–13. https://doi.org/10.1093/ije/dyr184.

(2) Wu, C. -t. Genes, Genetics, and Epigenetics: A Correspondence. *Science (80).* **2001**, *293* (5532), 1103–1105. https://doi.org/10.1126/science.293.5532.1103.

(3) Ganesan, A.; Arimondo, P. B.; Rots, M. G.; Jeronimo, C.; Berdasco, M. The Timeline of Epigenetic Drug Discovery: From Reality to Dreams. *Clin. Epigenetics* **2019**, *11* (1), 1–17. https://doi.org/10.1186/s13148-019-0776-0.

(4) Biswas, S.; Rao, C. M. Epigenetic Tools (The Writers, The Readers and The Erasers) and Their Implications in Cancer Therapy. *Eur. J. Pharmacol.* **2018**, *837* (June), 8–24. https://doi.org/10.1016/j.ejphar.2018.08.021.

(5) Burgess, R. J.; Zhang, Z. Histone Chaperones in Nucleosome Assembly and Human Disease. *Nat. Struct. Mol. Biol.* **2013**, *20* (1), 14–22. https://doi.org/10.1038/nsmb.2461.

(6) Teif, V. B.; Rippe, K. Predicting Nucleosome Positions on the DNA: Combining Intrinsic Sequence Preferences and Remodeler Activities. *Nucleic Acids Res.* **2009**, *37* (17), 5641–5655. https://doi.org/10.1093/nar/gkp610.

(7) Tyagi, M.; Imam, N.; Verma, K.; Patel, A. K. Chromatin Remodelers: We Are the Drivers!! *Nucleus* **2016**, *7* (4), 388–404. https://doi.org/10.1080/19491034.2016.1211217.

(8) Mayran, A.; Drouin, J. Pioneer Transcription Factors Shape the Epigenetic Landscape. *J. Biol. Chem.* **2018**, *293* (36), 13795–13804. https://doi.org/10.1074/jbc.R117.001232.

(9) Esteller, M. Epigenetics in Cancer. *N. Engl. J. Med.* **2008**, *358* (11), 1148–1159. https://doi.org/10.1056/NEJMra072067.

(10) Küçükali, C. İ.; Kürtüncü, M.; Çoban, A.; Çebi, M.; Tüzün, E. Epigenetics of Multiple Sclerosis: An Updated Review. *NeuroMolecular Med.* **2015**, *17* (2), 83–96. https://doi.org/10.1007/s12017-014-8298-6.

(11) Januar, V.; Saffery, R.; Ryan, J. Epigenetics and Depressive Disorders: A Review of Current Progress and Future Directions. *Int. J. Epidemiol.* **2015**, *44* (4), 1364–1387. https://doi.org/10.1093/ije/dyu273.

(12) Brindisi, M.; Saraswati, A. P.; Brogi, S.; Gemma, S.; Butini, S.; Campiani, G. Old but Gold: Tracking the New Guise of Histone Deacetylase 6 (HDAC6) Enzyme as a Biomarker and Therapeutic Target in Rare Diseases. *J. Med. Chem.* **2020**, *63* (1), 23–39. https://doi.org/10.1021/acs.jmedchem.9b00924.

(13) de Lera, A. R.; Ganesan, A. Two-Hit Wonders: The Expanding Universe of Multitargeting Epigenetic Agents. *Curr. Opin. Chem. Biol.* **2020**, *57*, 135–154. https://doi.org/10.1016/j.cbpa.2020.05.009.

(14) Huang, Z.; Jiang, H.; Liu, X.; Chen, Y.; Wong, J.; Wang, Q.; Huang, W.; Shi, T.; Zhang, J. HEMD: An Integrated Tool of Human Epigenetic Enzymes and Chemical Modulators for Therapeutics. *PLoS One* **2012**, *7* (6), e39917. https://doi.org/10.1371/journal.pone.0039917.

(15) Loharch, S.; Bhutani, I.; Jain, K.; Gupta, P.; Sahoo, D. K.; Parkesh, R. EpiDBase: A Manually Curated Database for Small Molecule Modulators of Epigenetic Landscape. *Database* **2015**, *2015*. https://doi.org/10.1093/database/bav013.

