Novel Inhibitors of Human Histone Deacetylase (HDAC) Identified by QSAR Modeling of Known Inhibitors, Virtual Screening, and Experimental Validation

Hao Tang,[†,‡] Xiang S. Wang,[†] Xi-Ping Huang,[§] Bryan L. Roth,[§] Kyle V. Butler,[||] Alan P. Kozikowski,[||] Mira Jung,[⊥] and Alexander Tropsha*[,†,‡]

Laboratory for Molecular Modeling, and Carolina Exploratory Center for Cheminformatics Research, Division of Medicinal Chemistry and Natural Products, School of Pharmacy, Biophysics Training Program, Department of Pharmacology, School of Medicine, University of North Carolina, Chapel Hill, North Carolina 27599, Department of Medicinal Chemistry and Pharmacognosy, University of Illinois, 833 South Wood Street, Chicago, Illinois 60612, and Department of Radiation Medicine, Georgetown University Medical Center, Washington, DC 20057

Received October 5, 2008

Inhibitors of histone deacetylases (HDACIs) have emerged as a new class of drugs for the treatment of human cancers and other diseases because of their effects on cell growth, differentiation, and apoptosis. In this study we have developed several quantitative structure−activity relationship (QSAR) models for 59 chemically diverse histone deacetylase class 1 (HDAC1) inhibitors. The variable selection *k* nearest neighbor (*k*NN) and support vector machines (SVM) QSAR modeling approaches using both MolconnZ and MOE chemical descriptors generated from two-dimensional rendering of compounds as chemical graphs have been employed. We have relied on a rigorous model development workflow including the division of the data set into training, test, and external sets and extensive internal and external validation. Highly predictive QSAR models were generated with leave-one-out cross-validated (LOO-CV) q^2 and external R^2 values as high as 0.80 and 0.87, respectively, using the *k*NN/MolconnZ approach and 0.93 and 0.87, respectively, using the SVM/MolconnZ approach. All validated QSAR models were employed concurrently for virtual screening (VS) of an in-house compound collection including 9.5 million molecules compiled from the ZINC7.0 database, the World Drug Index (WDI) database, the ASINEX Synergy libraries, and other commercial databases. VS resulted in 45 structurally unique consensus hits that were considered novel putative HDAC1 inhibitors. These computational hits had several novel structural features that were not present in the original data set. Four computational hits with novel scaffolds were tested experimentally, and three of them were confirmed active against HDAC1, with IC_{50} values for the most active compound of 1.00 μM. The fourth compound was later identified to be a selective inhibitor of HDAC6, a Class II HDAC. Moreover, two of the confirmed hits are marketed drugs, which could potentially facilitate their further development as anticancer agents. This study illustrates the power of the combined QSAR-VS method as a general approach for the effective identification of structurally novel bioactive compounds.

INTRODUCTION

The dynamic posttranslational modification of nucleosomal histones plays a critical role in transcriptional regulation. Hyperacetylation of core histones results in transcriptional activation, while hypoacetylation leads to expression repression.[1] This kind of regulation is considered to be the critical step in normal cell differentiation and chromatin condensation and is believed to be regulated by the balance between two groups of enzymes: histone deacetylases (HDACs) and histone acetyltransfereases (HATs).[2,3] Inhibition of HDACs represents a novel approach to interfere with cell cycle regulation; therefore, it has a great therapeutic potential in the treatment of diseases of aberrant cellular proliferation.[4] It has been reported that hyperacetylation of histones and nonhistone proteins induced by small molecule HDACs inhibitors (HDACI) leads to cell growth arrest, cellular differentiation, or apoptosis of malignant cells.[5−8] For these reasons, HDACI has become a promising class of chemical agents for the treatment of cancer and other diseases associated with uncontrolled cell proliferation.

To date, a number of structurally distinct classes of HDACI have been reported, including hydroxamates, cyclic peptides, aliphatic acids, and benzamides.[9,10] The natural product Trichostatin A (TSA)[11] is the most well-known member of the hydroxamates class; this compound is considered to be a mimetic of the natural substrate, that is, histone acetyl lysine side chain. Extensive structure−activity relationship (SAR) studies have been conducted for TSA and TSA-like compounds resulting in several potent HDACs inhibitors.[12−15] A TSA analog suberoylanilide hydroxamic

* To whom correspondence should be addressed. Address: CB #7360, Beard Hall, School of Pharmacy, University of North Carolina, Chapel Hill, NC 27599-7360. Phone: 919-966-2955. Fax: 919-966-0204. E-mail: alex_tropsha@unc.edu.

[†] School of Pharmacy, University of North Carolina at Chapel Hill.

[‡] Biophysics Training Program, University of North Carolina at Chapel Hill.

[§] School of Medicine, University of North Carolina at Chapel Hill.

[||] Department of Medicinal Chemistry and Pharmacognosy, University of Illinois at Chicago.

[⊥] Department of Radiation Medicine, Georgetown University Medical Center.

10.1021/ci800366f CCC: $40.75 © 2009 American Chemical Society
Published on Web 01/30/2009

Table 1. Summary of Previous QSAR Studies of HDACs Inhibitors

group	year	HDAC types	data set size	chemical class	model type	model validation
Chen et al.[78]	2008	HDAC1	training set (30), test set (25)	mostly hydroxamates	3D chemical-feature-based QSAR pharmacophore model by Hypogen	training set $R^2 = 0.924$, test set $R^2 = 0.896$
Kozikowski et al.[79]	2008	HDAC1−3, HDAC8; HDAC6, HDAC10;	20−23	biphenyl or phenylthiazoles analogues bearing hydroxamates or mercaptoacetamides	linear regression	best model with training set $R^2 = 0.943$
Ragno et al.[80]	2008	Class II HDAC, Maize HDAC1-A, HDAC1-B	25	(aryloxopropenyl) pyrrolyl hydroxamates	GRIND[a] 3-D QSAR, PLS[b]	R^2/q^2 values of 0.96/0.81 and 0.98/0.85 for HDAC1-B and HDAC1-A randomized variable selection
Juvale et al.[13]	2006	HDAC1	training set (40), test set (17)	hydroxamates	CoMFA[c], CoMSIA[d]	best CoMFA training set $R^2 = 0.910$, LOO-CV $R^2 = 0.502$, test set $R^2 = 0.327$ best CoMSIA training set $R^2 = 0.987$, LOO-CV $R^2 = 0.534$, test set $R^2 = 0.464$ *Y* randomization
Ragno et al.[23]	2006	Maize HDAC2	101		GRID[e]/GOLPE[f], PLS	R^2, cross validated q^2, and cross-validated SDEP (RMSE) values of 0.94, 0.83, and 0.41
Guo et al.[81]	2005	HDAC1	training set (25), test set (4)	indole amide analogues	CoMFA, CoMSIA	CoMFA $R^2 = 0.982$, LOO-CV $R^2 = 0.601$ CoMSIA $R^2 = 0.954$, LOO-CV $R^2 = 0.598$
Wagh et al.[82]	2006	HDACs	training set (39), test (17)	hydroxamic acid analogues	3D-QSAR by GFA[g]	training set $q^2 = 0.712$, $R^2 = 0.585$
Wang et al.[71]	2004	PC-3 cell line	19	TSA and SAHA like hydroxamic acid	stepwise multiple linear regression	LOO-CV $R^2 = 0.870$
Xie et al.[83]	2004	HDAC1, 4, 6 maize HDAC2	124	hydroxamic acid, short chain fatty acid, cyclic tetrapeptides, cyclic peptides without the AOE moiety, benzamides	PLS, binary classification	best model with R^2 of 0.76, and LOO-CV R^2 of 0.73 LOO-CV accuracy for binary model is 0.92

[a] GRIND: grid-independent descriptors. [b] PLS: partial least squares regression. [c] CoMFA: comparative molecular field analysis. [d] CoMSIA: comparative molecular similarity indices analysis. [e] GRID: GRID force field. [f] GOLPE: generating of optimal linear pls estimations. [g] GFA: genetic function approximation.

acid (SAHA)[16] was recently approved by the FDA for the treatment of cutaneous T cell lymphoma (CTCL), stimulating further investigations of HDACI in the treatment of various diseases.[17]

HDACs have been classified so far into four classes (Classes I−IV) depending on the sequence identity and domain organization. Among the Class I HDACs, HDACs 1, 2, and 8 are primarily found in the nucleus, whereas HDAC 3 is found in the nucleus, cytoplasm, and the membrane. In comparison, Class II HDACs subdivided into IIa (HDAC 4, 5, 7, 9) and IIb (HDAC 6, 10) are able to shuttle in and out of the nucleus depending on different signals. Class III HDACs include the SIRTs (sirtuins) or Sir2-related proteins; they are NAD-dependent[18] and are insensitive to TSA or other hydroxamate inhibitors. Class IV comprises HDAC 11, based on a phylogenetic analysis, and is the least characterized to date.[19] It has been considered important in recent years to develop class/subtype selective HDACI. Considering the number of pathways in which HDACs are involved, the HDACI that act exclusively on Class I or Class II enzymes are viewed as likely candidates as anticancer therapeutic agents.

The crystal structures of the histone deacetylase like protein (HDLP) both in the apo form and in complexes with TSA and SAHA were first published by Finnin et al. in 1999.[20] Five years later, Somoza's group and Di Marco's group both solved the X-ray structures of another class I histone deacetylase, histone deacetylase subtype 8 (HDAC8) in complex with several small molecule HDACI.[21,22] The crystallographic structures revealed that both HDLP and HDAC8 contain a conserved tunnel-like binding pocket with the polar active site at the bottom. In the X-ray structure of the HDLP/TSA complex, the long aliphatic chain of TSA (linker domain) spans the whole length of the tunnel-like pocket and the hydroxamic acid moiety interacts with the polar residues at the bottom of the pocket. The chelating atoms of hydroxamic acid coordinate zinc ion in a bidentate fashion and form hydrogen bonds with the His-Asp diad.[10,20,22] At the other end of the aliphatic chain, the aromatic group of TSA (surface recognition domain) interacts with the hydrophobic rim of the pocket.[10] Thus, SAR studies have been typically focused on three regions of HDACI: the metal binding group, the linker domain, and the surface recognition domain.[15]

Because of their potential clinical importance HDACI have been the subject of several quantitative structure−activity relationship (QSAR) modeling studies.[13,14,23] The results of these studies are summarized in Table 1. Most of the studies focused on a series of hydroxamates and employed 3D QSAR modeling methods. This preference was partially because a number of HDACs crystallographic structures have been solved in recent years and thus could be used for structural alignment of inhibitors to enable 3D QSAR modeling. The size of HDACI data sets varied among different reports, ranging from 19 to 124. The best reported models were characterized by leave-one-out cross-validation (LOO-CV) R^2 of 0.870 and R^2 of 0.987. For the test set, the R^2 was as high as 0.896. It should be pointed out that none of these earlier studies had employed an independent data

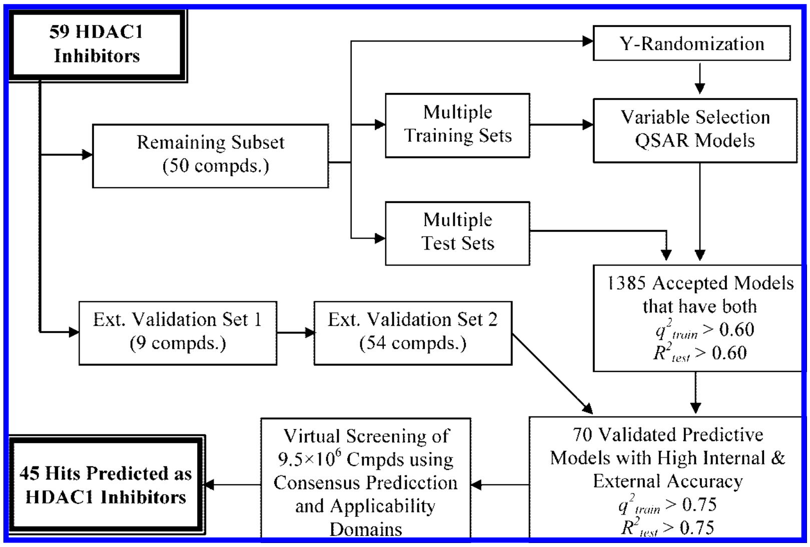

59 HDAC1 Inhibitors
Y-Randomization
Multiple Training Sets
Variable Selection QSAR Models
Remaining Subset (50 compds.)
Multiple Test Sets
1385 Accepted Models that have both $q^2_{train} > 0.60$ $R^2_{test} > 0.60$
Ext. Validation Set 1 (9 compds.)
Ext. Validation Set 2 (54 compds.)
45 Hits Predicted as HDAC1 Inhibitors
Virtual Screening of 9.5×10^6 Cmpds using Consensus Predicction and Applicability Domains
70 Validated Predictive Models with High Internal & External Accuracy $q^2_{train} > 0.75$ $R^2_{test} > 0.75$

Figure 1. Workflow of QSAR model building, validation, and virtual screening as applied to HDAC1 inhibitors. The specific data for *k*NN/MolconnZ modeling are used for illustration.

set for model validation, and none used models for virtual screening of chemical libraries to identify novel hits.

