



## Research article

# Deciphering molecular insights of HDAC6 inhibition through SHAP-based interpretation of optimized machine learning models

Nguyen Thi Thuan<sup>a</sup>, My-Kristyna Nguyen-Thao<sup>a</sup>, Le Thi Dung<sup>b</sup>, Do Thi Mai Dung<sup>a\*</sup>

<sup>a</sup>Hanoi University of Pharmacy, 13-15 Le Thanh Tong, Cua Nam, Hanoi, Vietnam

<sup>b</sup>Binh Duong Medical College, 529 Le Hong Phong, Phu Loi, Ho Chi Minh City

\* Corresponding author: [dungdtm@hup.edu.vn](mailto:dungdtm@hup.edu.vn)

## ARTICLE INFO

### Article history

Received 03 May 2025

Revised 12 July 2025

Accepted 04 August 2025

### Keywords

HDAC6 inhibitors

Machine Learning

SHAP

Explainable AI (XAI)

## ABSTRACT

Histone deacetylase 6 (HDAC6) is an important target for cancer treatment; however creating effective and selective inhibitors remains a considerable challenge. Machine learning (ML) can speed up drug discovery, though the interpretability of these models is limited. This study aimed to create optimized ML models to predict HDAC6 inhibitory activity, using SHapley Additive exPlanations (SHAP) to enhance interpretability. Bioactivity data (IC<sub>50</sub> values) for HDAC6 inhibitors were curated from ChEMBL and BindingDB. All inhibitors were classified as active or inactive based on comparison to SAHA. Five ML algorithms (Decision Tree, Random Forest, SVM, XGBoost, AdaBoost) were trained using five different molecular fingerprints: MACCS Keys, Morgan2, ECFP2, ECFP4, ECFP6. Hyperparameter tuning was conducted to optimize model performance. The best-performing model, ECFP6-RF, achieved high predictive performance on the test set (Accuracy: 90.20%, Precision: 91.53%, AUC-ROC: 96.25%) while maintaining minimal overfitting (train-test gap < 8%). SHAP analysis of the ECFP6-RF model identified key structural features that strongly contributed to HDAC6 inhibition. Notably, fragments associated with the hydroxamic acid zinc-binding group and specific aliphatic/aromatic linkers were highlighted as highly influential, consistent with established structure-activity relationships (SAR). This work demonstrates the successful application of optimized ML models combined with explainable AI (XAI), providing interpretable insights into the molecular determinants of activity.

\*Corresponding author: [dungdtm@hup.edu.vn](mailto:dungdtm@hup.edu.vn)

<http://doi.org/10.59882/1859-364X/311>

## INTRODUCTION

Histone deacetylases (HDACs) are an important group of enzymes involved in epigenetics and gene regulation primarily through removing acetyl groups from lysine residues on histones. This affects the interaction of transcription factors and post-translational modifications, thereby regulating gene expression while also maintaining protein stability and supporting other cellular functions. However, when HDACs are overexpressed, they can contribute to the development of serious diseases such as cancer, neurodegenerative disorders, and chronic inflammatory conditions [1]. So, HDACs have gained significant attention, pushing research into novel treatments that aim to reduce their influence on such diseases.

Among the different histone deacetylases, HDAC6 stands out because of its unique characteristics. Unlike most HDACs, which function in the nucleus, HDAC6 is mainly found in the cytoplasm, where it regulates non-histone proteins such as  $\alpha$ -tubulin, Hsp90, and cortactin. These interactions allow HDAC6 to influence various cellular processes, including those linked to tumor growth. Therefore, HDAC6 has become an attractive target for cancer therapy, offering a more selective approach that could reduce side effects compared to other broad-spectrum HDAC inhibitors [2]. Interest in HDAC6 continues to grow, with ongoing studies exploring its potential in new treatment strategies for various cancers.

Despite the increasing focus on HDAC6, the development of effective inhibitors still faces several challenges. While some selective HDAC6 inhibitors have entered clinical trials, none have received approval yet. One major problem is their limited

efficacy when used as monotherapies, often requiring higher doses, which in turn increases the risk of adverse effects. Additionally, issues such as low bioavailability, inadequate binding affinity, and drug resistance further complicate their development. To overcome these challenges, researchers are exploring alternative strategies, including combination therapies, dual-target inhibitors, and HDAC6-targeting degraders like PROTACs [3]. Nevertheless, the drug design process typically spans several months to years and demands substantial resources. To accelerate this workflow and reduce the cost and complexity of early-stage discovery, computational approaches are increasingly being used.

Particularly, machine learning (ML) and artificial intelligence (AI) are transforming the discovery of drug by streamlining drug identification and optimization. AI-driven models can analyze large chemical datasets, predict drug-target interactions, refine drug properties and identify promising compounds. The increasing availability of high-quality, digitized chemical data has further enhanced the applicability of AI across early-stage drug discovery. Additionally, recent developments in chemical informatics like machine-readable representations of molecular structures have improved how compounds are retrieved and analysed. Importantly, even in data-constrained settings, ML models can extract meaningful chemical patterns and generalize from limited examples. This is particularly advantageous in HDAC6 inhibitor research, where available data is not large [4, 5]. One of the key limitations of AI-driven drug discovery was limitation of interpretability, hindering researchers' ability to validate model outputs [6].

To address this, SHAP (SHapley Additive exPlanations) has emerged as a key tool for interpreting ML model predictions. By assigning importance values to features, SHAP helps researchers understand how different molecular properties affect the outputs. It breaks down predictions into local explanations, which clarify individual results, and global explanations, which highlight overall patterns in the data. By applying SHAP analysis, researchers can identify key structural characteristics that influence HDAC6 inhibition, refine model predictions, and enhance confidence in AI-generated results [7, 8].

In this study, we aim to identify potent HDAC6 inhibitors using machine learning and explainable AI, specifically SHAP. Applying computational methods, we seek to dive into its molecular insights, making the process more interpretable. To achieve this, SHAP analysis will be applied to the best-performing model to highlight key molecular features that lead to HDAC6 inhibition. This approach not only improves predictive accuracy but also provides clear insights into the underlying chemical properties that influence inhibition.

Additionally, SAHA (suberoylanilide hydroxamic acid) will be used as a reference compound to classify the compounds as active or inactive rather than based on a cut-off value. As a well-established HDAC inhibitor, SAHA serves as a benchmark for assessing new candidates, ensuring a structured comparison of their effectiveness.

## **MATERIALS AND METHODS**

### **DATA COLLECTION**

#### *Data Collecting*

The dataset for HDAC6 inhibitors was collected from two publicly available

databases, ChEMBL version 33 and BindingDB (UCSD) (detailed in SI.1). These databases collect bioassay results from published studies and patents, providing a reliable source of compounds activity data. The dataset consists of IC<sub>50</sub> values, which represent the concentration of a compound required to inhibit 50% of HDAC6 activity. Only IC<sub>50</sub> values were included, while other activity metrics like Ki or EC50 were excluded to maintain consistency in the dataset.

To ensure data quality, several preprocessing steps were applied. First, all compounds associated with bioassays targeting human HDAC6 were extracted from ChEMBL and BindingDB. Compounds missing IC<sub>50</sub> values or having undefined IC<sub>50</sub> values were removed. For each compound, the IC<sub>50</sub> value of SAHA from the same bioassay was used as a reference compound, allowing for a direct comparison under identical experimental conditions. Compounds were then classified as “active” or “inactive” based on their relative IC<sub>50</sub> values compared to SAHA in the same assay. Those with IC<sub>50</sub> values less than or equal to the reference were labeled as “active”, while those with higher IC<sub>50</sub> values were labeled as “inactive”. Additionally, compounds that could not be encoded using the RDKit library for fingerprint conversion were removed. To avoid redundancy, duplicate compounds appearing in databases were identified and removed. After completing these preprocessing steps, the final dataset contained 1,224 HDAC6 inhibitors.

#### *Training and Test Data Splitting*

After preprocessing, the dataset was randomly split into training and test sets at 80:20 ratio. Using 80% of the data for training ensures that the model has enough information to learn important patterns, while 20% for testing provides a sufficient sample

to assess how well the model performs on unseen data. To ensure that the data is split the same way in every run, a random seed (`random_state = 42`) was applied [9]. The training set was used to build the model, also when applying cross-validation, where it served both as the training set and the validation set for hyperparameter tuning. Meanwhile, the test set was reserved exclusively for final evaluation, ensuring an unbiased assessment of the model's performance on unseen data. The number of compounds and class distribution across the training and test sets were showed in Table 1.

**Table 1.** Distribution of Active and Inactive Compounds in Training and Test Sets

Dataset	Training Set	Test Set	Total
Active	490	122	612
Inactive	489	123	612
<b>Total</b>	979	245	1.224

### MODEL TRAINING AND EVALUATION

The model training and evaluation process involved several key steps, including feature representation, model selection, hyperparameter tuning, and performance evaluation.