(16) Medvedeva, Y. A.; Lennartsson, A.; Ehsani, R.; Kulakovskiy, I. V.; Vorontsov, I. E.; Panahandeh, P.; Khimulya, G.; Kasukawa, T.; Drabløs, F. EpiFactors: A Comprehensive Database of Human Epigenetic Factors and Complexes. *Database* **2015**, *2015*, bav067. https://doi.org/10.1093/database/bav067.

(17) Singh Nanda, J.; Kumar, R.; Raghava, G. P. S. DbEM: A Database of Epigenetic Modifiers Curated from Cancerous and Normal Genomes. *Sci. Rep.* **2016**, *6* (1), 19340. https://doi.org/10.1038/srep19340.

(18) Naveja, J. J.; Medina-Franco, J. L. Insights from Pharmacological Similarity of Epigenetic Targets in Epipolypharmacology. *Drug Discov. Today* **2018**, *23* (1), 141–

150. https://doi.org/10.1016/j.drudis.2017.10.006.

(19) Sessions, Z.; Sánchez-Cruz, N.; Prieto-Martínez, F. D.; Alves, V. M.; Santos, H. P.; Muratov, E.; Tropsha, A.; Medina-Franco, J. L. Recent Progress on Cheminformatics Approaches to Epigenetic Drug Discovery. *Drug Discov. Today* **2020**, *25* (12), 2268–2276. https://doi.org/10.1016/j.drudis.2020.09.021.

(20) Mendez, D.; Gaulton, A.; Bento, A. P.; Chambers, J.; De Veij, M.; Félix, E.; Magariños, M. P.; Mosquera, J. F.; Mutowo, P.; Nowotka, M.; Gordillo-Marañón, M.; Hunter, F.; Junco, L.; Mugumbate, G.; Rodriguez-Lopez, M.; Atkinson, F.; Bosc, N.; Radoux, C. J.; Segura-Cabrera, A.; Hersey, A.; Leach, A. R. ChEMBL: Towards Direct Deposition of Bioassay Data. *Nucleic Acids Res.* **2019**, *47* (D1), D930–D940. https://doi.org/10.1093/nar/gky1075.

(21) Wang, Y.; Zhang, S.; Li, F.; Zhou, Y.; Zhang, Y.; Wang, Z.; Zhang, R.; Zhu, J.; Ren, Y.; Tan, Y.; Qin, C.; Li, Y.; Li, X.; Chen, Y.; Zhu, F. Therapeutic Target Database 2020: Enriched Resource for Facilitating Research and Early Development of Targeted Therapeutics. *Nucleic Acids Res.* **2020**, *48* (D1), D1031–D1041. https://doi.org/10.1093/nar/gkz981.

(22) Naveja, J. J.; Oviedo-Osornio, C. I.; Medina-Franco, J. L. Computational Methods for Epigenetic Drug Discovery: A Focus on Activity Landscape Modeling. In *Advances in Protein Chemistry and Structural Biology*; 2018; Vol. 113, pp 65–83. https://doi.org/10.1016/bs.apcsb.2018.01.001.

(23) Keiser, M. J.; Roth, B. L.; Armbruster, B. N.; Ernsberger, P.; Irwin, J. J.; Shoichet, B. K. Relating Protein Pharmacology by Ligand Chemistry. *Nat. Biotechnol.* **2007**, *25* (2), 197–206. https://doi.org/10.1038/nbt1284.

(24) Liu, X.; Vogt, I.; Haque, T.; Campillos, M. HitPick: A Web Server for Hit Identification and Target Prediction of Chemical Screenings. *Bioinformatics* **2013**, *29* (15), 1910–1912. https://doi.org/10.1093/bioinformatics/btt303.

(25) Hamad, S.; Adornetto, G.; Naveja, J. J.; Chavan Ravindranath, A.; Raffler, J.; Campillos, M. HitPickV2: A Web Server to Predict Targets of Chemical Compounds. *Bioinformatics* **2019**, *35* (7), 1239–1240. https://doi.org/10.1093/bioinformatics/bty759.

(26) Awale, M.; Reymond, J.-L. The Polypharmacology Browser: A Web-Based Multi-Fingerprint Target Prediction Tool Using ChEMBL Bioactivity Data. *J. Cheminform.* **2017**, *9* (1), 11. https://doi.org/10.1186/s13321-017-0199-x.