In the present study of HDACI, we have applied the modeling strategy that has been under development in our laboratory for several years.[24] The important feature of our approach is that it combines validated QSAR modeling of historic data and virtual screening of available chemical libraries for the identification of novel active compounds, as illustrated in Figure 1. We have used experimental data for 59 histone deacetylase subtype 1 (HDAC1) inhibitors that were generated in one of our laboratories. All of the compounds in the data set were hydroxamates but incorporated many novel chemical modifications in the three major domains, that is, the hydroxamic acid, the linker domain and surface recognition domain. Our studies resulted in externally predictive QSAR models of HDAC1 inhibitors. Furthermore, the application of these models to virtual screening of a large (~9.5 mln) collection of commercially available chemical compounds identified several computational hits, and three of them were confirmed experimentally as novel active HDAC1 inhibitors.

MATERIALS AND METHODS

Data Sets for Model Building and Validation. Fifty-nine compounds with known HDAC1 inhibition activities were employed for the QSAR study (see Table 2). All data were generated in the laboratories of Dr. Kozikowski (chemistry) and Dr. Jung (biology), and most of them were reported earlier.[25−30] The data for eight compounds, BC-2-87, BC-3-63, BC-3-70, BC-3-94, BC-4-93, BC-6-30, BC-6-33, and BC-6-34, are reported here for the first time. The half-maximal (50%) inhibitory concentration of a substance (IC_{50}) was measured on HDAC1 from HeLa cell extracts. It was then converted to the pIC_{50} scale ($-\log IC_{50}$), in which higher values indicate exponentially greater potency.

Two independent external validation sets of different nature were employed in the later phase of our modeling workflow (cf. Figure 1): one included 9 HDAC1 inhibitors randomly selected from the original set of 59 compounds and another comprised 54 diverse HDAC1 inhibitors collected from two general reviews on HDACIs.[10,12] These external sets have covered most chemical classes of known HDACI.[15,31−35] Other compounds discussed in the reviews were excluded either because their HDAC1 binding affinity data were not reported or they were duplicates of compounds included in the modeling set. The observed pIC_{50} values of 54 compounds ranged between 4.0 and 8.0, which is similar to the activity range observed for the 50 compounds used for model development.

Libraries for Virtual Screening. The virtual screening was performed on our in-house collection of ~9 500 000 molecules, including the ZINC7.0 database of ~6 500 000 compounds,[36] the World Drug Index (WDI) database of ~59 000 compounds,[37] the ASINEX Synergy libraries (2006.04) of ~11 000 compounds,[38] the InterBioScreen screening libraries (2007.03) of ~400 000 compounds,[39] the Chemizon Progenitor databases (2006 v1.1) of ~3300 compounds,[40] and several other commercial databases. None of the compounds present in the modeling set were found in the screening libraries. MolconnZ4.09 (MZ4.09) descriptors were calculated for each compound in the databases and linearly normalized based on the maximum and minimum values of each descriptor type in the modeling data set of 59 HDAC1 inhibitors.

Generation of MolconnZ Descriptors. The MolconnZ4.05 (MZ4.05) software[41] affords the computation of a wide range of topological indices (descriptors) of molecular structure such as simple and valence path, cluster, path/cluster, and chain molecular connectivity indices, kappa molecular shape indices, topological and electrotopological state indices, differential connectivity indices, graph's radius and diameter, Wiener and Platt indices, Shannon and Bonchev−Trinajsti, information indices, counts of different vertices, and counts of paths and edges between different kinds of vertices.[42−49] Overall, MZ4.05 produces more than 400 different descriptors. In this study, only 262 chemically relevant descriptors were eventually used after removal of those with zero value or zero variance. MZ4.05 descriptors were range-scaled because the absolute values of individual types could differ by orders of magnitude.[50] Therefore, range scaling prevents undesirable overweighting of descriptors

Table 2. Structure and pIC_{50} of 59 HDAC1 Inhibitors Used for Model Building

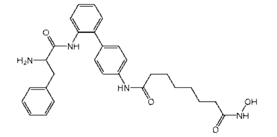
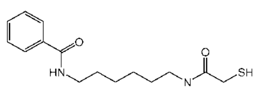
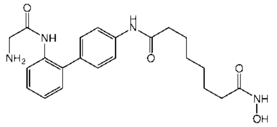
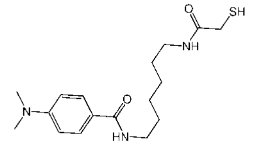
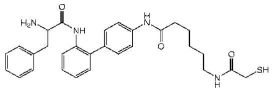
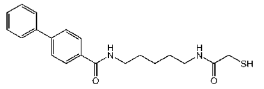
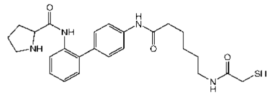
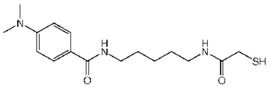
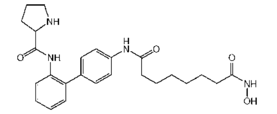
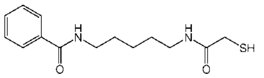
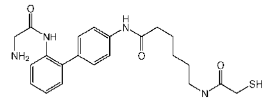
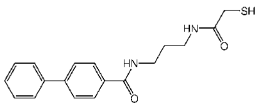
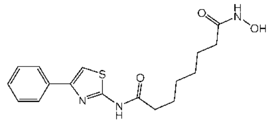
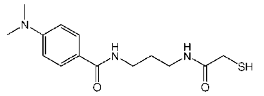
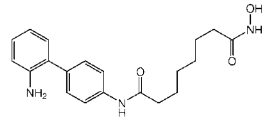
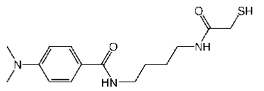
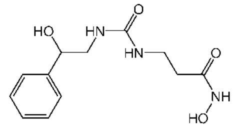
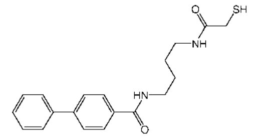
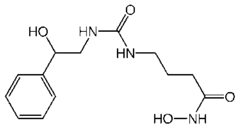
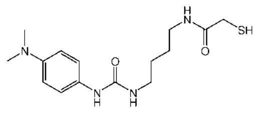
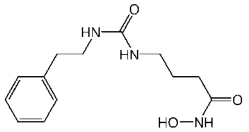
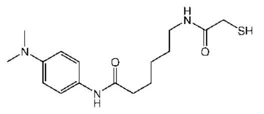
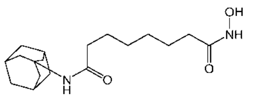
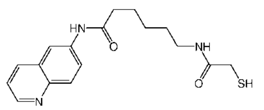
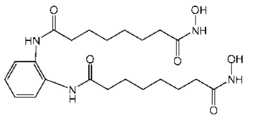
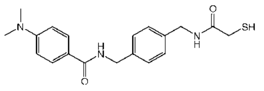
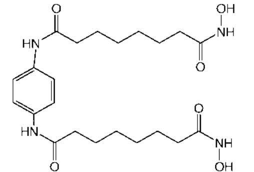
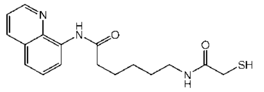
Mol.	Comp. ID	pIC_{50}	Mol.	Comp. ID	pIC_{50}
	AG-biph-08	7.10		BC-4-44	6.05
	AG-biph-15	6.70		BC-4-46	6.35
	AG-biph-35	5.15		BC-4-54	6.26
	AG-biph-36	5.82		BC-4-55	6.70
	AG-biph-38	7.10		BC-4-56	5.96
	AG-biph-40	5.52		BC-4-81	6.10
	Ag-THIA-01	7.82		BC-4-84	6.10
	Ag-b-57	7.10		BC-4-86	5.33
	BC-1-30-2	4.00		BC-4-87	4.89
	BC-2-45	4.00		BC-4-93	5.77
	BC-2-48	4.00		BC-4-96	6.22
	BC-2-83	6.00		BC-5-44	6.05
	BC-2-84	6.00		BC-6-12	6.70
	BC-2-87	6.00		BC-6-25	7.36

Table 2. Continued

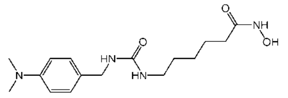
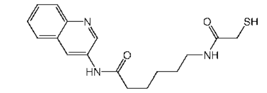
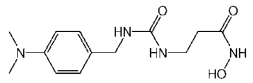
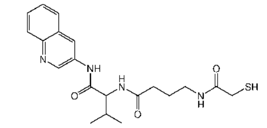
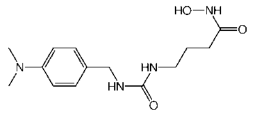
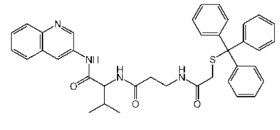
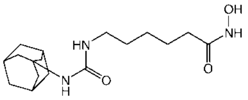
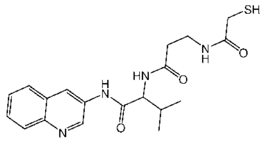
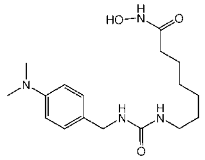
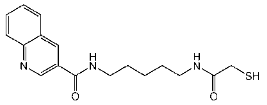
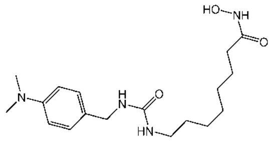
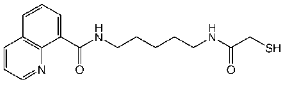
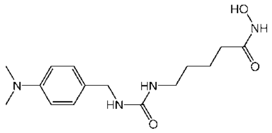
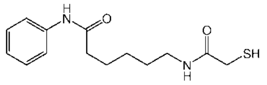
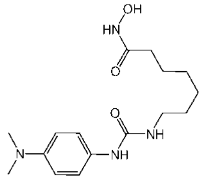
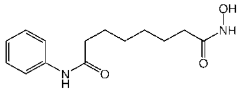
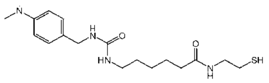
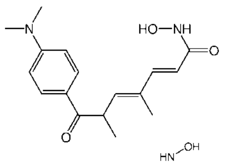
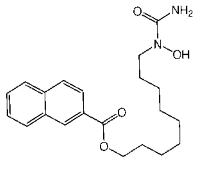
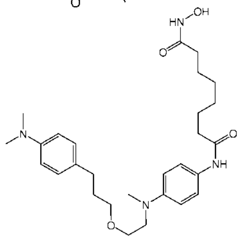
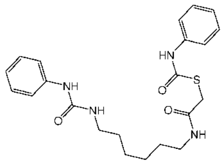
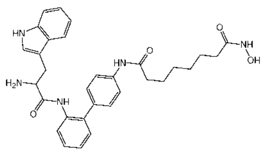
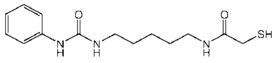
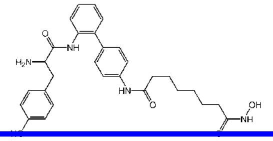
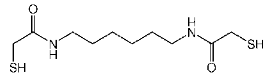
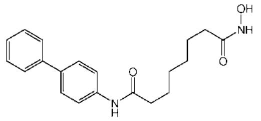
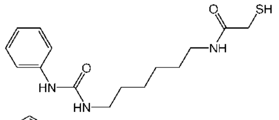
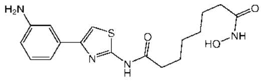
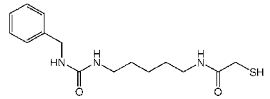
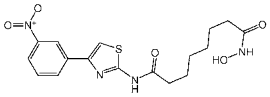
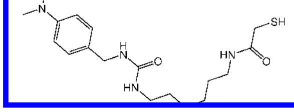
Mol.	Comp. ID	pIC_{50}	Mol.	Comp. ID	pIC_{50}
	BC-3-10	6.32		BC-6-26	7.30
	BC-3-14	5.00		BC-6-30	5.52
	BC-3-18	4.30		BC-6-33	4.30
	BC-3-22	6.38		BC-6-34	4.60
	BC-3-42	6.30		BC-6-38	6.40
	BC-3-46	6.70		BC-6-40	6.60
	BC-3-5	6.72		BC-6-83	6.52
	BC-3-52	6.87		SAHA	7.10
	BC-3-63	5.00		TSA	8.46
	BC-3-70	5.00		YC-03065	6.52
	BC-3-94	5.00		YChdac044	7.30
	BC-4-14	6.12		YChdac045	7.52

Table 2. Continued

Mol.	Comp. ID	*pIC*$_{50}$	Mol.	Comp. ID	*pIC*$_{50}$
	BC-4-2	5.00		Yc-II-84	7.26
	BC-4-3	6.20		Yc-II-88	8.10
	BC-4-31	6.00		Yc-II-90	7.66
	BC-4-4	5.30			

with high ranges of values in calculating compound similarities as part of QSAR modeling procedure.