For molecular representation, two types of descriptors were used: MACCS Key (166-bit) and ECFP2. MACCS keys are fixed-length binary fingerprints based on predefined structural patterns, while ECFP2 is a circular fingerprint that captures molecular substructures by considering atom neighborhoods within a radius of 1 [10-12]. Since MACCS keys are simpler and more interpretable, and ECFP2 provides more detailed molecular information, both were tested to clarify which worked better for the models. These fingerprints were generated using RDKit, a widely used cheminformatics library [13, 14].

Five machine learning models were tested: Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), Extreme Gradient Boosting (XGBoost), and Adaptive Boosting (AdaBoost). Each model applies a different approach to learning patterns in the data. All were implemented using scikit-learn and the XGBoost library [15, 16].

DT is a simple and interpretable model that splits data based on feature values, forming a tree-like structure where each path leads to a classification [17]. RF enhances this approach by combining multiple decision trees, training each on different parts of the data, and averaging their predictions to improve accuracy and reduce overfitting [18]. SVM works by finding the best decision boundary between two classes, making it effective for classification tasks [19]. XGBoost is an efficient boosting algorithm that builds decision trees sequentially, with each tree correcting previous errors to improve predictions while minimizing a loss function [20]. Meanwhile another boosting algorithm, AdaBoost, builds a strong classifier by sequentially training weak learners, increasing the weight of misclassified instances in each iteration to improve accuracy [21].

Since the initial models without hyperparameter tuning did not perform very well, a tuning process was applied to optimize each model's performance. For DT model, a grid search was used to determine the best values for tree depth (`max_depth`), minimum split size (`min_samples_split`), and minimum leaf size (`min_samples_leaf`). The search covered `max_depth` values from 5 to 50 in steps of 5, `min_samples_split` from 2 to 20, and `min_samples_leaf` from 1 to 10. Different values were tested, and the best combination was selected using 5-fold cross-validation.

For RF model, RandomizedSearchCV was applied to efficiently explore a broader range of values. A total of 100 random combinations were tested, each evaluated with 5-fold cross-validation, with the number of trees (`n_estimators`) ranging from 100 to 1000 in steps of 50, tree depth (`max_depth`) from 10 to 50 in steps of 5, minimum split size (`min_samples_split`) from 2 to 20, and minimum leaf size (`min_samples_leaf`) from 1 to 10. The number of features considered per split (`max_features`) was set to 'sqrt', 'log2', or None, and both bootstrapped and non-bootstrapped models (`bootstrap`: True or False) were tested.

For SVM model, a grid search was performed to optimize the regularization parameter (`C`) and kernel type (`kernel`). Model was tested with linear, radial basis function (`rbf`), and polynomial kernels. The search included `C` values of 0.01, 0.1, 1, 10, and 100. For polynomial kernels, the degree (`degree`) was varied from 2 to 5. Stratified 5-fold cross-validation was used to ensure a balanced class distribution across training and validation sets.

For XGBoost model, hyperparameter tuning focused on the number of boosting rounds (`n_estimators`), tree depth (`max_depth`), learning rate (`learning_rate`), data subsampling per boosting round (`subsample`), and feature subsampling per tree (`colsample_bytree`). The tested ranges included `n_estimators` from 100 to 1000 in steps of 50, `max_depth` from 3 to 15 in steps of 2, `learning_rate` values of 0.01, 0.05, 0.1, and 0.2, `subsample` from 0.5 to 1.0 in steps of 0.1, and `colsample_bytree` from 0.5 to 1.0 in steps of 0.1. Due to computational constraints, 3-fold cross-validation was used instead of 5-fold.

For AdaBoost models, Randomized-

SearchCV was used to explore 100 different parameter combinations, testing the number of weak learners (`n_estimators`) from 50 to 500 in steps of 50, learning rate (`learning_rate`) values of 0.01, 0.05, 0.1, 0.2, and 1.0, and boosting algorithm (`algorithm`) choices between 'SAMME' and 'SAMME.R'. Each model was evaluated using 5-fold stratified cross-validation.

Hyperparameter tuning was executed on a dual-socket workstation (2 × Xeon® Platinum 8180, 56 physical cores, 256 GB RAM, RTX 4080 SUPER, Samsung 990 PRO NVMe SSD 1T). To keep computational cost manageable we (i) employed RandomizedSearchCV with 100 draws for both Random Forest and AdaBoost, replacing exhaustive grids; (ii) reduced XGBoost cross-validation to 3-fold; and (iii) enabled full CPU parallelisation (`n_jobs` = -1). These measures limited the total compute to approximate 600 core-hours, while delivering the optimal settings reported in Tables 3 and 5.

To evaluate model performance, we used three key metrics: Accuracy, Precision, and AUC-ROC. Accuracy measures the proportion of correct predictions out of all predictions made, providing an overall assessment of correctness. Precision calculates the ratio of correctly predicted positive samples to the total predicted positives, helping to determine how reliable the model's positive predictions are. A higher precision indicates fewer false positives, meaning how well the model avoids false positives. AUC-ROC (Area Under the Receiver Operating Characteristic Curve) evaluates how well the model separates active and inactive compounds by comparing true positive and false positive rates at different thresholds [22].

After hyperparameter tuning, each model was further tested with different fingerprint representations, including Morgan2 [23], ECFP4, and ECFP6. Together with the initial models using MACCS Keys and ECFP2, this resulted in a total of 25 models. The best-performing model was selected for further interpretation then we used SHAP analysis to understand which molecular features were most important for classification.

### **SHAP VALUE ANALYSIS ON THE BEST MODEL**

After hyperparameter tuning and performance evaluation, the best model was selected to conduct SHAP analysis. The selection of the best model was based on two primary factors: performance metrics and confusion matrix analysis.

First, the chosen model should achieve strong performance on the test set, with Accuracy, Precision, and AUC-ROC all exceeding 90% while the difference between training and test performance metrics did not exceed 8% to reduce the risk of overfitting.

Second, confusion matrices were employed to further evaluate model performance. A confusion matrix is a table that compares predicted labels with actual labels, consisting of four key components: (1) True Positives (TP) - The number of active compounds correctly predicted as active; (2) True Negatives (TN) - The number of inactive compounds correctly predicted as inactive; (3) False Positives (FP) - The number of inactive compounds misclassified as active; (4) False Negatives (FN) - The number of active compounds misclassified as inactive. The primary criterion for model selection was the highest number of true positives (TP), as it reflects the model's accuracy in correctly identifying active compounds. Models were also evaluated

based on their false positives (FP), with a preference for those that minimized FP to reduce the risk of false classifications.

To further understand how the model makes predictions, SHAP analysis was used. SHAP values provide a way to measure the contribution of each feature to an individual prediction, making it possible to interpret the model's decision-making process. A SHAP value is the Shapley value of cooperative-game theory applied to features: it measures how much a given feature shifts the prediction for one molecule away from the dataset average. SHAP values satisfy three axioms that make them suitable for model interpretation: (i) local accuracy-the sum of all feature contributions exactly equals the model prediction; (ii) missingness-a feature that is absent (e.g., a fingerprint bit = 0) has zero contribution; and (iii) consistency-if a feature becomes more influential in a revised model, its SHAP value cannot decrease. In our context each fingerprint bit is treated as an independent "player," so positive SHAP values indicate that the presence of that substructure increases the predicted probability of HDAC6 inhibition, whereas negative values indicate the opposite. In this study, since the best-performing model was a Random Forest classifier, the TreeExplainer function from the SHAP library was used [24]. TreeExplainer is optimized for tree-based models, ensuring more accurate and efficient calculations of SHAP values. The mean absolute SHAP value over the test set was used to rank features in the summary plot shown in Figure 5.

The SHAP analysis was conducted on the test set, where SHAP values were computed for each data point. To visualize these values, a SHAP summary plot was generated. The SHAP summary plot provides a global

overview of how each feature contributes to the classification outcomes by displaying SHAP values for all samples in the dataset. Each feature is ranked by its mean absolute SHAP value, meaning that the most influential features appear at the top, while less important ones are ranked lower.

In the summary plot, each point represents a single prediction observation. The position of a point along the x-axis indicates the SHAP value, which reflects both the magnitude and direction of the feature's influence on the model's prediction. A positive SHAP value indicates that the feature increases the likelihood of a compound being classified as active, whereas a negative SHAP value suggests it pushes the prediction toward the inactive class. Meanwhile, the color of the point encodes the actual value of the feature for that specific data point. Because the fingerprint bits in our study are binary indicators, a red hue denotes the feature's presence (bit = 1), whereas a blue hue signifies its absence (bit = 0). Features are ranked from top to bottom based on their mean absolute SHAP value. Features that appear at the top are

considered to be more influential in the model's decision-making process.