(27) Awale, M.; Reymond, J.-L. Polypharmacology Browser PPB2: Target Prediction Combining Nearest Neighbors with Machine Learning. *J. Chem. Inf. Model.* **2019**, *59* (1), 10–17. https://doi.org/10.1021/acs.jcim.8b00524.

(28) Wang, L.; Ma, C.; Wipf, P.; Liu, H.; Su, W.; Xie, X.-Q. TargetHunter: An In Silico Target Identification Tool for Predicting Therapeutic Potential of Small Organic Molecules Based on Chemogenomic Database. *AAPS J.* **2013**, *15* (2), 395–406. https://doi.org/10.1208/s12248-012-9449-z.

(29) Gfeller, D.; Grosdidier, A.; Wirth, M.; Daina, A.; Michielin, O.; Zoete, V. SwissTargetPrediction: A Web Server for Target Prediction of Bioactive Small Molecules. *Nucleic Acids Res.* **2014**, *42* (W1), 32–38. https://doi.org/10.1093/nar/gku293.

(30) Daina, A.; Michielin, O.; Zoete, V. SwissTargetPrediction: Updated Data and New Features for Efficient Prediction of Protein Targets of Small Molecules. *Nucleic Acids Res.* **2019**, *47* (W1), W357–W364. https://doi.org/10.1093/nar/gkz382.

(31) Zdrazil, B.; Richter, L.; Brown, N.; Guha, R. Moving Targets in Drug Discovery. *Sci. Rep.* **2020**, *10* (1), 20213. https://doi.org/10.1038/s41598-020-77033-x.

(32) Oprea, T. I.; Bologa, C. G.; Brunak, S.; Campbell, A.; Gan, G. N.; Gaulton, A.; Gomez, S. M.; Guha, R.; Hersey, A.; Holmes, J.; Jadhav, A.; Jensen, L. J.; Johnson, G. L.; Karlson, A.; Leach, A. R.; Ma'ayan, A.; Malovannaya, A.; Mani, S.; Mathias, S. L.; McManus, M. T.; Meehan, T. F.; von Mering, C.; Muthas, D.; Nguyen, D.-T.;

Overington, J. P.; Papadatos, G.; Qin, J.; Reich, C.; Roth, B. L.; Schürer, S. C.; Simeonov, A.; Sklar, L. A.; Southall, N.; Tomita, S.; Tudose, I.; Ursu, O.; Vidović, D.; Waller, A.; Westergaard, D.; Yang, J. J.; Zahoránszky-Köhalmi, G. Unexplored Therapeutic Opportunities in the Human Genome. *Nat. Rev. Drug Discov.* **2018**, *17* (5), 317–332. https://doi.org/10.1038/nrd.2018.14.

(33) Lo, Y.-C.; Rensi, S. E.; Torng, W.; Altman, R. B. Machine Learning in Chemoinformatics and Drug Discovery. *Drug Discov. Today* **2018**, *23* (8), 1538–1546. https://doi.org/10.1016/j.drudis.2018.05.010.

(34) Miljković, F.; Rodríguez-Pérez, R.; Bajorath, J. Machine Learning Models for Accurate Prediction of Kinase Inhibitors with Different Binding Modes. *J. Med. Chem.* **2020**, *63* (16), 8738–8748. https://doi.org/10.1021/acs.jmedchem.9b00867.

(35) Li, H.; Sze, K.; Lu, G.; Ballester, P. J. Machine-learning Scoring Functions for Structure-based Drug Lead Optimization. *WIREs Comput. Mol. Sci.* **2020**, *10* (5), 1–20. https://doi.org/10.1002/wcms.1465.

(36) Gaulton, A.; Hersey, A.; Nowotka, M.; Bento, A. P.; Chambers, J.; Mendez, D.; Mutowo, P.; Atkinson, F.; Bellis, L. J.; Cibrián-Uhalte, E.; Davies, M.; Dedman, N.; Karlsson, A.; Magariños, M. P.; Overington, J. P.; Papadatos, G.; Smit, I.; Leach, A. R. The ChEMBL Database in 2017. *Nucleic Acids Res.* **2017**, *45* (D1), D945–D954. https://doi.org/10.1093/nar/gkw1074.

(37) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50* (5), 742–754. https://doi.org/10.1021/ci100050t.