Generation of MOE Descriptors. The MOE2006.08 software[51] generates both 2D and 3D descriptors. 2D molecular descriptors include physical properties, subdivided surface areas, atom counts and bond counts, Kier and Hall connectivity and kappa shape indices, adjacency and distance matrix descriptors, pharmacophore feature descriptors, and partial charge descriptors.[43,44,48,52−55] 3D molecular descriptors include potential energy descriptors, surface area, volume and shape descriptors, and conformation-dependent charge descriptors.[56] For model generation, we used 179 MOE descriptors with nonzero value and variance that were range-scaled.

Prior to the model building, we did not carry out correlation analysis on the descriptors in this case. Our previous experience showed that the intercorrelations among individual types in MZ4.05 and MOE descriptors were relatively low. Moreover, the variable selection approaches in our both *k*NN and SVM QSAR methods can minimize its influence on the model quality effectively.

Selection of Training and Test Sets. The data set was subdivided into multiple training/test set pairs using the sphere exclusion method developed in our laboratory.[57,58] By default, fifty different training/test set splits were initially tried using probe sphere radii defined by the minimum and maximum elements, $D_{\min}$ and $D_{\max}$, of the distance matrix *D* between compound-vectors in the descriptor space and forty-two splits were ultimately accepted. The number of compounds in the test set was varied to achieve the largest possible size of the test set, while ensuring that the training set models were still able to accurately predict the binding affinity of the test set compounds.

***k*NN Regression Method.** The *k* nearest neighbor (*k*NN) QSAR method used in this study employs the *k*NN pattern recognition principle[59] and variable selection method. In short, a subset of variables (descriptors) is selected randomly as a hypothetical descriptor pharmacophore (HDP).[60] The HDP is validated by LOO-CV, where each compound is eliminated from the training set and its HDAC1 inhibition activity is predicted as the weighted average of the activities of the *k* most similar molecules (*k* varies from 1 to 5). The weighted molecular similarity is represented by the modified Euclidean distance between compounds in HDP multidimensional space as shown in eqs 1 and 2. Essentially, the neighbor with the smaller distance from a compound is given a higher weight in calculating the predicted activity.

$$w_i = \frac{\mathrm{e}^{-d_i}}{\sum_i \mathrm{e}^{-d_i}} \tag{1}$$

$$\tilde{y} = \sum w_i y_i \tag{2}$$

where d_i is the Euclidean distance between the compound *i* and its *k*th nearest neighbors, w_i is the weight for the *k*th nearest neighbor, y_i is the experimentally measured activity value for the *k*th nearest neighbor, and $\tilde{y}$ is the predicted activity value.

Simulated annealing and Metropolis-like acceptance criteria were used to optimize the variables. Details of the *k*NN method implementation including the description of the simulated annealing procedure used for stochastic sampling of the descriptor space, are given elsewhere.[50] The statistical significance of the models were estimated by the LOO-CV q^2 in the training set, a coefficient of determination $R_0{}^2$ (eq 3) and a linear fit predictive R^2 for both internal and external test sets.

$$q^2(R_0{}^2) = 1 - \frac{\sum (\tilde{y}_i - y_i)^2}{\sum (\bar{y} - y_i)^2} \tag{3}$$

Here y_i and $\tilde{y}_i$ are the observed and predicted activities of compound *i*, respectively, and $\bar{y}$ is the average activity of all compounds. Model acceptability cutoffs were $q^2 > 0.60$ for training set and correlation coefficient $R^2 > 0.60$ for the internal test set.[57] All models that satisfied both criteria were applied to external validation sets.

SVM Regression Method. Support vector machines (SVM) was originally developed by Vapnik[61] as a general data modeling methodology where the training set error and the model complexity are incorporated into a special loss function and simultaneously minimized during model development. The importance of the prediction error versus the model complexity can be tuned during the optimization process, in order to generate models with reasonable complexity and avoid overfitting. SVM was later extended to afford the development of SVM regression models for data sets with noninteger variables.

We have implemented SVM for QSAR modeling as described earlier.[62] In brief, given a training set of pairs (x_i, y_i), $i = 1... m$, where x_i is an array of descriptors of each compound and y_i is its biological activity (e.g., IC_{50} value), the sought correlation between structure and activity can be represented as $y_i = f(x_i)$. For simplicity, we define $f(x_i)$ as a linear function

$$f(x_i) = \langle w_i, x_i \rangle + b \tag{4}$$

where w_i is the coefficient vector of the linear function and b is the bias. A major goal of the SVM regression algorithm is to minimize the loss function, which is a combination of prediction error defined by$_{\xi_i}$ and the magnitude of the coefficient C in the following equation

$$\text{loss}_{\min} = \frac{\|w\|}{2} + C\sum_{i=1}^{m} \xi_i \tag{5}$$

with the constraint

$$|y_i - (w\varphi(x_i) + b)| = \xi_i \tag{6}$$

Here, the training vectors x_i are mapped onto a high dimensional space by a kernel function ϕ. In the end, SVM regression is expected to find a linear correlation between the actual activity and this high dimensional space $\phi(x_i)$. For this study, we have implemented a linear kernel. C is a penalty parameter of the error term that controls the weight between two terms in the SVM optimization process.

In many cases, the biological activity may contain small errors or the kernel function may not be capable of perfectly representing the training compounds in a simplified manner. To penalize against complex models, we have added a slack variable ε to the loss function[62] in addition to the penalty parameter C. It is a threshold of prediction error for any compound's activity before the algorithm is penalized for the poor prediction. Beyond this threshold the algorithm is penalized by the value of $\xi_i - \varepsilon$. When combining the SVM optimization process defined in eq 7 with this slack variable, the following loss function is obtained:

$$\text{loss}_{\min} = \frac{\|w\|}{2} + C\sum_{i=1}^{m} \begin{cases} 0 & \text{if } \xi_i \le \varepsilon \\ \xi_i - \varepsilon & \text{if } \xi_i > \varepsilon \end{cases} \tag{7}$$

The nature of SVM regression requires one to specify the values of C and ε a priori since it is not known beforehand which values may work the best for one particular data set; thus, a parameter tuning must be performed. The goal is to identify optimal values of C and ε in that the model can give the best prediction for the test set. For this study, we have chosen to use a "grid-search" scheme on C and ε. It starts with randomly choosing a training/test set split of the data set, conducting a grid-search using those compounds, then fine-tuning the complete data set over the parameter value ranges that exhibited the best results. Our coarse grid-search of C varied from 50 to 1000 with an increment of 80, and ε varied from 0 to 1.5 with an increment of 0.15. Once the best parameter ranges were found, a fine-tuned search was carried out to search values within 200 and 0.3 units for C and ε, with the steps of 5 and 0.05, respectively.

Applicability Domain. Ideally, a QSAR model can predict the target property for any compound for which chemical descriptors can be calculated. However, since *k*NN QSAR modeling predicts test set compound activities by interpolating those of the nearest neighbor compounds in the training set, a special applicability domain, or similarity threshold, should be introduced to avoid extreme model extrapolation by making predictions for compounds that are significantly dissimilar to members of the training set.[50] To measure similarity, each compound is represented by a point in the M-dimensional descriptor space (where M is the total number of descriptors selected in the descriptor pharmacophore) with the coordinates X_{i1}, X_{i2}, ...X_{iM}, where X_{id} are the values of individual descriptors for compound i. The similarity between any two molecules is characterized by the Euclidean distance between their representative points. The Euclidean distance between two points i and j in M-dimensional space can be calculated as follows:

$$d_{ij} = \sqrt{\sum_k (X_{ik} - X_{jk})^2} \tag{8}$$

Compounds with the smallest distance between them are considered to have the highest similarity. The distribution of pairwise compound similarity in the training set is analyzed to produce an applicability domain threshold, D_T, as follows:

$$D_T = \bar{y} + Z\sigma \tag{9}$$

Here, $\bar{y}$ is the average Euclidean distance d_{ij} of the k nearest neighbors of each compound within the training set, σ is the standard deviation of these Euclidean distances, and Z is an arbitrary parameter to control the significance level. On the basis of previous studies in our laboratory, we set the default value of Z as 0.5, which formally places the boundary for which compounds will be predicted at one-half of the standard deviation (assuming a Gaussian distribution between k nearest neighbor compounds in the training set). Thus, if the distance of an external compound from at least one of its nearest neighbors in the training set exceeds this threshold, the prediction is considered unreliable.

As discussed above, our *k*NN QSAR utilizes only descriptors selected by variable selection to limit a model's applicability domain. For SVM QSAR, we chose to take a similar approach. SVM's does not employ a definitive variable selection technique like *k*NN, where the weights of each descriptor are assigned a value of 0 or 1. Instead, the weight of each descriptor is a noninteger number that may have a positive (directly correlated with the biological property) or negative (inversely related to the biological property) number. Knowing this, we wanted to select a subset of descriptors that the model found to correlate well with compound activity. Each model assigns a weight used to correlate descriptor values with a biological property. Since both high positive weights and low negative weights may be vital for activity prediction, we chose to use the absolute value of each of these weights as a measure of how important they were for predicting binding. Of course, the range of these weights may vary drastically between models, so we also normalized the weights between zero and one and then implemented a similar applicability domain compared to the one used by *k*NN. To do that, the calculated Euclidean distances were weighted based on the normalized, absolute weights of each descriptor assigned by the SVM model. This produces a similar pseudoboundary to *k*NN, except the boundary would be extended for descriptors whose weight was close to zero and the boundary would be narrow for descriptors whose normalized, absolute weight was close to one.

External Validation and *Y*-Randomization Test. It is critical to validate a QSAR model by assessing its prediction accuracy for an external set that was not used in model building. We have conducted extensive external validations on both *k*NN and SVM models using two external data sets as described above. In both cases, the prediction accuracy had to satisfy the conditions

$$R^2 > 0.60 \tag{10}$$

$$(R^2 - R_0^2)/R^2 < 0.10 \text{ and } 0.85 < k < 1.15 \tag{11}$$

where *k* is the slope of the regression lines (predicted versus observed activities) through the origin. The predictions were generated using consensus models, and the model coverage for each external data set was calculated as well (vide infra).

Our previous experience suggests that more accurate results are obtained by consensus, that is, by averaging predictions from multiple QSAR models.[62,63] Thus, the consensus QSAR prediction scheme was applied to all validation set compounds found within individual applicability domains of models used in consensus prediction. The averaged predicted activity, the fraction of models that predict the activity, and the variance of the prediction values have been calculated for each compound.