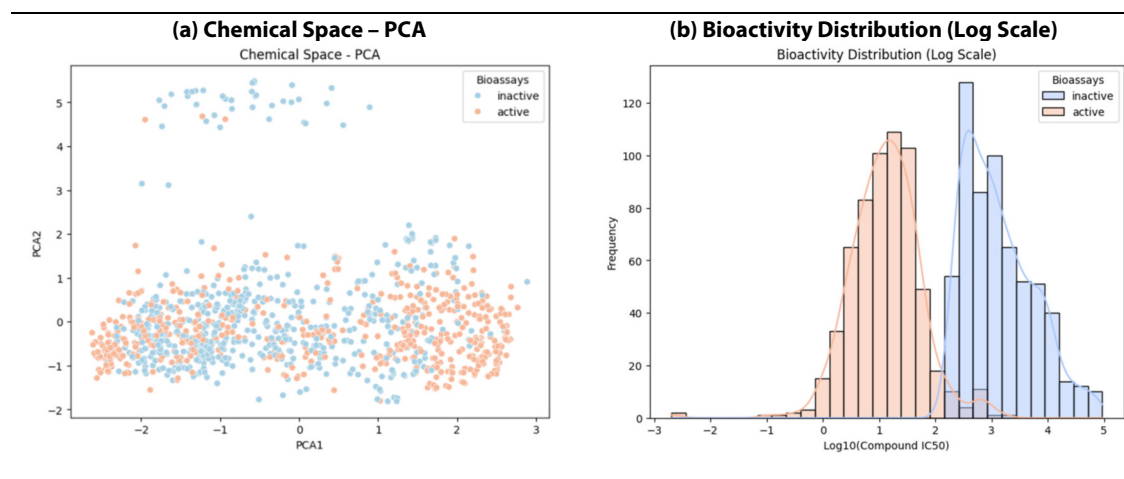
We exploited SHAP analysis to examine whether our machine learning predictions from the best-performing model align with experimental results. This approach allowed us to identify the structural fragments most influential in determining whether a compound is classified as active or inactive. By analyzing SHAP values, we can assess our best model's correctness, reliability and effectiveness.

## RESULTS AND DISCUSSIONS

### DATA ANALYSIS

We analyzed chemical space and bioactivity distribution to evaluate the structural diversity and data distribution of the dataset. The results are shown in the figures below.

The chemical space of the dataset was visualized using Principal Component Analyses (PCA). PCA helps simplify complex molecular data by reducing it to two main components, making it easier to see overall pattern [25]. The PCA results revealed



**Figure 1.** Chemical Space and Bioactivity Distribution of the Dataset

that although some overlap exists between active and inactive compounds, there are still visible areas where one group is more dominant (Figure 1a). The widespread distribution of data points across the PCA space indicates a diverse chemical dataset, which is beneficial for model generalization.

The bioactivity distribution was assessed by visualizing  $IC_{50}$  values on a logarithmic scale ( $\log_{10}$ ). Using a log scale improved the representation of the wide  $IC_{50}$  range, especially when large variations existed among compounds. The histogram showed that most compounds had  $IC_{50}$  values concentrated within a specific range, with "active" compounds exhibiting significantly lower  $IC_{50}$  values than "inactive" ones. Although there is a small overlap, the histogram shows two distinct peaks, indicating a clear separation between the two groups – active and inactive - in the dataset (Figure 1b).

To further make sure the models were trained on a diverse set of molecular structures, we used the Tanimoto coefficient ( $T_c$ ) to measure how similar or different the compounds are.  $T_c$  ranges from 0 to 1, where a value close to 1 means two compounds are very similar, and a lower value means they are more different [26, 27]. This analysis helps confirm that our dataset includes a wide variety of structures, making the model more reliable for different cases (Figure 2).

The Figure 2 presents the distribution of Tanimoto similarity ( $T_c$ ) values for pairwise comparisons within the dataset using five different molecular fingerprints: MACCS keys, Morgan2, ECFP2, ECFP4, and ECFP6. Each histogram represents the frequency distribution of  $T_c$  values, helping us understand the structural diversity of the dataset based on different fingerprinting methods.

The similarity distribution varies depending on the fingerprint used. MACCS Keys exhibit the broadest and most evenly spread distribution, with values ranging from 0.2 to 0.8, indicating a mix of both structurally similar and diverse compounds (Figure 2a). In contrast, ECFP2 and Morgan2 display sharp peaks at mid-range similarity scores, suggesting that they capture fewer structural similarities than MACCS Keys and represent a more diverse chemical space (Figure 2b, 2c). Meanwhile, ECFP4 and ECFP6 exhibit highly asymmetric distributions, with most similarity scores being low, reflecting a higher level of structural diversity (Figure 2d, 2e).

These differences matter when selecting compounds for further evaluation. Fingerprints that produce high similarity scores in a narrow range might not be ideal for identifying new or diverse compounds. In contrast, fingerprints with lower similarity scores or a wider spread indicate a more diverse dataset, which can help identify novel compounds.

## **MODEL TRAINING AND EVALUATION RESULTS**

### *Performance of Default Models*

Table 2. Performance of Default Models

Most of the models exhibited near-perfect accuracy on the training set, often exceeding 97%, regardless of the fingerprint type. However, the performance on the test set dropped significantly (as low as ~76% on accuracy for SVM with ECFP2), indicating a strong overfitting effect, where the models learn specific patterns from the training data but fail to generalize effectively to unseen data.

AdaBoost generally underperformed compared to other models, particularly for

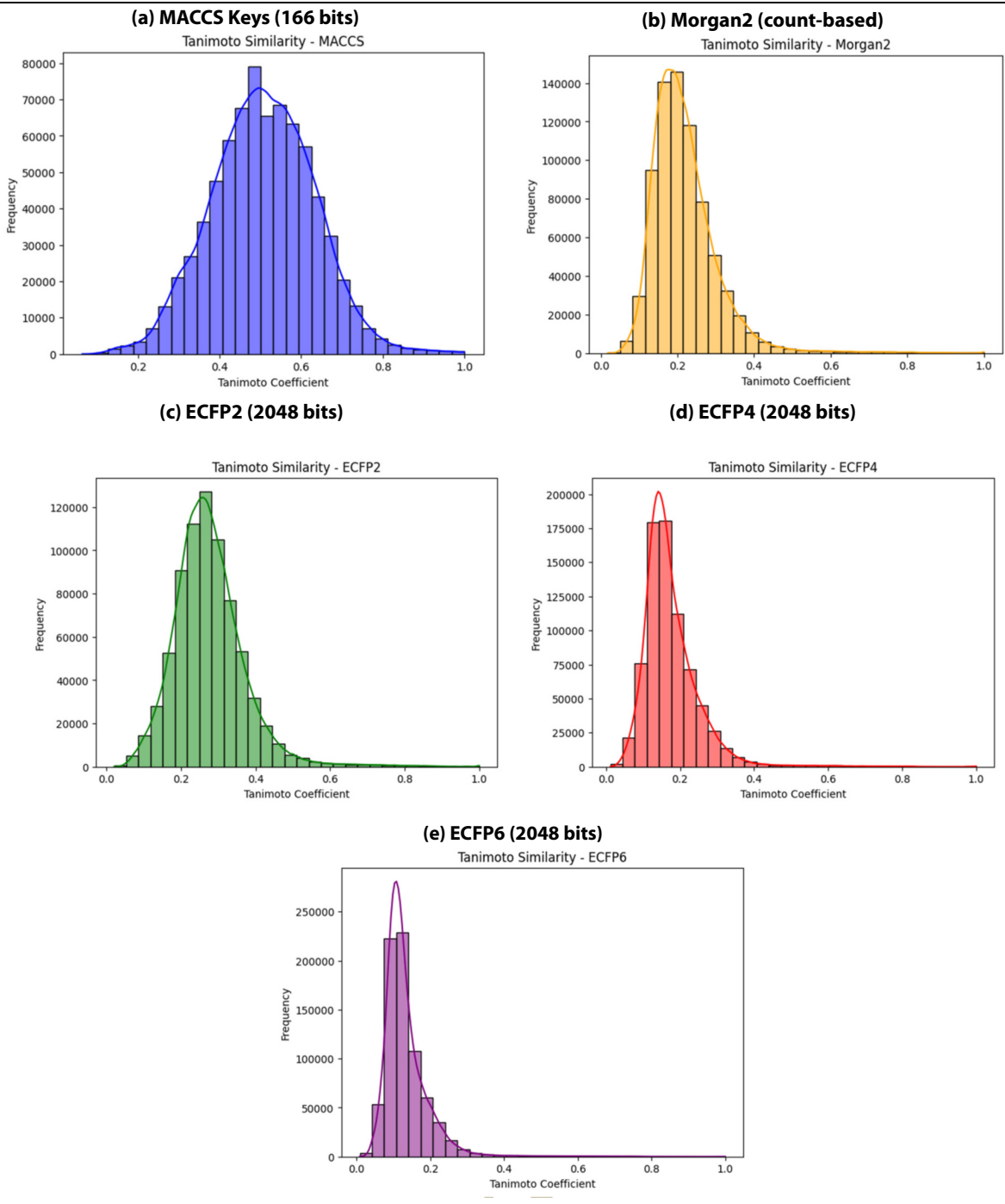


Figure 2. Pairwise Tc values of the dataset used to build the models using fingerprints.