(38) Mayr, A.; Klambauer, G.; Unterthiner, T.; Steijaert, M.; Wegner, J. K.; Ceulemans, H.; Clevert, D.-A.; Hochreiter, S. Large-Scale Comparison of Machine Learning Methods for Drug Target Prediction on ChEMBL. *Chem. Sci.* **2018**, *9* (24), 5441–5451. https://doi.org/10.1039/C8SC00148K.

(39) Sirous, H.; Campiani, G.; Brogi, S.; Calderone, V.; Chemi, G. Computer-Driven

Development of an in Silico Tool for Finding Selective Histone Deacetylase 1 Inhibitors. *Molecules* **2020**, *25* (8), 1952. https://doi.org/10.3390/molecules25081952.

(40) Li, S.; Ding, Y.; Chen, M.; Chen, Y.; Kirchmair, J.; Zhu, Z.; Wu, S.; Xia, J. HDAC3i-Finder: A Machine Learning-based Computational Tool to Screen for HDAC3 Inhibitors. *Mol. Inform.* **2020**, minf.202000105. https://doi.org/10.1002/minf.202000105.

(41) Norinder, U.; Naveja, J. J.; López-López, E.; Mucs, D.; Medina-Franco, J. L. Conformal Prediction of HDAC Inhibitors. *SAR QSAR Environ. Res.* **2019**, *30* (4), 265–277. https://doi.org/10.1080/1062936X.2019.1591503.

(42) Speck-Planche, A.; Scotti, M. T. BET Bromodomain Inhibitors: Fragment-Based in Silico Design Using Multi-Target QSAR Models. *Mol. Divers.* **2019**, *23* (3), 555–572. https://doi.org/10.1007/s11030-018-9890-8.

(43) Kim, S.; Chen, J.; Cheng, T.; Gindulyte, A.; He, J.; He, S.; Li, Q.; Shoemaker, B. A.; Thiessen, P. A.; Yu, B.; Zaslavsky, L.; Zhang, J.; Bolton, E. E. PubChem 2019 Update: Improved Access to Chemical Data. *Nucleic Acids Res.* **2019**, *47* (D1), D1102–D1109. https://doi.org/10.1093/nar/gky1033.

(44) Golbraikh, A.; Muratov, E.; Fourches, D.; Tropsha, A. Data Set Modelability by QSAR. *J. Chem. Inf. Model.* **2014**, *54* (1), 1–4. https://doi.org/10.1021/ci400572x.

(45) Altman, N. S. An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression. *Am. Stat.* **1992**, *46* (3), 175–185. https://doi.org/10.1080/00031305.1992.10475879.

(46) Tin Kam Ho. The Random Subspace Method for Constructing Decision Forests. *IEEE Trans. Pattern Anal. Mach. Intell.* **1998**, *20* (8), 832–844. https://doi.org/10.1109/34.709601.

(47) Friedman, J. Greedy Function Approximation : A Gradient Boosting Machine Author (s): Jerome H . Friedman Source : The Annals of Statistics , Vol . 29 , No . 5 (Oct ., 2001), Pp . 1189-1232 Published by : Institute of Mathematical Statistics Stable URL :

Http://Www. *Ann. Stat.* **2001**, *29* (5), 1189–1232.

(48) Cortes, C.; Vapnik, V. Support-Vector Networks. *Mach. Learn.* **1995**, *20* (3), 273–297. https://doi.org/10.1007/BF00994018.

(49) Hopfield, J. J. Neural Networks and Physical Systems with Emergent Collective Computational Abilities. *Proc. Natl. Acad. Sci.* **1982**, *79* (8), 2554–2558. https://doi.org/10.1073/pnas.79.8.2554.

(50) 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* (6), 1273–1280. https://doi.org/10.1021/ci010132r.