In addition to external validation, *Y*-randomization test was carried out to establish model robustness. The test consists of rebuilding models using shuffled activities of the training set and evaluation of such models' predictive accuracy in comparison with the original model. It is expected that models obtained for the training set with randomized activities should have significantly lower values of statistical parameters such as q^2, R^2, $R_0{}^2$, etc., for training and, especially, test sets. Therefore, if most QSAR models generated in the *Y*-randomization test exhibit relatively high values of the statistical parameters for both training and test sets, it implies that a reliable QSAR model cannot be obtained for the given data set. This test was applied to all QSAR approaches in this study and was repeated twice for each division.

QSAR-Based Virtual Screening. As illustrated in the workflow in Figure 1, the rigorously validated QSAR models were employed for virtual screening. A global applicability domain was applied first in the complete descriptor space to filter out compounds that differed in their structure from the modeling set compounds. All 59 known inhibitors were exploited as the probes during the calculation. During the consensus prediction, the results were accepted only when the compound was found within the applicability domains of more than 50% of all models used in consensus prediction and the standard deviation of estimated means across all models was small. Furthermore, we restricted ourselves to the most conservative applicability domain for each model using $Z_{\text{cutoff}} = 0.5$.

Principle Component Analysis (PCA). The PCA calculations were carried out using the entire set of MolconnZ4.05 descriptors calcualted for all compounds in the modeling set, two external validadtion sets, and virtual screening hits. The purpose of these calculations was to provide a visual means to compare relative positioning of the three data sets plus hits in the chemistry (i.e., dmultidimensional descriptor) space. The programs in the kernlab package[64] of the latest version of R 2.8.0[65] were employed. Using PCA, the distribution of compounds in the original descriptor space could be visualized in a lower dimensional space, normally in the 3D space of the first three principal components.

Experimental Validation of Screening Hits. Recombinant HDACs were purchased from either BIOMOL International (Plymouth Meeting, PA) or PBS Bioscience (San Diego, CA). The inhibitor activity was determined using an HDAC Fluorimetric Assay/Drug Discovery Kit from BIOMOL International according to manufacturer's protocols. Briefly, reactions were set up in 96-well plates in a total of 50 μL HDACs assay buffer (50 mM Tris-HCl of pH 8.0, 137 mM NaCl, 2.7 mM KCl, 1 mM $MgCl_2$) containing HDAC1 (or HDAC6), testing compounds, and HDACs substrate. Trichostatin A served as the positive controlm, and the vehicle, 1% DMSO, was employed as the negative control. The reaction was initiated by the addition of HDACs substrate at room temperature and lasted for 30 min. The final concentration of HDACs substrate was around its apparent K_m. For HDAC1, 50 μM of substrate was used, and for HDAC6, 10 or 30 μM was used. The reaction was then stopped by adding 50 μL of Fluor de Lys (TM) Assay Developer, and the mixture had been incubated for another 15 min at room temperature. The Assay Developer was added to stop the deacetylation reaction and produce fluorophore from the deacetylated substrate. The fluorophore can be excited at 360 nm and emits light at 460 nm. The relative fluorescence is read by a FlexStation II plate reader (Molecular Devices, Sunnyvale, CA). Initial screening concentrations were 100 μM with samples with over 50% inhibition further tested in dose response assays. The raw data (relative fluorescence units) were plotted as a function of the molar concentration of test compounds (in logarithm) and fitted to the three-parameter logistic function to calculate pIC_{50} by Prism 5.0 (GraphPad Software, La Jolla, CA). Here the pIC_{50} is defined as the logarithm of molar concentration of test compound that inhibits the fluorescence production by 50%.

RESULTS AND DISCUSSIONS

***k*NN QSAR Regression Modeling.** The statistical results for the 10 best *k*NN QSAR models using MZ4.05 descriptors are summarized in Table 3; 1385 models, that is, ~50% of the total number of models generated, were accepted since they had both the LOO-CV q^2 values for the training set and linear fit predictive R^2 values for the test set greater than 0.60. Seventy models with q^2/R^2 values exceeding 0.75/0.75 were retained for consensus prediction. As shown in Figure 2A, the most predictive model afforded q^2 value of 0.81 for 34 compounds and R^2 values of 0.80 for 16 compounds (RMSE = 0.38). For models built with MOE descriptors, the best q^2/R^2 values were as high as 0.70/0.76 (RMSE = 0.45, see Figure 2C). The statistics of the top 10 *k*NN/MOE models are summarized in Table S1 of Supporting Information. Similarly, thirteen models with q^2/R^2 values exceeding 0.70/0.70 were employed for consensus prediction. These results suggest that the intrinsic structure-binding affinity relationships exist for HDAC1 inhibi-

Table 3. Statistics for Ten Best *k*NN Models for All Test Sets Using MolconnZ Descriptors

model no.	training set size	test set size	descriptor no.	nearest neighbor no.	q^2 (training set)	R^2 (test set)	R_0^2 (test set)	RMSE (test set)
1	45	5	22	1	0.80	0.87	0.69	0.27
2	41	9	20	2	0.80	0.81	0.77	0.49
3	34	16	14	1	0.81	0.80	0.76	0.38
4	35	15	12	2	0.82	0.79	0.70	0.48
5	42	8	14	1	0.81	0.79	0.73	0.35
6	34	16	26	1	0.80	0.79	0.78	0.37
7	28	22	36	1	0.83	0.77	0.67	0.42
8	40	10	12	2	0.81	0.77	0.77	0.43
9	29	21	20	1	0.79	0.77	0.76	0.47
10	34	16	16	1	0.79	0.76	0.74	0.40

tors that can be best described by *k*NN models using both independent descriptor sets.

To ensure that the models did not merely capture noise, *Y*-randomization test was carried out as described above. As expected, the best models using MZ4.05 descriptors and shuffled activities only produced training set models with q^2 of less than 0.40 (data not shown). Also the best *k*NN/MOE models using randomized activity only yielded the q^2/R^2 values less than 0.40/0.40. These results confirmed that *k*NN models uncovered nonspurious correlations between both MolconnZ and MOE descriptors and compound inhibition activity.

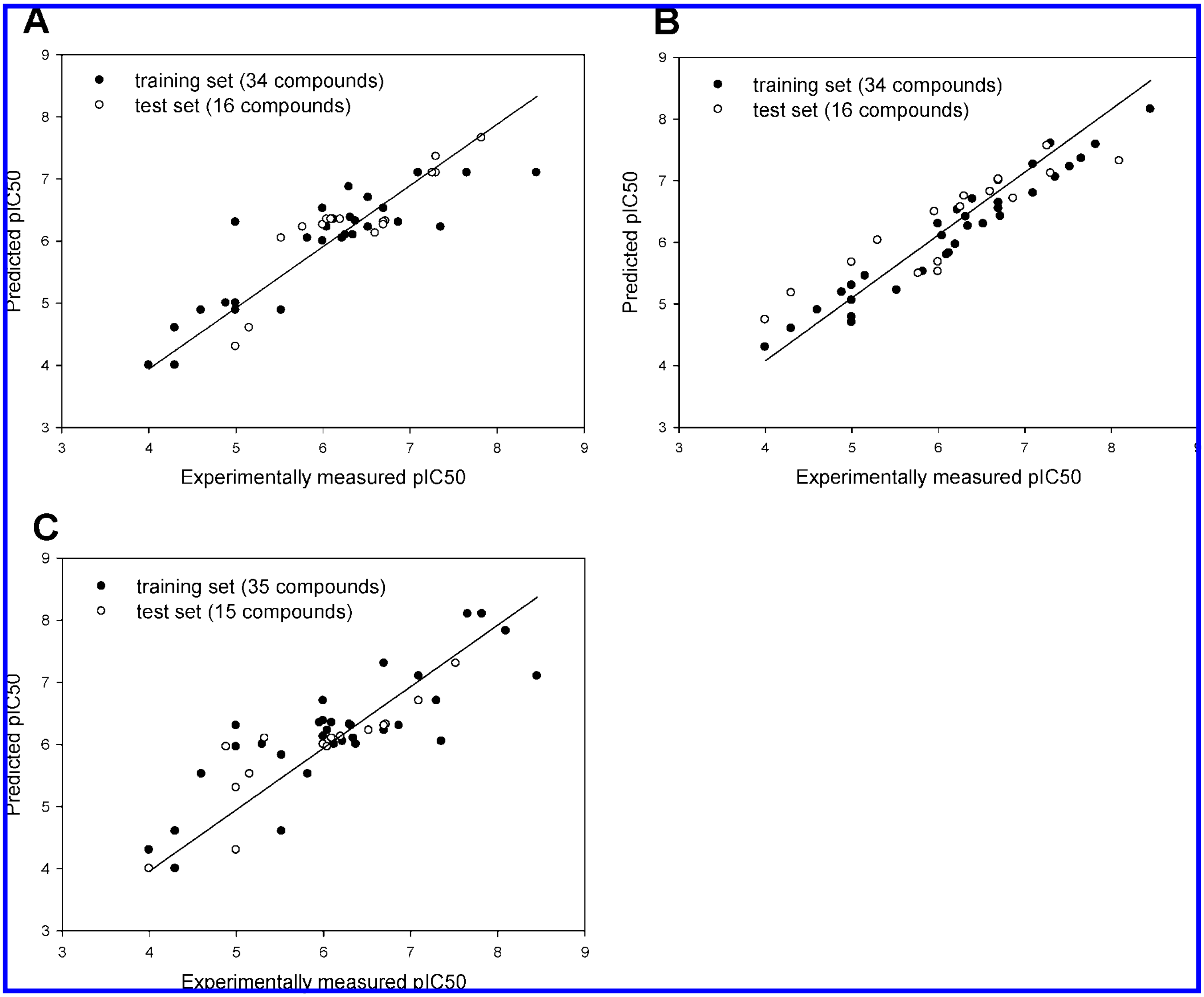


Figure 2. Comparison of actual vs predicted inhibition efficiency (pIC_{50}) values for the best QSAR model for each combination of statistical modeling approach and descriptor type. A. For *k*NN/MolconnZ method (q^2 = 0.81, R^2 = 0.80). The training set contains 34 compounds (●), and the test set contains 16 compounds (○). B. For SVM/MolconnZ method (q^2 = 0.94, R^2 = 0.81). The training set contains 34 compounds (●), and the test set contains 16 compounds (○). C. For *k*NN/MOE models (q^2 = 0.70, R^2 = 0.76). The training set contains 35 compounds (●), and the test set contains 15 compounds (○).

Table 4. Statistics for Ten Best SVM Models for All Test Sets Using MolconnZ Descriptors

model no.	training set size	test set size	C	ε	q^2 (training set)	R^2 (test set)	R_0^2 (test set)	RMSE (test set)
1	37	13	200	0.40	0.93	0.87	0.62	0.36
2	37	13	200	0.50	0.91	0.86	0.66	0.34
3	37	13	200	0.35	0.94	0.85	0.59	0.38
4	37	13	200	0.55	0.90	0.85	0.67	0.34
5	37	13	200	0.60	0.89	0.84	0.68	0.33
6	34	16	200	0.30	0.94	0.81	0.76	0.51
7	35	15	200	0.30	0.95	0.72	0.66	0.49
8	39	11	200	0.30	0.94	0.72	0.71	0.51
9	29	21	200	0.30	0.96	0.71	0.66	0.49
10	35	15	200	0.35	0.94	0.71	0.66	0.49

SVM QSAR Regression Modeling. The statistical results for top 10 SVM QSAR models using MZ4.05 descriptors are summarized in Table 4. The best q^2, R^2, R_0^2 values are as high as 0.93, 0.87, and 0.62, respectively. Figure 2B shows the best predictive model with q^2 value of 0.94 for 34 compounds and R^2 values of 0.81 for 16 compounds (RMSE = 0.51). For this model, the optimum values of C and ε were found to be 200 and 0.30, respectively. The value of 0.30 is reasonable for ε, because it is common to observe a 0.30 log unit error in enzyme/inhibition assays. Seventeen models of SVM/MZ4.05 combination with q^2/R^2 values exceeding 0.70/0.70 were retained for consensus prediction. In comparison, the performance of SVM/MOE combination was much less satisfactory. The best R^2 and R_0^2 value were as low as 0.64 and 0.53, respectively. Meanwhile, the number of acceptable models was drastically small. Thus, we did not employ SVM/MOE models for consensus prediction because of their poor accuracy ($R_{\text{test}}^2 < 0.75$).