**Table 2** presents the performance of the default machine learning models trained on two molecular representations (MACCS Keys and ECFP2).

Finger-print	Model	Training Set			Test Set		
		Accuracy	Precision	AUC-ROC	Accuracy	Precision	AUC-ROC
MACCS Keys	DT	98.77	99.59	99.97	83.67	85.34	85.57
	RF	98.77	98.38	99.91	85.71	87.83	94.04
	SVM	98.77	85.99	93.88	84.01	85.47	90.18
	XGBoost	98.77	98.19	99.95	86.53	89.38	93.40
	AdaBoost	77.94	76.76	84.27	80.41	79.84	83.93
ECFP2	DT	98.67	99.18	99.96	81.63	83.48	82.98
	RF	98.67	98.18	99.91	85.71	85.95	93.11
	SVM	97.65	97.36	99.04	75.92	76.92	85.05
	XGBoost	97.96	97.96	99.82	85.71	88.50	93.05
	AdaBoost	81.00	78.46	88.68	77.96	80.36	86.86

**Table 3.** Performance of Models After Hyperparameter Tuning

Finger-print	Model	Training Set			Test Set		
		Accuracy	Precision	AUC-ROC	Accuracy	Precision	AUC-ROC
MACCS Keys	DT	86.11	87.03	94.38	80.41	80.83	89.26
	RF	<b>97.34</b>	<b>96.59</b>	<b>99.7</b>	<b>86.94</b>	<b>88.14</b>	<b>94.58</b>
	SVM	89.79	88.09	96.71	82.86	82.79	91.67
	XGBoost	<b>98.77</b>	<b>98.38</b>	<b>99.95</b>	<b>87.35</b>	<b>90.99</b>	<b>93.96</b>
	AdaBoost	82.23	80.27	90.35	80.41	79.84	87.67
ECFP2	DT	87.54	87.40	95.13	82.04	81.97	87.55
	RF	<b>93.77</b>	<b>92.48</b>	<b>98.59</b>	<b>86.94</b>	<b>86.89</b>	<b>93.58</b>
	SVM	92.13	90.73	97.48	81.63	85.32	90.34
	XGBoost	<b>97.75</b>	<b>97.18</b>	<b>99.83</b>	<b>86.12</b>	<b>87.93</b>	<b>93.83</b>
	AdaBoost	86.01	84.68	92.84	80.00	82.88	89.24

MACCS Keys, where it had the training accuracy of 77.94% and test accuracy (80.41%).

*Performance After Hyperparameter Tuning*

As we have discussed before, the models

showed signs of overfitting. When a model overfits, it memorizes noise and dataset-specific details that do not translate to unseen data, leading to high false positive and false negative rates, which result in lower metrics on the test set. To address this problem, we

conducted hyperparameter tuning to optimize the models, reduce overfitting, and enhance their generalization ability. The results are shown in the Table 3, presenting the performance of the models after applying hyperparameter tuning. The best hyperparameters for each model are detailed in Table 4.

It can be seen from Table 3 that the performance gap between the training and test sets was reduced, meaning that the models have generalized better. Moreover, the metrics of the test set slightly increased across most models, indicating that overfitting was minimized to some extent.

For DT, overfitting was notably reduced. Using MACCS Keys, the training-test accuracy gap dropped from 15.1% to 5.7%. A similar trend can be seen for ECFP2, where the gap decreased from 17.04% to 5.50%, confirming that tuning made the model less prone to overfitting.

For SVM, tuning significantly improved its test set performance. With MACCS Keys, the training-test accuracy gap decreased from 14.76% to 6.93%. For ECFP2, the gap shrank substantially from 21.73% to 10.50%.

Ensemble models, RF and XGBoost, also showed improved generalization. Using MACCS Keys, RF's accuracy gap decreased from 13.06% to 10.40%, while XGBoost's gap saw a marginal reduction from 12.24% to 11.42%. With ECFP2, the improvement was more noticeable for RF, with the gap decreasing from 12.96% to 6.83%, and XGBoost saw a reduction from 12.25% to 11.63%.

For AdaBoost with MACCS Keys, the initially small accuracy gap dropped further from 2.53% to 1.82%. In the case of ECFP2, although the gap between the training and test

sets did not significantly narrow, it remained relatively acceptable at 6.01%.

These results highlight the importance of optimizing hyperparameters to achieve reliable predictive performance in bioactivity classification for novel compounds.

On the other hand, comparing between models, RF and XGBoost consistently delivered the highest test set performance across all evaluation metrics.

In terms of accuracy, RF achieved 86.94% with MACCS Keys and 86.94% with ECFP2, while XGBoost reached 87.35% and 86.12%, respectively. These values were notably higher than those of DT (80.41% for MACCS Keys, 82.04% for ECFP2) and SVM (82.86% for MACCS Keys, 81.63% for ECFP2), demonstrating the enhanced predictive capability of ensemble methods.

Precision further highlights the advantage of RF and XGBoost. Using MACCS Keys, RF obtained a precision of 88.14%, higher than SVM (82.79%) and DT (80.83%), while XGBoost performed even better at 90.99%. The pattern held for ECFP2, where RF (86.89%) and XGBoost (87.93%) outperformed SVM (85.32%) and DT (81.97%).

Looking at AUC-ROC, RF and XGBoost outperformed the other classifiers. With MACCS Keys, RF achieved an AUC-ROC of 94.58, while XGBoost followed closely at 93.96, both surpassing SVM (91.67) and DT (89.26). A similar trend was observed for ECFP2, where RF (93.58) and XGBoost (93.83) exceeded the performance of SVM (90.34) and DT (87.55).

Overall, RF and XGBoost emerged as the most reliable models with strong metrics while overfitting was minimized.

#### *Fingerprint Comparison*

Table 5 extends Table 3 by incorporating

**Table 4.** Best Hyperparameter of Each Model

Fingerprint	Model	Best Hyperparameter
MACCS Keys	DT	'max_depth': 15, 'min_samples_leaf': 10, 'min_samples_split': 2
	RF	'bootstrap': False, 'max_depth': 30, 'max_features': 'sqrt', 'min_samples_leaf': 2, 'min_samples_split': 2, 'n_estimators': 509
	SVM	'svc_C': 1, 'svc_kernel': 'rbf'
	XGBoost	'colsample_bytree': 0.9, 'learning_rate': 0.1, 'max_depth': 6, 'n_estimators': 300, 'subsample': 0.8
	AdaBoost	'algorithm': 'SAMME', 'learning_rate': np.float64(0.8065429868602328), 'n_estimators': 1344
ECFP2	DT	max_depth': None, 'min_samples_leaf': 5, 'min_samples_split': 20
	RF	'bootstrap': False, 'max_depth': 30, 'max_features': 'sqrt', 'min_samples_leaf': 2, 'min_samples_split': 2, 'n_estimators': 509
	SVM	'svc_C': 1, 'svc_kernel': 'rbf'
	XGBoost	'colsample_bytree': 0.8, 'learning_rate': 0.1, 'max_depth': 6, 'n_estimators': 300, 'subsample': 0.8
	AdaBoost	'algorithm': 'SAMME', 'learning_rate': np.float64(0.8065429868602328), 'n_estimators': 1344