(51) Mansouri, K.; Abdelaziz, A.; Rybacka, A.; Roncaglioni, A.; Tropsha, A.; Varnek, A.; Zakharov, A.; Worth, A.; Richard, A. M.; Grulke, C. M.; Trisciuzzi, D.; Fourches, D.; Horvath, D.; Benfenati, E.; Muratov, E.; Wedebye, E. B.; Grisoni, F.; Mangiatordi, G. F.; Incisivo, G. M.; Hong, H.; Ng, H. W.; Tetko, I. V.; Balabin, I.; Kancherla, J.; Shen, J.; Burton, J.; Nicklaus, M.; Cassotti, M.; Nikolov, N. G.; Nicolotti, O.; Andersson, P. L.; Zang, Q.; Politi, R.; Beger, R. D.; Todeschini, R.; Huang, R.; Farag, S.; Rosenberg, S. A.; Slavov, S.; Hu, X.; Judson, R. S. CERAPP: Collaborative Estrogen Receptor Activity Prediction Project. *Environ. Health Perspect.* **2016**, *124* (7), 1023–1033. https://doi.org/10.1289/ehp.1510267.

(52) Alves, V. M.; Golbraikh, A.; Capuzzi, S. J.; Liu, K.; Lam, W. I.; Korn, D. R.; Pozefsky, D.; Andrade, C. H.; Muratov, E. N.; Tropsha, A. Multi-Descriptor Read Across (MuDRA): A Simple and Transparent Approach for Developing Accurate Quantitative Structure–Activity Relationship Models. *J. Chem. Inf. Model.* **2018**, *58* (6), 1214–1223. https://doi.org/10.1021/acs.jcim.8b00124.

(53) MEDINA-FRANCO, J. L.; Sánchez-Cruz, N. {Supporting Information For. https://doi.org/10.6084/m9.figshare.13519580.

(54) Sánchez-Cruz, N.; Pilón-Jiménez, B. A.; Medina-Franco, J. L. Functional Group and

Diversity Analysis of BIOFACQUIM: A Mexican Natural Product Database. *F1000Research* **2020**, *8*, 2071. https://doi.org/10.12688/f1000research.21540.2.

(55) Chávez-Hernández, A. L.; Sánchez-Cruz, N.; Medina-Franco, J. L. A Fragment Library of Natural Products and Its Comparative Chemoinformatic Characterization. *Mol. Inform.* **2020**, *39* (11), minf.202000050. https://doi.org/10.1002/minf.202000050.

(56) Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Model.* **1988**, *28* (1), 31–36. https://doi.org/10.1021/ci00057a005.

(57) 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 {P}ython. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.

(58) Tetko, I. V.; Sushko, I.; Pandey, A. K.; Zhu, H.; Tropsha, A.; Papa, E.; Öberg, 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* (9), 1733–1746. https://doi.org/10.1021/ci800151m.

(59) Sushko, I.; Novotarskyi, S.; Körner, R.; Pandey, A. K.; Cherkasov, A.; Li, J.; Gramatica, P.; Hansen, K.; Schroeter, T.; Müller, K.-R.; Xi, L.; Liu, H.; Yao, X.; Öberg, T.; Hormozdiari, F.; Dao, P.; Sahinalp, C.; Todeschini, R.; Polishchuk, P.; Artemenko, A.; Kuz'min, V.; Martin, T. M.; Young, D. M.; Fourches, D.; Muratov, E.; Tropsha, A.; Baskin, I.; Horvath, D.; Marcou, G.; Muller, C.; Varnek, A.; Prokopenko, V. V.; Tetko, I. V. Applicability Domains for Classification Problems: Benchmarking of Distance to Models for Ames Mutagenicity Set. *J. Chem. Inf. Model.* **2010**, *50* (12), 2094–2111. https://doi.org/10.1021/ci100253r.

(60) Wilson, J. E.; Patel, G.; Patel, C.; Brucelle, F.; Huhn, A.; Gardberg, A. S.; Poy, F.;

Cantone, N.; Bommi-Reddy, A.; Sims, R. J.; Cummings, R. T.; Levell, J. R. Discovery of CPI-1612: A Potent, Selective, and Orally Bioavailable EP300/CBP Histone Acetyltransferase Inhibitor. *ACS Med. Chem. Lett.* **2020**, *11* (6), 1324–1329. https://doi.org/10.1021/acsmedchemlett.0c00155.

(61) Chen, J.; Li, Y.; Zhang, J.; Zhang, M.; Wei, A.; Liu, H.; Xie, Z.; Ren, W.; Duan, W.; Zhang, Z.; Shen, A.; Hu, Y. Discovery of Selective HDAC/BRD4 Dual Inhibitors as Epigenetic Probes. *Eur. J. Med. Chem.* **2021**, *209*, 112868. https://doi.org/10.1016/j.ejmech.2020.112868.