To ensure that our SVM QSAR modeling was based on nonspurious structure/activity relationship, the inhibition activities were randomly shuffled for the training set and all calculations were repeated following exactly the same protocol. The best models using randomized data only produced an R^2 of 0.20 for the test set (data not shown), suggesting that the high R^2 is not due to a chance correlation and our accepted SVM models were robust.

Model Validation Using External Data Sets. Both *k*NN and SVM QSAR models validated by test sets were used to predict the inhibition activity of two external validation sets (Tables 5 and 6). For consensus prediction we have employed 70 best *k*NN/MolconnZ models and 17 best SVM/MolconnZ models. For the external validation set 1, the data reported in Table S2−S4 of Supporting Information suggests that both *k*NN/MolconnZ and SVM/MolconnZ consensus models afforded reasonable results. Figure 3 shows the correlation between experimentally measured and calculated activities of the external validation set 1 using three types of consensus models. Among the three, *k*NN/MolconnZ consensus models showed the best performance, with the R^2 of 0.87, R_0^2 of 0.78, and RMSE of 0.59 for 8 compounds (BC-2-83 was found to be out of applicability domain of most *k*NN/MolconnZ models, see Table S2 of Supporting Information). For 7 out of these 8 compounds, the predicted activities were within a reasonable range of 0.5 log unit. However, one compound corresponding to the black circle in Figure 3A was predicted with a large error (>1.0 log unit). A possible explanation for this observation is that this compound is the only one that contains two metal binding groups but no aromatic group. The latter is known to be important for the inhibition activity as suggested by many SAR studies.[15] The SVM/MolconnZ models performed slightly worse than the *k*NN/MolconnZ models, despite the fact that the SVM/MolconnZ combination had better performance for both training and test sets. The R^2 and R_0^2 of consensus prediction by SVM/MolconnZ models was 0.71 and 0.68, respectively, for all 9 compounds (Figure 3B). Interestingly, *k*NN/MOE models showed much worse statistics for the external set 1: The R^2 was 0.60, but the R_0^2 was only 0.25, and the RMSE was as high as 0.84 for 9 compounds (Figure 3C). Though satisfying eq 10, the statistics is not acceptable for *k*NN/MOE combination because the value of $(R^2 - R_0^2)/R^2$ (see eq 11) is too large. Thus, we did not apply this combination to external validation set 2 and the later virtual screening. These results demonstrate the critical need of external validation set for evaluating the model robustness, as well as illustrate a known phenomenon that training set accuracy does not necessarily correlate with model performance for external data sets.[66]

In a typical implementation of our modeling workflow (Figure 1), we only select randomly the external validation set (designated as set 1 in this study) once. The rationale behind this approach is that it emulates a situation faced by experimental medicinal chemists in ongoing projects when they deal with only one "external set", that is, compounds that they plan to synthesize. However, in our studies we have a liberty of selecting several samples of external compounds to test the stability of training set models. Thus, we have made several random splits of the data into modeling and external sets; Table S5 of Supporting Information reports the statistical parameters of models for *k*NN/MolconnZ approach. We found that the prediction accuracy for the external data sets was consistent for all splits although the number of eligible models varied in different runs. The latter is expected considering the chemical diversity of 59 HDAC1 inhibitors. Furthermore, the external prediction accuracy was not affected by applying our consensus prediction scheme and the predictions were consistent for the independent external validation set 2.

We have used the statistical index R_0^2 (eq 3) and RMSE to evaluate model robustness in addition to the correlation coefficient R^2. Traditionally, the latter is considered as a good indicator of predictive power of models. In fact, this coefficient reflects the similarity in relative ranking of compounds based on actual versus the calculated activities rather than the accuracy of the activity prediction. On the other hand, R_0^2 directly compares the actual versus predicted activities because it estimates the fitness of the data to the line with the intercept of zero and the slope of one. It thus gives a better measurement of how well the model predicts compounds' activities, which is why we advocated its use as an important model accuracy metric in our previous studies.[67,68] The above case of *k*NN/MOE consensus prediction illustrates the difference between R^2 and R_0^2, as underscored by eq 11. This suggests that R_0^2 and RMSE are also important indicators of model robustness especially when the size of the test set is small.

External validation set 2 contains HDAC1 inhibitors of different chemical scaffolds; therefore, it can be considered as a real test of the predictability of QSAR models. In

Table 5. Summary of Combinatorial QSAR Modeling for the HDAC1 Inhibitor Modeling Set and the External Validation Set 1 (EV1)

	*k*NN				SVM			
type of descriptor	q^2(CV)	R^2	R_0^2	RMSE	q^2(CV)	R^2	R_0^2	RMSE
MZ4.05	0.80	0.87	0.69	0.27	0.93	0.87	0.62	0.36
		0.87(EV1)	0.78(EV1)	0.59(EV1)		0.71(EV1)	0.68(EV1)	0.52(EV1)
MOE2006.08	0.70	0.76	0.76	0.45	0.91	0.64	0.53	0.54
		0.60(EV1)	0.25(EV1)	0.84(EV1)		N/A	N/A	N/A

Table 6. Consensus Predictions of Inhibition Efficacy for the External Validation Set 2 (EV2) by *k*NN/MolconnZ and SVM/MolconnZ Models

	SVM/MolconnZ			*k*NN/MolconnZ				
[a]comp ID	models cov	pred value	std dev	models cov	pred value	std dev	consensus value	exp pIC_{50}
[b]16c_AE[12]	17/17	6.14	0.15	32/70	6.31	0.77	6.26	6.96
17a_AE	17/17	6.50	0.18	42/70	6.65	0.69	6.61	6.47
17b_AE	17/17	6.60	0.15	42/70	6.78	0.48	6.73	6.68
17d_AE	17/17	6.80	0.15	41/70	6.81	0.49	6.81	8.40
17e_AE	17/17	6.75	0.19	34/70	6.81	0.58	6.79	8.05
17f_AE	17/17	7.36	0.20	33/70	7.06	0.64	7.16	7.74
17g_AE	17/17	7.24	0.24	32/70	7.29	0.81	7.27	8.40
17h_AE	17/17	7.11	0.26	24/70	7.24	0.72	7.18	8.05
17i_AE	17/17	7.68	0.26	31/70	7.25	0.82	7.40	8.40
17j_AE	17/17	7.15	0.24	15/70	6.62	0.83	6.91	8.40
1g_AE	17/17	6.38	0.28	56/70	6.75	0.92	6.66	6.26
1h_AE	17/17	6.56	0.22	57/70	6.99	0.66	6.89	7.77
1i_AE	17/17	6.51	0.21	59/70	7.03	0.37	6.91	6.75
1j_AE	17/17	6.57	0.17	57/70	6.90	0.70	6.83	7.51
1k_AE	17/17	6.75	0.24	39/70	6.96	1.01	6.90	8.22
1l_AE	17/17	6.39	0.22	53/70	6.67	0.94	6.60	6.26
6d_AE	17/17	6.68	0.17	66/70	7.14	0.31	7.05	7.64
6e_AE	17/17	6.66	0.16	65/70	7.16	0.30	7.06	8.52
6f_AE	17/17	7.06	0.18	64/70	7.66	0.43	7.53	8.40
6g_AE	17/17	7.04	0.16	58/70	7.18	0.41	7.15	7.66
6h_AE	17/17	6.65	0.15	58/70	7.19	0.41	7.07	8.30
6i_AE	17/17	6.67	0.18	61/70	7.64	0.43	7.43	8.40
7c_AE	17/17	7.01	0.21	58/70	7.08	0.27	7.06	6.28
7d_AE	17/17	7.31	0.18	56/70	7.17	0.39	7.20	7.82
7e_AE	17/17	6.86	0.17	55/70	7.16	0.40	7.09	7.52
7f_AE	17/17	7.33	0.19	63/70	7.20	0.41	7.23	7.42
7g_AE	17/17	7.04	0.16	56/70	7.24	0.48	7.19	7.25
7h_AE	17/17	7.18	0.20	47/70	7.17	0.33	7.18	8.40
9a_AE	17/17	7.10	0.21	45/70	7.43	0.55	7.34	6.59
9b_AE	17/17	7.27	0.26	47/70	7.50	0.59	7.44	6.19
9d_AE	17/17	6.97	0.16	42/70	7.61	0.71	7.43	6.35
9e_AE	17/17	6.94	0.32	36/70	7.21	0.46	7.12	7.01
9g_AE	17/17	6.83	0.21	27/70	7.09	0.66	6.99	7.82
9h_AE	17/17	7.58	0.31	28/70	7.45	0.88	7.50	8.22
[c]6_CB[10]	17/17	6.50	0.23	62/70	7.28	0.57	7.11	7.36
11_CB	17/17	6.50	0.20	57/70	6.66	0.38	6.62	7.00
17_CB	17/17	7.11	0.16	48/70	7.36	0.45	7.30	8.70
26_CB	17/17	5.93	0.16	1/70	7.02	N/A	5.99	5.32
30_CB	17/17	6.14	0.15	32/70	6.31	0.77	6.26	6.96
31_CB	17/17	6.37	0.18	48/70	6.07	0.21	6.15	6.68
15_CB	17/17	7.44	0.29	70/70	7.12	0.14	7.18	6.70
RMSE		0.92			0.86		0.86	

[a] Only 41 out of the total of 54 compounds are listed. The other 13 compounds were found to be out of applicability domain of QSAR models. [b] All compounds labeled with AE are from ref.[12] [c] All compounds labeled with CB are from ref.[10]

addition, it is fully independent from the 59 compounds of modeling set. Among all 54 inhibitors, 41 could be predicted by the majority of consensus models and the results are summarized in Table 6. For both *k*NN/MolconnZ and SVM/MolconnZ models, 28 out of these 41 compounds had the errors of their predicted activities of less than 1.0 log unit. The RMSE was 0.86 for *k*NN/MolconnZ models, 0.92 for SVM/MolconnZ models, and 0.86 for the consensus averaged value of all models combined. It was shown in our recent study that consensus models afford higher prediction accuracy for the external validation data sets with the highest space coverage as compared to individual constituent models.[69] The same pattern was observed in the present study as well. The RMSE of the consensus score is superior to constituent SVM/MolconnZ models and on par with constituent *k*NN/MolconnZ models. In addition, there is only one compound with a relatively large margin of error (>1.5 log unit) when the consensus prediction is used. For individual constituent models; however, there are three compounds with similarly large errors of prediction with *k*NN/MolconnZ models and five compounds with SVM/MolconnZ models.