**Table 5.** Performance of Hyperparameter-Tuned Models with Different Fingerprints

Finger-print	Model	Training Set			Test Set		
		Accuracy	Precision	AUC-ROC	Accuracy	Precision	AUC-ROC
Morgan 2	DT	91.52	91.79	98.07	84.49	85.59	88.80
	RF	93.77	92.48	98.59	86.94	86.89	93.58
	SVM	91.93	90.69	97.23	81.63	84.68	90.38
	XGBoost	97.55	97.55	99.73	89.80	93.69	95.99
	AdaBoost	89.68	88.98	95.99	84.90	87.61	91.17
ECFP4	DT	90.50	91.27	97.50	86.94	89.47	91.93
	RF	<b>97.24</b>	<b>96.77</b>	<b>99.67</b>	<b>89.80</b>	<b>90.08</b>	<b>96.27</b>
	SVM	98.98	98.58	99.93	86.53	87.39	91.72
	XGBoost	<b>97.85</b>	<b>96.99</b>	<b>99.84</b>	<b>89.39</b>	<b>90.68</b>	<b>96.38</b>
	AdaBoost	88.05	87.83	95.97	84.08	86.73	92.69
ECFP6	DT	86.31	87.08	94.60	84.08	84.87	88.53
	RF	<b>98.06</b>	<b>97.77</b>	<b>99.85</b>	<b>90.20</b>	<b>91.53</b>	<b>96.25</b>
	SVM	96.63	96.35	99.63	86.53	88.70	92.10
	XGBoost	<b>98.57</b>	<b>98.37</b>	<b>99.90</b>	<b>89.80</b>	<b>92.17</b>	<b>96.35</b>
	AdaBoost	90.70	90.14	97.84	84.90	85.71	91.00

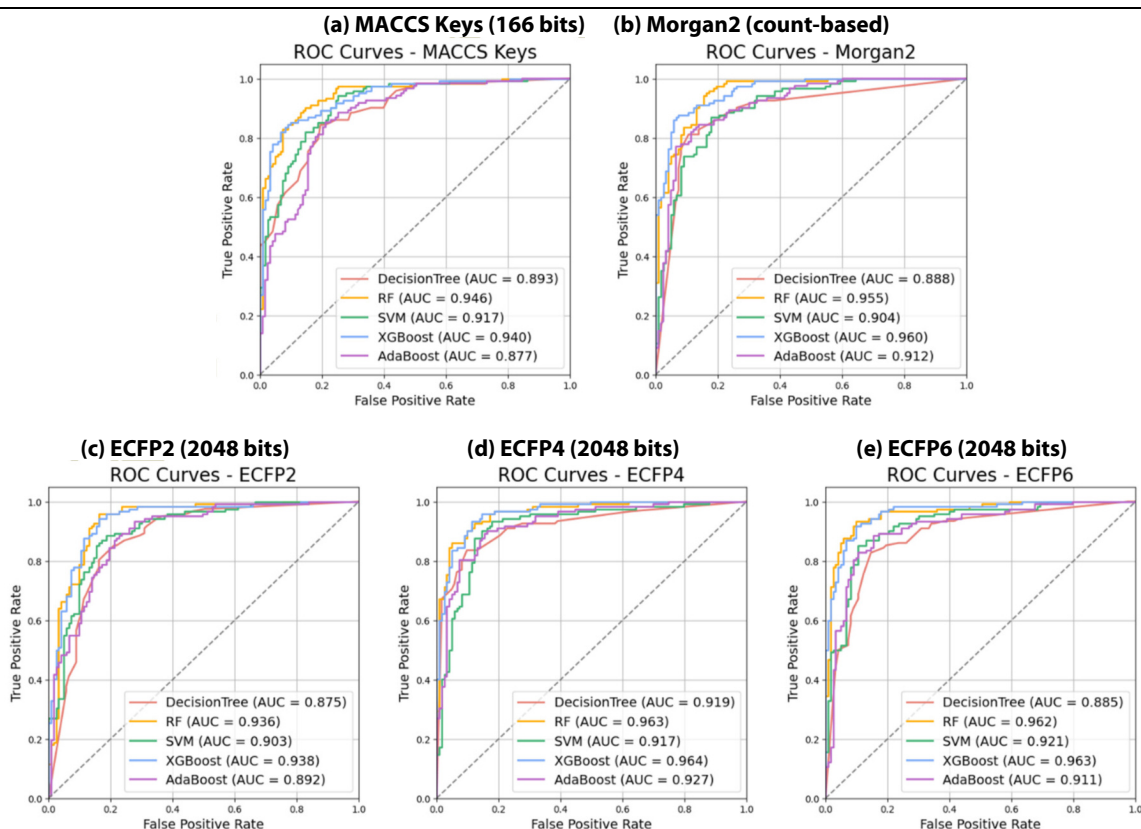


Figure 3. ROC-Curves of the Fingerprints

three additional fingerprints namely Morgan 2, ECFP4, and ECFP6. These tables compare the performance of hyperparameter-tuned models across different fingerprints on both the training and test sets. Meanwhile, Figure 3 illustrates the ROC curves, emphasizing the influence of fingerprint selection on model performance.

From what we can see, ECFP4 and ECFP6 outperform other fingerprints across most models, particularly in AUC-ROC and accuracy. With these fingerprints, RF and XGBoost consistently achieve the highest AUC scores on both training and test sets. Notably, ECFP4 reaches a test AUC of 96.38% in XGBoost and 96.27% in RF, while ECFP6 achieves 96.35% and 96.25%,

respectively. On the other hand, ECFP2 underperforms compared to its ECFP families, especially in generalization to the test set, with a slight reduction in metrics. This difference may be due to the fact that ECFP4 and ECFP6 use larger radii ( $r = 2$  and  $r = 3$ , respectively) compared to ECFP2 ( $r = 1$ ). Larger radii allow the fingerprints to capture bigger and more informative molecular substructures, which is very beneficial for identifying relevant patterns associated with HDAC6 inhibition.

Morgan2 performs slightly worse than ECFP variants but better than ECFP2 and MACCS Keys. RF and XGBoost still achieve strong results with Morgan2, yielding AUC scores of 95.99% (XGBoost) and 93.58% (RF).

MACCS Keys stays behind in most cases, particularly in test set performance. While MACCS Keys - RF and MACCS Keys - XGBoost still perform well (AUC of 94.58% and 93.96% respectively), MACCS Keys - DT and MACCS Keys - AdaBoost show quite lower test AUC (89.26% and 87.67%, respectively).

Overall, ECFP4 and ECFP6 provide the most optimal performance, particularly with RF and XGBoost. Morgan2 remains strong also but slightly behind, while MACCS Keys and ECFP2 demonstrate weaker predictive power, especially on the test set.

### **SHAP VALUE ANALYSIS RESULTS ON THE BEST MODEL**

#### *Choosing the Best Model*

The best model was chosen based on key metrics and confusion matrix analysis.

First, the models were evaluated using accuracy, precision, and AUC-ROC to identify the top performers. As previously mentioned, RF and XGBoost emerged as the top performers, with ECFP4 and ECFP6 surpassing other fingerprints in performance. After further review of the evaluation metrics in Tables 3, 5, we have selected the ECFP6 - RF model as the best performer, as explained below:

Both RF and XGBoost achieved high accuracy, precision, and AUC-ROC (>90%). However, ECFP6 - RF demonstrated better generalization, with a smaller gap between training and test performance. It achieved a test accuracy of 90.20%, precision of 91.53%, and AUC-ROC of 96.25%, while its training accuracy was 98.06%, resulting in a 7.86% gap, which remains within an acceptable range (<8%) to minimize overfitting.

Although XGBoost attained comparable or slightly higher test metrics, it exhibited a

larger train-test performance gap (accuracy gap >8%). Similarly, ECFP4 performed well with both RF (89.80% accuracy, 90.08% precision, 96.27% AUC-ROC) and XGBoost (89.80% accuracy, 92.17% precision, 96.35% AUC-ROC), but still slightly underperformed comparing to ECFP6.

Given the need for a model that maximizes predictive performance while maintaining minimal overfitting, ECFP6 - RF was the best option.

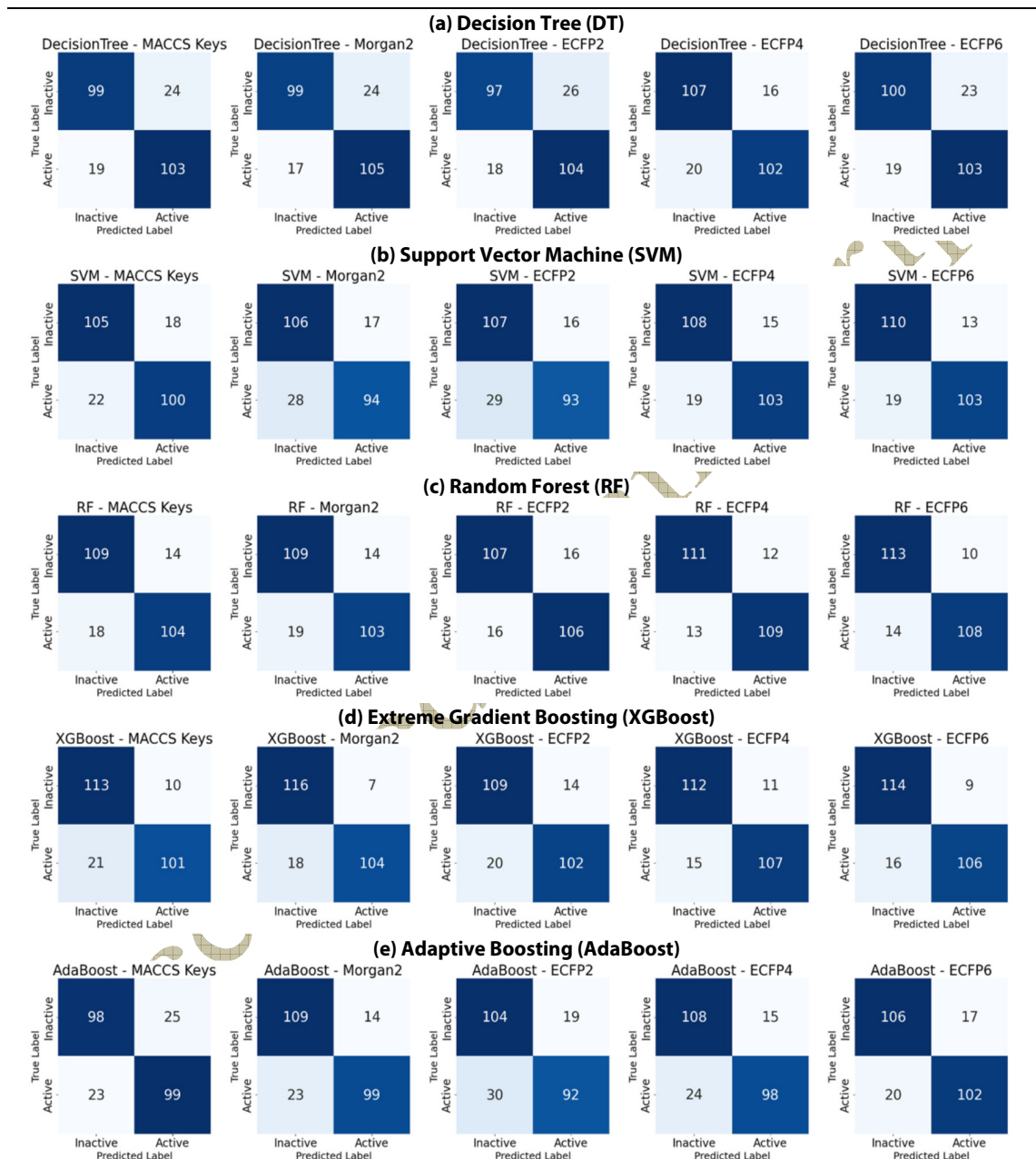
Second, the confusion matrix was analyzed for each model to validate the selection. The results confirm that the ECFP6 - RF model provides the most balanced performance in distinguishing active and inactive compounds compared to other models.

The models with the highest true positives (TP, active compounds correctly predicted as active) are ECFP4 - RF (TP = 109), ECFP6 - RF (TP = 108), and ECFP4 - XGBoost (TP = 107). While ECFP4 - RF has the highest TP, ECFP6 - RF remains the better choice. Although its TP is slightly lower, it compensates with a lower false positive (FP) rate (FP = 10 compared to FP = 12 in ECFP4 - RF). This means fewer inactive compounds are misclassified as active. This reduction in false positives, coupled with a relatively high TP, makes ECFP6 - RF the optimal model for minimizing errors.

In general, ECFP6 - RF remains the best choice, offering both high test set performance and a controlled gap between training and test metrics while maintaining a balanced classification between active and inactive compounds.

#### *SHAP Value Analysis Results*

The results of our SHapley Additive exPlanations (SHAP) analysis are presented in Figure 5, which highlights the most

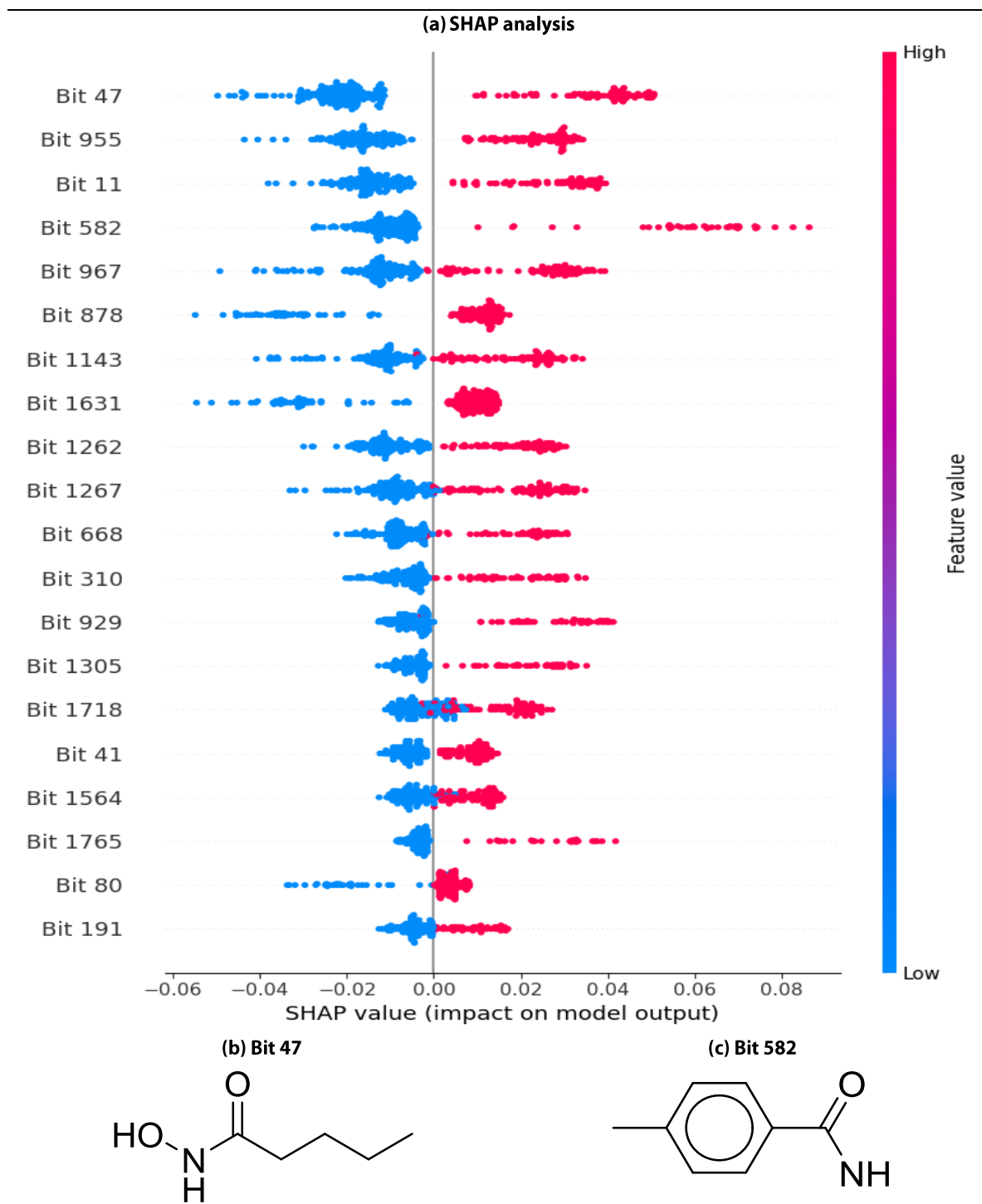


**Figure 4.** The Confusion Matrices of the Models

influential molecular fingerprints identified by the model chosen as having the best performance.

The SHAP summary plot in Figure 5a provides insights into how individual

molecular fingerprints contribute to the model's predictions. Bit 47, bit 955, and bit 11 have the highest SHAP values indicated by their red dots, meaning that these molecular fragments strongly contribute to



**Figure 5.** SHAP Value Analysis Results on the Best Model.

predicting a compound as active. This suggests that when these structural patterns are present in a molecule, the model is more confident in classifying it as active compound.

Some lower ranked features like bit 878 and bit 1631 show a clear distinction between positive and negative SHAP values. This means that these bits are also influential in classification by strongly pushing the prediction in one direction (either active or inactive) when they are present in a compound, rather than having a neutral or weak effect.

On the other hand, bit 1267 and bit 668 show a different pattern. Instead of having a strong, consistent impact in one direction, their SHAP values have a little mix in the middle. This suggests that their influence on classification depends on the context in which they appear. In some molecules, they may contribute to activity, while in others, they may push the prediction toward inactivity. This could be due to interactions with other molecular features that modify their effects.