Compounds 6e_AE, 17j_AE, and 17d_AE are among those with a large margin of error (~1.5 log unit). They could be analyzed to explore the reasons for QSAR prediction errors. It should be noted that both *k*NN and SVM methods converged

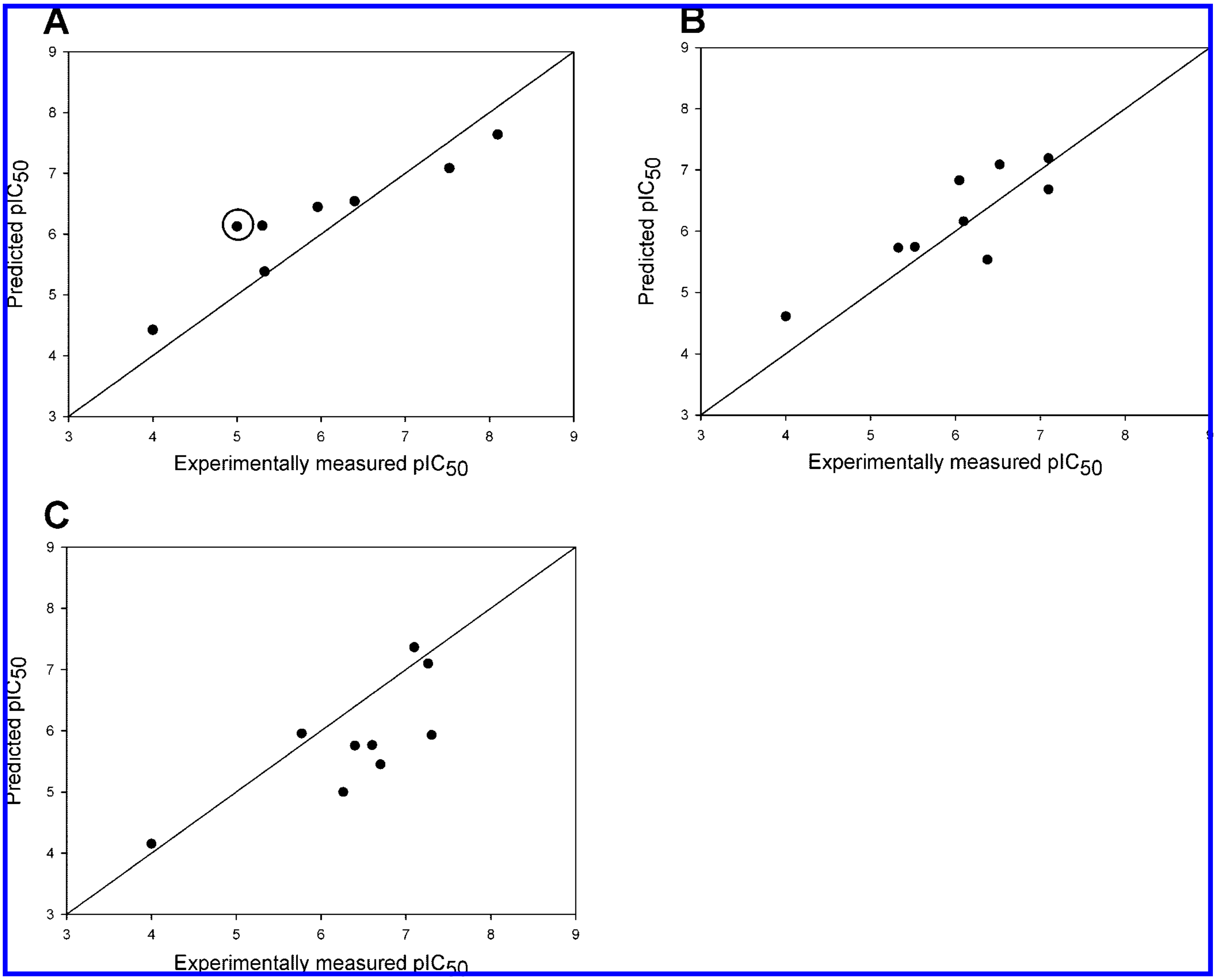


Figure 3. Comparison of actual vs predicted inhibition efficiency (*pIC_{50}*) values for the best QSAR model as applied to the external validation set 1. A. For the *k*NN/MolconnZ method ($R^2 = 0.87$, 8 compounds). The compound with the black circle is the possible structural outlier that has been discussed in the results. B. For the SVM/MolconnZ method ($R^2 = 0.71$, 9 compounds). C. For the *k*NN/MOE method ($R^2 = 0.60$, 9 compounds).

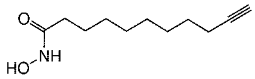
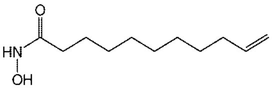
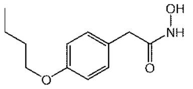
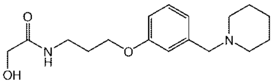
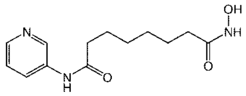
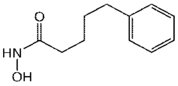
on these three compounds and showed the similar trend of errors (see Table 6). It is feasible that these compounds could be the activity outliers because of experimental errors.

To further assess the chemical diversity of two external validation tests with respect to the modeling set, the PCA was conducted to enable the data set visulization in the space of first three principal components (Figure 5). As shown in Table S7 of Supporting Information, the first three principle components (PCs) explained 54.7% of the total variance of the 59 HDAC1 inhibitors. As observed, the compounds in the external validation set 1 occupied the same regions as the modeling set. Similarly, the distribution of compounds in the external validation set 2 was generally as broad as that for the 59 inhibitors in the modleing set but with less variance at the third PC. It should be pointed out that their pIC_{50} (6.3−8.4) are at the upper range of the modeling set.

QSAR-Based Virtual Screening. On the basis of the results of model validation in the previous section, only *k*NN/MolconnZ and SVM/MolconnZ approaches were used for virtual screening because of their good performances on both modeling set and two external validation sets. Therefore, 70 *k*NN/MolconnZ models and 17 SVM/MolconnZ models with defined applicability domains were applied concurrently toward virtual screening of our chemical libraries. Prior to the consensus predictions, our initial filtering using global applicability domain of modeling set reduced the total number of compounds from approximately 9.5×10^6 to 3.2×10^3. The predicted activities from individual models were averaged to yield a consensus pIC_{50} value. Finally, 45 hits were selected to be of high predicted activities (6.68−7.43 for *k*NN/MolconnZ and 5.94−7.77 for SVM/MolconnZ) and structural uniqueness.

As expected, the predicted activities of HDAC1 inhibitors by two different types of models were not identical but differed by less than 1.0 log unit in most cases. For each of the *k*NN/MolconnZ and SVM/MolconnZ consensus hit, we searched published literature to find out if any of these compounds was reported independently as HDAC1 inhibitors. We found that compounds 33 and 38 have been indeed cited as potential HDAC1 inhibitors (see Table 7).[70,71] Both compounds are structurally similar to SAHA which is a strong HDAC1 inhibitor included in the modeling data set. Furthermore, compounds 2, 28, and 34 were reported to have anti-inflammatory activity that is commonly associated with

Table 7. Consensus Predictions of Inhibition Efficacy for the Confirmed Screening Hits identified by *k*NN/MolconnZ and SVM/MolconnZ Models

Mol.	Compd. #	CAS #	*k*NN/MolconnZ			SVM/MolconnZ			Consens. Value	HDAC1 *pIC*$_{50}$	HDAC6 Inhibition[a]
			Model Cov.	Pred. Value	Std. Dev.	Model Cov.	Pred. Value	Std. Dev.			
	2	123041 -56-5	37/70	7.00	0.46	17/17	6.28	0.30	6.77	5.90	105%
	28	16791 -35-8	36/70	7.04	0.53	17/17	6.49	0.29	6.86	6.00	101%
	34	2438 -72-4	46/70	7.15	0.46	17/17	6.67	0.26	7.02	4.00	99%
	45	78273 -80-0	36/70	6.68	0.67	17/17	5.94	0.19	6.44	< 4.00	42.6%
	33	382180 -17-8	70/70	7.12	0.35	17/17	7.60	0.27	7.21	6.70[70]	
	38	114918 -01-3	42/70	7.12	0.47	17/17	6.59	0.27	6.97	7.62[71]	
RMSE				1.59			1.05		1.50		

[a] The inhibition assay against HDAC6 at the single concentration of 30 μM.

HDAC1 inhibition, which may be viewed as indirect evidence in support of the prediction.

In general, as shown in Table S6 of the Supporting Information, most hit compounds contain long aliphatic chain that permits the chelating group to reach the bottom of the binding pocket and coordinate with the zinc ion. An aromatic group at the opposite end of the chelating group is supposed to enhance inhibition through hydrophobic interaction with the capping region of the active site. These are actually the common structural features known for HDAC1 inhibitors. Furthermore, many additional features are also found in the hit compounds, such as triple bonds (compound 2) and 3-bromo-4-hydroxy-phenyl group (compounds 11, 14), which exist in HDAC1 inhibitors such as Oxamflatin and Psammaplin A.[10,31] It should be pointed out that these functional groups were not present in the original modeling data set, which demonstrates the ability of QSAR-based virtual screening to uncover computational hits with novel chemical features. The existence of unsaturated bonds in the linker region between the chelating group and the cap region has been observed frequently among many screening hits. However, this feature is only found in TSA (that was included in the training set), which has the highest inhibitory activity (pIC$_{50}$) of 8.46. Since this feature is not often seen in other known inhibitors, this observation should be additionally explored for lead optimization in future studies. The unsaturated bonds in the linker region likely restraint the conformational freedom of the long aliphatic chain, which could help decrease the unfavorable entropy change during the inhibitor binding.

In comparison with 59 HDAC1 inhibitors in the modeling set, the 45 screening hits occupied only the fraction of the space covered by the modeling set (see Figure 5). This should be expected given that only the hits with high predcted inhibition efficacy (pIC$_{50}$ > 6.50) were chosen as the final hits. Notably, 5−7 screening hits, including the experimentally confirmed ones (2/123041-56-5, 28/16791-35-8) were found to be separated from other compounds. They were located at the region with coordinates of 30.0 for PC1, 5.0 for PC2, and 0 for PC3. These uniqueness in the PCA chemical space were consistent with the novel structural features found in the two hits, that is, the triple bond or unsaturated bond, as well as the long aliphatic chain at the linker region.

In recent years, our group has explored the hit identification strategy that combines rigorously validated QSAR models and virtual screening.[62,72−76] It has been shown that our current workflow is capable of identifying potent compounds of novel chemical scaffolds as compared to modeling set compounds, especially in the cases of anticonvulsant agents[63] and D1 dopaminergic antagonists.[62] There are several aspects of our current protocol for QSAR based virtual screening that need to be highlighted. First, models built using variable selection approaches only include a subset of all descriptors, that is, those identified as significant in the process of model optimization. This feature of individual models coupled with the applicability domain threshold could result in misannotation of some structurally diverse molecules in the virtual screening databases as inactives. Consensus prediction scheme provides a viable solution to this problem because each model has its own limitations, but the ensemble of models covers much greater chemical feature space and, consequently, could identify putatively active compounds of greater chemical diversity. Second, the dependent variable in the current data set is the continuous value of inhibition potency. During model building, all descriptors with constant values have been eliminated, and only the descriptor types that are used in predictive QSAR

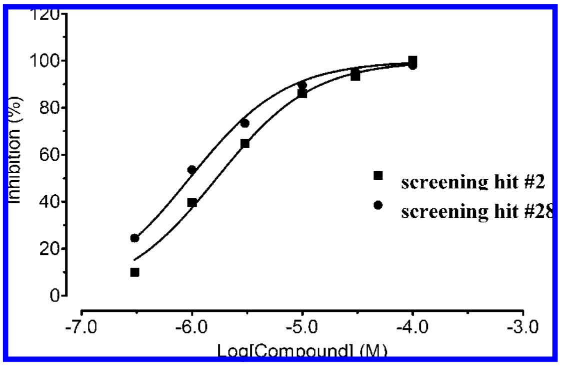


Figure 4. Full dose response curve for hit compounds 2 and 28 in human HDAC1 inhibition assay.

models were retained. Obviously, descriptors with the same values for all compounds in the training set could not contribute to the QSAR model that always correlates changes in chemical structure to changes in biological activity. However, there is a possibility that some of these eliminated descriptors (that apparently describe chemical features common to all inhibitors) are essential for discriminating inhibitors from nonbinders. Thus, if these descriptors are not considered in virtual screening there is a probability of identifying false positives. To circumvent this problem, we have applied global applicability domain in the preliminary screening step to filter out compounds that are generally structurally dissimilar from the modeling set compounds.

Experimental Validation. Four structurally diverse hits with moderate to high predicted activity were selected from the 45 consensus virtual screening hits for experimental validation taking into account commercial availability. To our satisfaction, compounds 2, 28, and 34 were confirmed to be micromolar inhibitors against HDAC1 (Figure 4 and Table 7). Among them, compound 28 showed the best inhibitory activity with pIC_{50} values of 6.00. The fourth compound, 45, did not inhibit HDAC1 at the concentration of 300 μM. However, interestingly enough this compound was later identified by us as a selective inhibitor for HDAC6, a class II HDACs enzyme. At the concentration of 30 μM, 45 inhibited about 42.6% of HDAC6 activity, while other three compounds (2, 28, and 34) showed 105%, 101%, and 99% inhibition, respectively. Moreover, it is of notice that the chelating functional group in 45 is unique compared to other hits. This observation could be further explored for rational design of class/subtype selective HDACI.