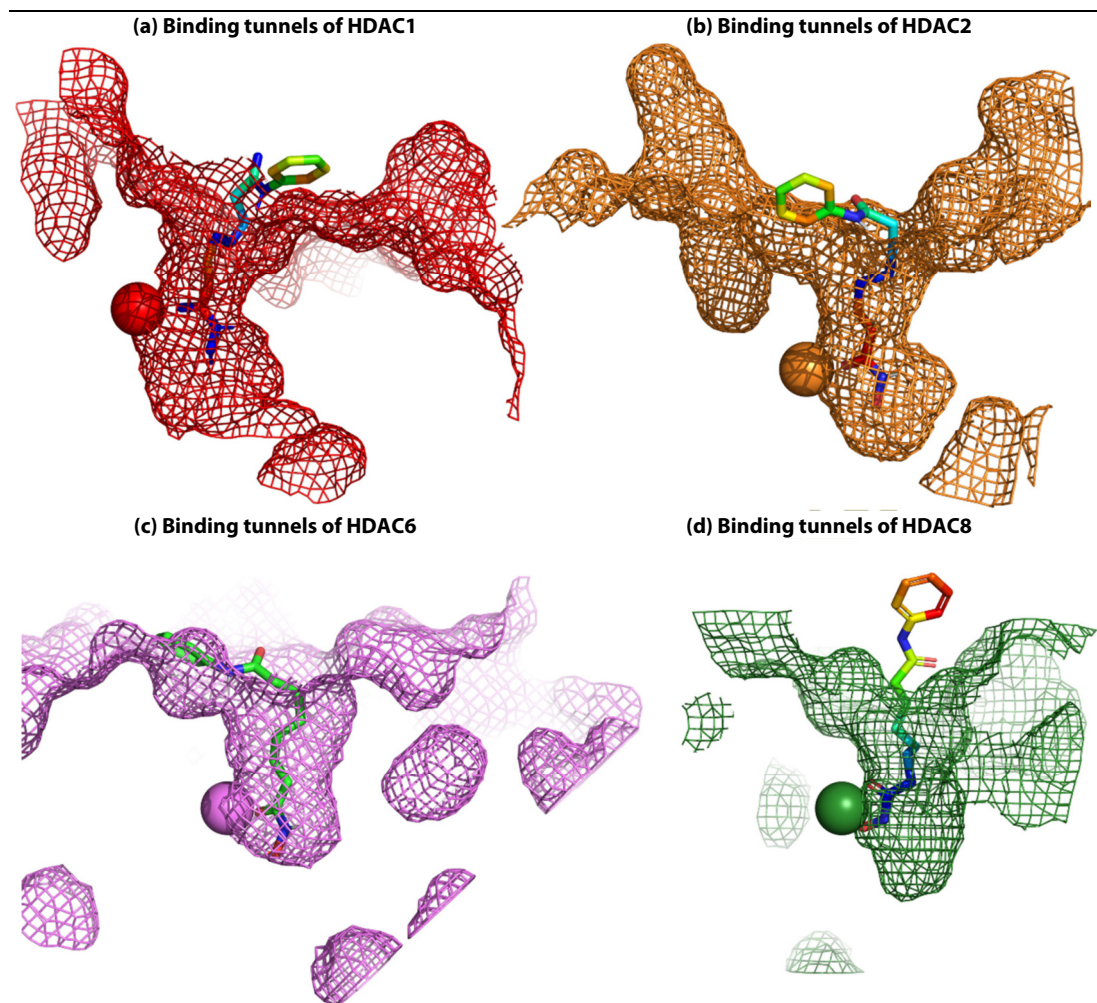
Interestingly, when we analyzed the top nine molecular fragments that contribute most to HDAC6 inhibition using SHAP, we found that all of them were related to the hydroxamic acid scaffold. This list is provided in Supplementary Information 3 (SI.3).

As each bit in ECFP fingerprint can represent either a full substructure or just a small part of it, some bits, such as the top 3, bit 47, bit 955, and bit 11, all captured the key hydroxamic acid group quite completely, including the  $-C(=O)-N(-OH)-$  part and the aliphatic chain, while others represented smaller parts of the same structure. Overall, this shows that our best model is focusing on the important parts of the structure that are key to predicting bioactivity.

These results match many previous studies that confirm hydroxamic acids are strong zinc-binding groups (ZBG). Hydroxamic acids form stable complexes with the zinc ion in HDACs, thereby exerting effective inhibitory activity against HDAC6 [28-30]. This is interesting because our machine learning model was never explicitly provided with structure-activity relationships information in the input data, yet it successfully identified the key scaffold driving HDAC6 inhibition. This highlights the model's ability to detect essential pharmacophores purely from data.

Among all identified fragments, the top-ranked feature by mean SHAP value is bit 47, indicating its strong global contribution. Compared to other bits, bit 47 shows more than just encoding the hydroxamic acid or ZBG ( $-C(=O)-N(-OH)-$ ), it also captures a short aliphatic linker. Prior studies indicate that aliphatic linkers in HDAC6 inhibitors are typically short, unbranched hydrocarbon chains ranging from 4 to 7 carbons, as this appears to fit into the binding pocket of HDAC6 [28, 31]. This means that the most influential bit, Bit 47, aligns with experimental results. This further reinforces the validity and potential of machine learning in predicting the biological properties of HDAC6 inhibitors.

Although bit 47 was the highest overall contribution to model predictions, the highest positive SHAP value corresponds to bit 582 (Figure 5). This discrepancy indicates a difference in how each bit contributes to model predictions: bit 47 exerts a broad and reliable influence across many compounds, while bit 582, though less frequent, causes a sharp increase in predicted activity whenever it appears. In other words, bit 47 functions as a common, foundational signal in the



**Figure 6.** Comparison of HDAC binding tunnel architecture.

prediction of active compounds, whereas bit 582 acts as a rare but potent activity enhancer. Recognizing both types of features is essential for rational drug design, as it enables researchers to balance generalizable scaffolds with high-impact structural motifs during lead optimization.

In details, bit 582 appears to encode a linker region that includes an aromatic ring, indicating it represents an aromatic linker. This is particularly noteworthy, as previous studies have shown that incorporating

aromatic linkers can significantly enhance the conformational stability of HDAC6 inhibitors. Due to the shallower and wider binding channel of HDAC6 compared to other isoforms (as we can also see from Figure 6), inhibitors with bulky and short aromatic linkers exhibit greater selectivity for HDAC6 amongst all HDAC compared to traditional alkyl chain, as this modification helps restrict the hydroxamate group's proximity to  $Zn^{2+}$  [32-34].

Another important aspect is that the

aromatic ring captured by bit 582 corresponds to a benzyl group, which has been shown to provide strong and selective HDAC6 inhibition. Compounds containing benzyl linkers often exhibit very low IC<sub>50</sub> values, indicating potent bioactivity [35]. This also helps explain why bit 582 shows such a strong influence in the SHAP analysis.

SHAP analysis confirms that our best-performing model successfully identifies structural features crucial to HDAC6 inhibition. The recognition of hydroxamic acid further supports the model's predictions and aligns with experimental data.

Beyond this specific case, our findings underscore the broader utility of machine learning in rational drug design. The ability of AI-driven models to pinpoint essential pharmacophores without prior knowledge of structure-activity relationships suggests that this approach can be extended to other drug targets. By identifying key molecular fragments, machine learning can facilitate the development of more selective and potent therapeutic agents. Our study reinforces the growing importance of AI-driven molecular analysis in early-stage drug discovery.

#### Abbreviations

AdaBoost: Adaptive Boosting

AI: Artificial Intelligence

AUC-ROC: Area Under the Receiver Operating Characteristic Curve

DT: Decision Tree

ECFP: Extended-Connectivity Fingerprints (ECFP2, ECFP4, ECFP6 used)

FN: False Negatives

FP: False Positives

HDAC(s): Histone Deacetylase(s)

HDAC6: Histone Deacetylase 6

IC<sub>50</sub>: Half maximal inhibitory concentration

MACCS: Molecular ACCESS System

ML: Machine Learning

PCA: Principal Component Analyses

RF: Random Forest

SAHA: Suberoylanilide Hydroxamic Acid

SHAP: SHapley Additive exPlanations

SVM: Support Vector Machine

Tc: Tanimoto coefficient

TN: True Negatives

TP: True Positives

XAI: Explainable AI

XGBoost: Extreme Gradient Boosting

ZBG: Zinc-Binding Group

#### CONCLUSION

This study successfully developed and optimized machine learning models for the classification of HDAC6 inhibitors based on their bioactivity relative. Through systematic evaluation and hyperparameter tuning, RF and XGBoost models, particularly when combined with ECFP4 and ECFP6 fingerprints, demonstrated superior predictive performance and generalization capabilities compared to DT, SVM, and AdaBoost models. The ECFP6-RF model was identified as the optimal choice, achieving excellent test set metrics (Accuracy >90%, AUC-ROC >96%) while maintaining a controlled gap between training and test performance, thereby minimizing the risk of overfitting.