Our current screening libraries include the WDI database, which contains approximately 59 000 approved or investigational drugs in the world. It has become a practical strategy to screen this database during the early phase of drug development. The hits identified in this library could be placed on the fast track and avoid the risk and length of preclinical/clinical studies. In our study, two hits that were submitted for experimental validation were actually identified from the WDI database. Compound 34 is Bufexamac, a marketed drug used for joint and muscular pain, while compound 45 is Roxatidine, a widely used competitive H2 receptor antagonist for the treatment of peptic ulcer. These two hits will enrich the candidates pool of HDACI and potentially facilitate the pipeline of drug development, a strategy known as repurposing.[77]

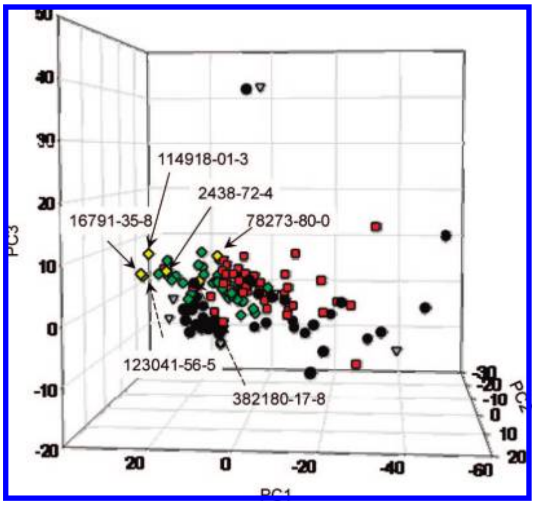


Figure 5. Plot of HDAC1 inhibitors in the coordinate system formed by the first three principle components resulting from the Principle Component Analysis (PCA) compounds represented by MolconnZ4.05 descriptors. The distribution plot includes modeling set (59, black circles), external validation set 1 (9, white triangle), external validation set 2 (54, red squares), and 45 virtual screening hits including 6 confirmed hits (yellow diamonds, labeled by compounds' CAS numbers) and 39 untested hits (green diamonds).

CONCLUSIONS

We have employed a combinatorial QSAR approach to generate models for 59 chemically diverse compounds tested for their inhibitory activity against HDAC1. The SVM and *k*NN QSAR methods were used in combination with MolconnZ and MOE descriptors independently to identify the best approach with the highest external predictive power. Highly predictive QSAR models were generated with *k*NN/MolconnZ and SVM/MolconnZ approaches. Rigorously validated QSAR models were then used to screen our in-house database collection of a total of over 9.5 million compounds. This study resulted in 45 consensus hits that were predicted to be potent HDAC1 inhibitors. Two hit compounds that were not present in the original data set were nevertheless reported recently as HDAC1 inhibitors.[70,71] Four hit compounds with interesting chemical features were purchased and experimentally validated. Three of them were confirmed to have inhibitory activities to HDAC1 (Class I HDACs), and the best activity obtained was IC_{50} of 1.00 μM. The fourth compound was later identified to be a selective inhibitor to HDAC6, a Class II HDACs. Moreover, two of the confirmed hits are marketed drugs which could potentially expedite their development as anticancer drugs acting via HDAC1 inhibition. This study illustrates that validated QSAR models have the ability of identifying novel structurally diverse hits by the means of virtual screening. We believe that the technology described in this study could be used for data analysis and hypothesis generation in many computational drug discovery studies.

ACKNOWLEDGMENT

We would like to thank Dr. Alexander Golbraikh for helpful discussions concerning QSAR modeling and virtual screening and constructive criticism throughout this research project. We are also grateful to Tripos, Chemical Computing Group and eduSoft for software grants. Finally, we acknowledge the access to the computing facilities at the ITS Research Computing Division of the University of North Carolina at Chapel Hill. The studies reported in this paper were supported in part by the NIH research grant GM066940 (A.T.), the planning grant HG003898 (A.T.), the NIMH Psychoactive Drug Screening Program contract HHSN-271-2008-00025-C (B.L.R.), and the Alzheimers's Disease Drug Discovery Foundation/Elan grant No. 271210 (A.P.K.). H.T. and X.S.W. contributed to experimental design, model building, virtual screening, and manuscript drafting. X.H. ran validation assays under the direction of B.L.R. K.V.B. and A.P.K. provided chemical data for HDACI modeling set, while M.J. generated the corresponding biological data. A.T. supervised computational modeling studies and is responsible for the overall study design.

Supporting Information Available: The consensus predictions for the external validation set, model statistics of multiple splits, predictions of inhibition efficacy for screening hits, and other supplementary tables indicated in the text. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Hassig, C. A.; Schreiber, S. L. Nuclear histone acetylases and deacetylases and transcriptional regulation: HATs off to HDACs. *Curr. Opin. Chem. Biol.* **1997**, *1*, 300–308.

(2) Wolffe, A. P. Histone deacetylase: a regulator of transcription. *Science* **1996**, *272*, 371–372.

(3) Yamagoe, S.; Kanno, T.; Kanno, Y.; Sasaki, S.; Siegel, R. M.; Lenardo, M. J.; Humphrey, G.; Wang, Y.; Nakatani, Y.; Howard, B. H.; Ozato, K. Interaction of histone acetylases and deacetylases in vivo. *Mol. Cell. Biol.* **2003**, *23*, 1025–1033.

(4) Mork, C. N.; Faller, D. V.; Spanjaard, R. A. A mechanistic approach to anticancer therapy: Targeting the cell cycle with histone deacetylase inhibitors. *Curr. Pharm. Des* **2005**, *11*, 1091–1104.

(5) Gui, C. Y.; Ngo, L.; Xu, W. S.; Richon, V. M.; Marks, P. A. Histone deacetylase (HDAC) inhibitor activation of p21WAF1 involves changes in promoter-associated proteins, including HDAC1. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 1241–1246.

(6) Johnstone, R. W. Histone-deacetylase inhibitors: novel drugs for the treatment of cancer. *Nat. Rev. Drug Discovery* **2002**, *1*, 287–299.

(7) Johnstone, R. W.; Licht, J. D. Histone deacetylase inhibitors in cancer therapy: is transcription the primary target. *Cancer Cell* **2003**, *4*, 13–18.

(8) Wolffe, A. P. Chromatin remodeling: why it is important in cancer. *Oncogene* **2001**, *20*, 2988–2990.

(9) Bolden, J. E.; Peart, M. J.; Johnstone, R. W. Anticancer activities of histone deacetylase inhibitors. *Nat. Rev. Drug Discovery* **2006**, *5*, 769–784.

(10) Rodriquez, M.; Aquino, M.; Bruno, I.; De Martino, G.; Taddei, M.; Gomez-Paloma, L. Chemistry and biology of chromatin remodeling agents: state of art and future perspectives of HDAC inhibitors. *Curr. Med. Chem.* **2006**, *13*, 1119–1139.

(11) Yoshida, M.; Kijima, M.; Akita, M.; Beppu, T. Potent and specific inhibition of mammalian histone deacetylase both in vivo and in vitro by trichostatin A. *J. Biol. Chem.* **1990**, *265*, 17174–17179.

(12) Curtin, M.; Glaser, K. Histone deacetylase inhibitors: the Abbott experience. *Curr. Med. Chem.* **2003**, *10*, 2373–2392.

(13) Juvale, D. C.; Kulkarni, V. V.; Deokar, H. S.; Wagh, N. K.; Padhye, S. B.; Kulkarni, V. M. 3D-QSAR of histone deacetylase inhibitors: hydroxamate analogues. *Org. Biomol. Chem.* **2006**, *4*, 2858–2868.

(14) Wang, D. F.; Helquist, P.; Wiech, N. L.; Wiest, O. Toward selective histone deacetylase inhibitor design: homology modeling, docking studies, and molecular dynamics simulations of human class I histone deacetylases. *J. Med. Chem.* **2005**, *48*, 6936–6947.

(15) Woo, S. H.; Frechette, S.; Abou, K. E.; Bouchain, G.; Vaisburg, A.; Bernstein, N.; Moradei, O.; Leit, S.; Allan, M.; Fournel, M.; Trachy-Bourget, M. C.; Li, Z.; Besterman, J. M.; Delorme, D. Structurally simple trichostatin A-like straight chain hydroxamates as potent histone deacetylase inhibitors. *J. Med. Chem.* **2002**, *45*, 2877–2885.

(16) Richon, V. M.; Webb, Y.; Merger, R.; Sheppard, T.; Jursic, B.; Ngo, L.; Civoli, F.; Breslow, R.; Rifkind, R. A.; Marks, P. A. Second generation hybrid polar compounds are potent inducers of transformed cell differentiation. *Proc. Natl. Acad. Sci. U.S.A* **1996**, *93*, 5705–5708.

(17) Grant, S.; Easley, C.; Kirkpatrick, P. Vorinostat. *Nat. Rev. Drug Discovery* **2007**, *6*, 21–22.

(18) Landry, J.; Sutton, A.; Tafrov, S. T.; Heller, R. C.; Stebbins, J.; Pillus, L.; Sternglanz, R. The silencing protein SIR2 and its homologs are NAD-dependent protein deacetylases. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 5807–5811.

(19) Gregoretti, I. V.; Lee, Y. M.; Goodson, H. V. Molecular evolution of the histone deacetylase family: Functional implications of phylogenetic analysis. *J. Mol. Biol.* **2004**, *338*, 17–31.

(20) Finnin, M. S.; Donigian, J. R.; Cohen, A.; Richon, V. M.; Rifkind, R. A.; Marks, P. A.; Breslow, R.; Pavletich, N. P. Structures of a histone deacetylase homologue bound to the TSA and SAHA inhibitors. *Nature* **1999**, *401*, 188–193.

(21) Somoza, J. R.; Skene, R. J.; Katz, B. A.; Mol, C.; Ho, J. D.; Jennings, A. J.; Luong, C.; Arvai, A.; Buggy, J. J.; Chi, E.; Tang, J.; Sang, B. C.; Verner, E.; Wynands, R.; Leahy, E. M.; Dougan, D. R.; Snell, G.; Navre, M.; Knuth, M. W.; Swanson, R. V.; McRee, D. E.; Tari, L. W. Structural snapshots of human HDAC8 provide insights into the class I histone deacetylases. *Structure.* **2004**, *12*, 1325–1334.

(22) Vannini, A.; Volpari, C.; Filocamo, G.; Casavola, E. C.; Brunetti, M.; Renzoni, D.; Chakravarty, P.; Paolini, C.; De Francesco, R.; Gallinari, P.; Steinkuhler, C.; Di Marco, S. Crystal structure of a eukaryotic zinc-dependent histone deacetylase, human HDAC8, complexed with a hydroxamic acid inhibitor. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 15064–15069.

(23) Ragno, R.; Simeoni, S.; Valente, S.; Massa, S.; Mai, A. 3-D QSAR studies on histone deacetylase inhibitors. A GOLPE/GRID approach on different series of compounds. *J. Chem. Inf. Model.* **2006**, *46*, 1420–1430.

(24) Tropsha, A.; Golbraikh, A. Predictive QSAR Modeling Workflow, Model Applicability Domains, and Virtual Screening. *Curr. Pharm. Des.* **2007**, *13*, 3494–3504.

(25) Chen, B.; Petukhov, P. A.; Jung, M.; Velena, A.; Eliseeva, E.; Dritschilo, A.; Kozikowski, A. P. Chemistry and biology of mercaptoacetamides as novel histone deacetylase inhibitors. *Bioorg. Med. Chem. Lett.* **2005**, *15*, 1389–1392.

(26) Kozikowski, A. P.; Chen, Y.; Gaysin, A.; Chen, B.; D'Annibale, M. A.; Suto, C. M.; Langley, B. C. Functional Differences in Epigenetic Modulators-Superiority of Mercaptoacetamide-Based Histone Deacetylase Inhibitors Relative to Hydroxamates in Cortical Neuron Neuroprotection Studies. *J. Med. Chem.* **2007**, *50*, 3054–3061.

(27) Kozikowski, A. P., Dritschilo, A., Jung, M., Petukhov, P. A., Chen, B. Histone deacetylase inhibitors for treatment of neurological diseases and cancer. PCT Int. Appl. 2005007091, 2005.

(28) Kozikowski, A. P., Dritschilo, A., Jung, M., Petukhov, P. A., Chen, B. Preparation of w-ureido alkanohydroxamic acid and related urea derivatives as histone deacetylase inhibitors. U.S. Pat. Appl. Publ. 2005014839, 2005.

(29) Kozikowski, A. P., Chen, B. Preparation of hydroxyamides and mercaptoacetamides as histone deacetylase inhibitors for treatment of neurological diseases and cancer. U. S. Pat. Appl. Publ. 2005032831, 2005.

(30) Kozikowski, A. P., Jung, M., Dritschilo, A. Isoform-selective HDAC inhibitors including biphenyl hydroxamic acid- and mercaptoacetamide-containing amino acid amides, their preparation and use for treating cancer, neurological diseases and malaria. PCT Int. Appl. 2008019025, 2008.

(31) Ohtani, M.; Matsuura, T.; Shirahase, K.; Sugita, K. (2*E*)-5-[3-[(phenylsulfonyl)amino]phenyl]-pent-2-en-4-ynohydroxamic acid and its derivatives as novel and potent inhibitors of ras transformation. *J. Med. Chem.* **1996**, *39*, 2871–2873.

(32) Jung, M.; Brosch, G.; Kolle, D.; Scherf, H.; Gerhauser, C.; Loidl, P. Amide analogues of trichostatin A as inhibitors of histone deacetylase and inducers of terminal cell differentiation. *J. Med. Chem.* **1999**, *42*, 4669–4679.

(33) Lu, Q.; Yang, Y. T.; Chen, C. S.; Davis, M.; Byrd, J. C.; Etherton, M. R.; Umar, A.; Chen, C. S. Zn^{2+}-chelating motif-tethered short-chain fatty acids as a novel class of histone deacetylase inhibitors. *J. Med. Chem.* **2004**, *47*, 467–474.

(34) Suzuki, T.; Ando, T.; Tsuchiya, K.; Fukazawa, N.; Saito, A.; Mariko, Y.; Yamashita, T.; Nakanishi, O. Synthesis and histone deacetylase inhibitory activity of new benzamide derivatives. *J. Med. Chem.* **1999**, *42*, 3001–3003.

(35) Jimenez, C.; Crews, P. Novel Marine Sponge Derived Amino-Acids 0.13. Additional Psammaplin Derivatives from Psammaplysilla-Purpurea. *Tetrahedron* **1991**, *47*, 2097–2102.

(36) Irwin, J. J.; Shoichet, B. K. ZINC--a free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model.* **2005**, *45*, 177–182.

(37) World Drug Index (WDI). http://www.daylight.com/products/wdi.html.

(38) Synergy Libraries. http://www.asinex.com/libraries_synergy.html (accessed Jan 1,2006).

(39) InterBioScreen Libraries. http://www.ibscreen.com (accessed Jan 1,2007).

(40) Progenitor Databases. http://www.chemizon.com (accessed Jan 1,2006).

(41) *MolconnZ*, version 4.05; eduSoft, LC: Ashland, VA, 2006.

(42) Randi, M. On characterization on molecular branching. *J. Am. Chem. Soc.* **1975**, *97*, 6609–6615.

(43) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Structure−Activity Analysis*; Wiley: New York, 1986.

(44) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Chemistry and Drug Research*; Academic Press: New York, 1976.

(45) Kier, L. B.; Hall, L. H. *Molecular Structure Description: The Electrotopological State*; Academic Press: New York, 1999.

(46) Kier, L. B.; Hall, L. H. An index of electrotopological state of atoms in molecules. *J. Math. Chem.* **1991**, *7*, 229.

(47) Kier, L. B.; Hall, L. H. An electrotopological state index for atoms in molecules. *Pharm. Res.* **1990**, *7*, 801.

(48) Wiener, H. Structural determination of paraffin boiling points. *J. Am. Chem. Soc.* **1947**, *69*, 17–20.

(49) Petitjean, M. Applications of the radius−diameter diagram to the classification of topological and geometrical shapes of chemical compounds. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 331–337.

(50) Zheng, W.; Tropsha, A. Novel variable selection quantitative structure−property relationship approach based on the *k*-nearest-neighbor principle. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 185–194.

(51) *Molecular Operation Eenvironment*, version 2006.08; Chemical Computing Group Inc.:Montreal, Quebec, Canada, 2006.

(52) Wiener, H. Correlation of heats of isomerization, and differences in heats of vaporization of isomers, among the paraffin hydrocarbons. *J. Am. Chem. Soc.* **1947**, *69*, 2636–2638.

(53) Balaban, A. T. Highly discriminating distance-based topological index. *Chem. Phys. Lett.* **1982**, *89*, 399–404.

(54) Balaban, A. T. Five new topological indices for the branching of tree-like graphs. *Theor. Chim. Acta* **1979**, *53*, 355–375.

(55) Gasteiger, J.; Marsili, M. Iterative partial equalization of orbital electronegativity—A rapid access to atomic charges. *Tetrahedron* **1980**, *36*, 3219–3228.

(56) Stanton, D.; Jurs, P. Development and use of charged partial surface area structural descriptors in computer-assisted quantitative structure−property relationship studies. *Anal. Chem.* **1990**, *62*, 2323–2329.

(57) Golbraikh, A.; Tropsha, A. Predictive QSAR modeling based on diversity sampling of experimental datasets for the training and test set selection. *J. Comput.-Aided Mol. Des* **2002**, *16*, 357–369.

(58) Golbraikh, A.; Shen, M.; Xiao, Z.; Xiao, Y. D.; Lee, K. H.; Tropsha, A. Rational selection of training and test sets for the development of validated QSAR models. *J. Comput.-Aided Mol. Des* **2003**, *17*, 241–253.

(59) Sharaf, M. A.; Illman, D. L.; Kowalski, B. R. *Chemometrics*; John Wiley & Sons: New York, 1986.

(60) Tropsha, A.; Zhang, W. F. Identification of the descriptor pharmacophores using variable selection QSAR: Applications to database mining. *Curr. Pharm. Des.* **2001**, *7*, 599–612.

(61) Vapnik, V. *The Nature of Statistical Learning Theory*; Springer-Verlag: New York, 1995.

(62) Oloff, S.; Mailman, R. B.; Tropsha, A. Application of validated QSAR models of D1 dopaminergic antagonists for database mining. *J. Med. Chem.* **2005**, *48*, 7322–7332.

(63) Shen, M.; Beguin, C.; Golbraikh, A.; Stables, J. P.; Kohn, H.; Tropsha, A. Application of predictive QSAR models to database mining: identification and experimental validation of novel anticonvulsant compounds. *J. Med. Chem.* **2004**, *47*, 2356–2364.

(64) Karatzoglou, A.; Smola, A.; Hornik, K.; Zeileis, A. kernlab—An S4 package for kernel methods in R. *J Stat. Software* **2004**, *11*, 1–20.

(65) R Development Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, 2004.

(66) Golbraikh, A.; Tropsha, A. Beware of q2. *J. Mol. Graph. Modell.* **2002**, *20*, 269–276.

(67) Votano, J. R.; Parham, M.; Hall, L. M.; Hall, L. H.; Kier, L. B.; Oloff, S.; Tropsha, A. QSAR modeling of human serum protein binding with several modeling techniques utilizing structure-information representation. *J. Med. Chem.* **2006**, *49*, 7169–7181.

(68) Zhang, S. X.; Golbraikh, A.; Oloff, S.; Kohn, H.; Tropsha, A. A novel automated lazy learning QSAR (ALL-QSAR) approach: Method development, applications, and virtual screening of chemical databases using validated ALL-QSAR models. *J. Chem. Inf. Model.* **2006**, *46*, 1984–1995.

(69) Zhu, H.; Tropsha, A.; Fourches, D.; Varnek, A.; Papa, E.; Gramatica, P.; Oberg, T.; Dao, P.; Cherkasov, A.; Tetko, I. V. Combinatorial QSAR modeling of chemical toxicants tested against Tetrahymena pyriformis. *J. Chem. Inf. Model.* **2008**, *48*, 766–784.

(70) Remiszewski, S. W.; Sambucetti, L. C.; Atadja, P.; Bair, K. W.; Cornell, W. D.; Green, M. A.; Howell, K. L.; Jung, M.; Kwon, P.; Trogani, N.; Walker, H. Inhibitors of human histone deacetylase: synthesis and enzyme and cellular activity of straight chain hydroxamates. *J. Med. Chem.* **2002**, *45*, 753–757.

(71) Wang, D. F.; Wiest, O.; Helquist, P.; Lan-Hargest, H. Y.; Wiech, N. L. QSAR studies of PC-3 cell line inhibition activity of TSA and SAHA-like hydroxamic acids. *Bioorg. Med. Chem. Lett.* **2004**, *14*, 707–711.

(72) Hsieh, J. H.; Wang, X. S.; Teotico, D.; Golbraikh, A.; Tropsha, A. Differentiation of AmpC beta-lactamase binders vs. decoys using classification kNN QSAR modeling and application of the QSAR classifier to virtual screening. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 593–609.

(73) Medina-Franco, J. L.; Golbraikh, A.; Oloff, S.; Castillo, R.; Tropsha, A. Quantitative structure-activity relationship analysis of pyridinone HIV-1 reverse transcriptase inhibitors using the k nearest neighbor method and QSAR-based database mining. *J Comput.-Aided Mol. Des* **2005**, *19*, 229–242.

(74) Shen, M.; LeTiran, A.; Xiao, Y.; Golbraikh, A.; Kohn, H.; Tropsha, A. Quantitative structure−activity relationship analysis of functionalized amino acid anticonvulsant agents using *k* nearest neighbor and simulated annealing PLS methods. *J. Med. Chem.* **2002**, *45*, 2811–2823.

(75) Wang, X. S.; Tang, H.; Golbraikh, A.; Tropsha, A. Combinatorial QSAR modeling of specificity and subtype selectivity of ligands binding to serotonin receptors 5HT1E and 5HT1F. *J. Chem. Inf. Model* **2008**, *48*, 997–1013.

(76) Zhang, S.; Wei, L.; Bastow, K.; Zheng, W.; Brossi, A.; Lee, K. H.; Tropsha, A. Antitumor Agents 252. Application of validated QSAR models to database mining: discovery of novel tylophorine derivatives as potential anticancer agents. *J. Comput.-Aided Mol. Des* **2007**, *21*, 97–112.

(77) O'Connor, K. A.; Roth, B. L. Finding new tricks for old drugs: An efficient route for public-sector drug discovery. *Nat. Rev. Drug Discovery* **2005**, *4*, 1005–1014.

(78) Chen, Y. D.; Jiang, Y. J.; Zhou, J. W.; Yu, Q. S.; You, Q. D. Identification of ligand features essential for HDACs inhibitors by pharmacophore modeling. *J. Mol. Graph. Modell.* **2008**, *26*, 1160–1168.

(79) Kozikowski, A. P.; Chen, Y.; Gaysin, A. M.; Savoy, D. N.; Billadeau, D. D.; Kim, K. H. Chemistry, biology, and QSAR studies of substituted biaryl hydroxamates and mercaptoacetamides as HDAC inhibitors-nanomolar-potency inhibitors of pancreatic cancer cell growth. *ChemMedChem.* **2008**, *3*, 487–501.

(80) Ragno, R.; Simeoni, S.; Rotili, D.; Caroli, A.; Botta, G.; Brosch, G.; Massa, S.; Mai, A. Class II-selective histone deacetylase inhibitors. Part 2: alignment-independent GRIND 3-D QSAR, homology and docking studies. *Eur. J. Med. Chem.* **2008**, *43*, 621–632.

(81) Guo, Y.; Xiao, J.; Guo, Z.; Chu, F.; Cheng, Y.; Wu, S. Exploration of a binding mode of indole amide analogues as potent histone deacetylase inhibitors and 3D-QSAR analyses. *Bioorg. Med. Chem.* **2005**, *13*, 5424–5434.

(82) Wagh, N. K.; Deokar, H. S.; Juvale, D. C.; Kadam, S. S.; Kulkarni, V. M. 3D-QSAR of histone deacetylase inhibitors as anticancer agents by genetic function approximation. *Indian J. Biochem. Biophys* **2006**, *43*, 360–371.

(83) Xie, A.; Liao, C.; Li, Z.; Ning, Z.; Hu, W.; Lu, X.; Shi, L.; Zhou, J. Quantitative structure−activity relationship study of histone deacetylase inhibitors. *Curr. Med. Chem. Anticancer Agents* **2004**, *4*, 273–299.

CI800366F