Crucially, the application of SHAP analysis to the best-performing model provided valuable insights into its decision-making process, bridging the gap often associated with "black box" models. The SHAP analysis successfully identified molecular fragments critical for HDAC6 inhibition, prominently featuring substructures related to the hydroxamic acid moiety and distinct linker types. The model's ability to pinpoint these essential pharmacophores, consistent with established experimental findings and structure-activity

relationships, without prior explicit encoding of this knowledge, underscores the power of data-driven approaches.

Beside, the model relies solely on two-dimensional binary fingerprints, omitting three-dimensional conformational and physicochemical factors that are known to shape zinc-binding and, ultimately, HDAC6 inhibition. Future work will therefore incorporate graph-based or explicit 3-D descriptors and pair them with interaction-aware explainability methods-such as SHAP-GNN or Integrated Gradients-to capture spatial features and fragment cooperativity more faithfully, thereby broadening predictive reach and deepening mechanistic insight.

In summary, this research demonstrates the value of combining optimized machine learning models with explainable AI (SHAP) for HDAC6 inhibitor discovery. This approach enables efficient screening and offers interpretable insights for rational drug design, emphasizing the growing role of AI in developing novel therapeutics. The findings encourage further use of such models to identify effective and selective drug candidates for HDAC6 and other complex targets.

#### CONFLICTS OF INTEREST

None

#### REFERENCES

1. Zhang W, Ge L, Zhang Y, Zhang Z, Zhang W, Song F, et al. Targeted intervention of tumor microenvironment with HDAC inhibitors and their combination therapy strategies. *European Journal of Medical Research*. 2025;30(1).
2. Kaur S, Rajoria P, Chopra M. HDAC6: A unique HDAC family member as a cancer target. *Cell Oncol (Dordr)*. 2022;45(5):779-829.
3. Huang Z, Li L, Cheng B, Li D. Small molecules targeting HDAC6 for cancer treatment: Current progress and novel strategies. *Biomed Pharmacother*. 2024;178:117218.
4. Gupta R, Srivastava D, Sahu M, Tiwari S, Ambasta RK, Kumar P. Artificial intelligence to deep learning: machine intelligence approach for drug discovery. *Molecular Diversity*. 2021;25(3):1315-60.
5. Sarkar C, Das B, Rawat VS, Wahlang JB, Nongpiur A, Tiewsoh I, et al. Artificial Intelligence and Machine Learning Technology Driven Modern Drug Discovery and Development. *International Journal of Molecular Sciences*. 2023;24(3):2026.
6. Yang G, Ye Q, Xia J. Unbox the black-box for the medical explainable AI via multi-modal and multi-centre data fusion: A mini-review, two showcases and beyond. *Inf Fusion*. 2022;77:29-52.
7. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. *Advances in neural information processing systems*. 2017;30.
8. Ponce-Bobadilla AV, Schmitt V, Maier CS, Mensing S, Stodtmann S. Practical guide to SHAP analysis: Explaining supervised machine learning model predictions in drug development. *Clin Transl Sci*. 2024;17(11):e70056.
9. Scikit-learn. `sklearn.model_selection.train_test_split` 2025 [Available from: [https://scikit-learn.org/stable/modules/generated/sklearn.model\\_selection.train\\_test\\_split.html](https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html)].
10. Durant JL, Leland BA, Henry DR, Nourse JG. Reoptimization of MDL keys for use in drug discovery. *J Chem Inf Comput Sci*. 2002;42(6):1273-80.
11. Cereto-Massagué A, Ojeda MJ, Valls C, Mulero M, Garcia-Vallvé S, Pujadas G. Molecular fingerprint similarity search in virtual screening. *Methods*. 2015;71:58-63.
12. Rogers D, Hahn M. Extended-connectivity fingerprints. *J Chem Inf Model*. 2010;50(5):742-54.
13. Landrum G. MACCS keys implementation in RDKit 2010 [Available from: <https://github.com/rdkit/rdkit/blob/master/rdkit/Chem/MACCSkeys.py>].
14. RDKit: Open-source cheminformatics [Available from: <https://www.rdkit.org/>].

15. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. *J Mach Learn Res.* 2011;12(null):2825-30.
16. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; San Francisco, California, USA: Association for Computing Machinery; 2016. p. 785-94.
17. Che D, Liu Q, Rasheed K, Tao X. Decision Tree and Ensemble Learning Algorithms with Their Applications in Bioinformatics. In: Arabnia HR, Tran Q-N, editors. Software Tools and Algorithms for Biological Systems. New York, NY: Springer New York; 2011. p. 191-9.
18. Rigatti SJ. Random Forest. *Journal of Insurance Medicine.* 2017;47(1):31-9.
19. Noble W. What is a support vector machine? *Nat Biotechnol.* 2006;24:1565-7.
20. Wiens M, Verone-Boyle A, Henscheid N, Podichetty JT, Burton J. A Tutorial and Use Case Example of the eXtreme Gradient Boosting (XGBoost) Artificial Intelligence Algorithm for Drug Development Applications. *Clinical and Translational Science.* 2025;18(3).
21. Md Shahri NHH, Lai S, Mohamad M, Rahman H, Rambli A. Comparing the Performance of AdaBoost, XGBoost, and Logistic Regression for Imbalanced Data. *Mathematics and Statistics.* 2021;9:379-85.
22. Scikit-learn. sklearn.metrics - Metrics and scoring [Available from: <https://scikit-learn.org/stable/api/sklearn.metrics.html>].
23. Morgan HL. The Generation of a Unique Machine Description for Chemical Structures-A Technique Developed at Chemical Abstracts Service. *Journal of Chemical Documentation.* 1965;5(2):107-13.
24. Lundberg S. shap.TreeExplainer - SHAP documentation [Available from: <https://shap.readthedocs.io/en/latest/generated/shap.TreeExplainer.html>].
25. Elhaik E. Principal Component Analyses (PCA)-based findings in population genetic studies are highly biased and must be reevaluated. *Scientific Reports.* 2022;12(1).
26. Godden JW, Xue L, Bajorath J. Combinatorial preferences affect molecular similarity/diversity calculations using binary fingerprints and Tanimoto coefficients. *J Chem Inf Comput Sci.* 2000;40(1):163-6.
27. Rácz A, Bajusz D, Héberger K. Life beyond the Tanimoto coefficient: similarity measures for interaction fingerprints. *Journal of Cheminformatics.* 2018;10(1).
28. Wang XX, Wan RZ, Liu ZP. Recent advances in the discovery of potent and selective HDAC6 inhibitors. *Eur J Med Chem.* 2018;143:1406-18.
29. Rajak H, Singh A, Raghuvanshi K, Kumar R, Dewangan PK, Veerasamy R, et al. A structural insight into hydroxamic acid based histone deacetylase inhibitors for the presence of anticancer activity. *Curr Med Chem.* 2014;21(23):2642-64.
30. Sixto-López Y, Gómez-Vidal JA, De Pedro N, Bello M, Rosales-Hernández MC, Correa-Basurto J. Hydroxamic acid derivatives as HDAC1, HDAC6 and HDAC8 inhibitors with antiproliferative activity in cancer cell lines. *Scientific Reports.* 2020;10(1).
31. Segretti MCF, Vallerini GP, Brochier C, Langley B, Wang L, Hancock WW, Kozikowski AP. Thiol-Based Potent and Selective HDAC6 Inhibitors Promote Tubulin Acetylation and T-Regulatory Cell Suppressive Function. *ACS Medicinal Chemistry Letters.* 2015;6(11):1156-61.
32. Butler KV, Kalin J, Brochier C, Vistoli G, Langley B, Kozikowski AP. Rational design and simple chemistry yield a superior, neuroprotective HDAC6 inhibitor, tubastatin A. *J Am Chem Soc.* 2010;132(31):10842-6.
33. Ruzic D, Ellinger B, Djokovic N, Santibanez JF, Gul S, Beljkas M, et al. Discovery of 1-Benzhydryl-Piperazine-Based HDAC Inhibitors with Anti-Breast Cancer Activity: Synthesis, Molecular Modeling, In Vitro and In Vivo Biological Evaluation. *Pharmaceutics.* 2022;14(12):2600.
34. Yang F, Zhao N, Ge D, Chen Y. Next-generation of selective histone deacetylase inhibitors. *RSC Advances.* 2019;9(34):19571-83.
35. Ho Y-H, Wang K-J, Hung P-Y, Cheng Y-S, Liu J-R, Fung S-T, et al. A highly HDAC6-selective inhibitor acts as a fluorescent probe. *Organic & Biomolecular Chemistry.* 2018;16(42):7820-32.