Drug Discovery Study on HDAC8 Inhibitors Using Machine Learning

Atika Nurani

2023

Osaka University

1 Abstract

The use of machine learning in drug discovery has been quite popular in recent years and a lot of state-of-the-art models have been built to predict and find new molecules. However, the drug discovery dataset, like many other datasets, has a class imbalance problem. This class imbalance problem could make the exploration and search of chemical space for new drug molecules difficult. This work used SMOTE to combat the data imbalance problem on the HDAC8 dataset from ChEMBL to find a new inhibitor for HDAC8. HDAC8 is a histone deacetylase that plays an important role in many types of cancers and possesses zinc(II) ion on its active site, so the existence of zinc-binding group for HDAC8 is important and this work would focus on finding new candidates for HDAC8 inhibitor. A random forest model using SMOTE was built and the usage of SMOTE was found to be able to increase the precision, recall, and F1 score of the model, compared to using a normal distribution. Using this model, compounds from Osaka University Library were screened, and after filtering based on ADMET and docking score, compound 12 was found to have inhibition activity on HDAC8 with IC50 of around 842 nM. Compound 12 was also found to have inhibitory activity against HDAC1 and HDAC3 with IC50 of 38 µM and 12 µM respectively, making it selective towards HDAC8. This work shows that using SMOTE could help a model explore a wider chemical space and find new drug candidates.

2 Acknowledgement

I would like to express my gratitude for Prof.Takayoshi Suzuki and staffs, associate Professor Yukihiro Itoh, and assistant professor Yasunobu Yamashita and Yuri Takada for their insight and guidance on this project. I am very grateful for the opportunities to learn many things from them.

I would also like to express my gratitude for the member of Suzuki laboratory for their support and company throughout this journey.

Last but not least, I am also very grateful for my friends and family for being my support system, I definitely could not have done this without them.

"There is nothing more lonely than an action taken quietly on your own, and nothing more comforting than doing that same quiet action in parallel with fellow humans doing the same action, everyone alone next to each other."

(Joseph Fink and Jeffery Cranor, Welcome to Night Vale)

"They did what all scientists do when following proper scientific method: (1) hypothesis; (2) argument; (3) fight; (4) cry; (5) hug."

(Joseph Fink and Jeffery Cranor, It Devours!)

2023

Atika Nurani

Contents

List of Figures

List of Tables

3 Introduction

3.1 Epigenetics

Epigenetics is a field of study of stable and heritable change in gene expression or cellular phenotype that occurs without changes in DNA sequence [4]. The term was first coined in 1957 by Conrad Waddington who proposed the concept of epigenetic landscape to explain the process of changes in cell development [5]. Nowadays, research on epigenetic involves the study of post translational modifications in DNA and histone proteins and how these changes influence chromatin structure and gene regulation.

3.2 Histone deacetylases

One of the post translational modifications that could happen to histone proteins is acetylation. This modification occurs at the ε-amino group of lysines, mostly on the amino-tail of histones and plays roles in the regulation of gene transcription. The acetylation of histone is highly reversible. Lysine residue is acetylated by histone/lysine acetyltransferases (HATs/KATs), while the removal of acetyl are done by histone deacetylases (HDACs) [6]. Deacetylation of histones could be signal for other histone modifications to occur which would lead to the change in gene transcription (Figure 1).

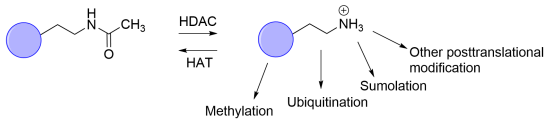


Figure 1: Scheme of HDAC and HAT. HDAC allows other histone modifications to occur in histones.

In humans, the HDAC enzymes are divided into four classes based on their sequence similarities, shown in Table 1. The class I proteins have similiarity to the Rpd3 protein in yeast, while class II proteins are similar to the Hda1 protein. Class III are similar to the yeast Sir2 protein, while Class IV protein shares similarity to both Class I and II proteins Class I, II, and IV belong to the arginase/deacetylase superfamily, while Class III belongs to deoxyhypusine synthase like NAD/FAD-binding domain superfamily [6].

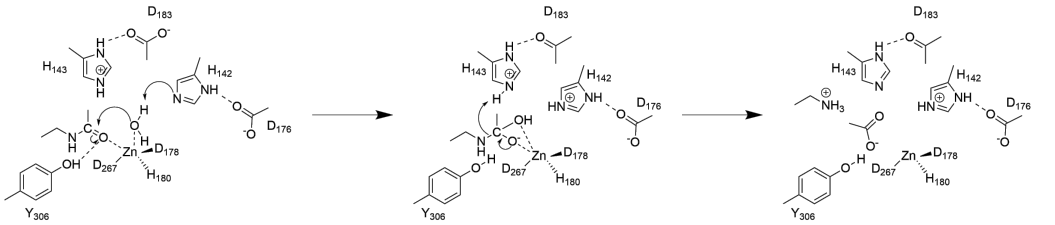
Class	Enzymes
Class I	HDAC1, HDAC2, HDAC3, HDAC8
Class II	HDAC4, HDAC5, HDAC6, HDAC7, HDAC9, HDAC10
Class III	SIRT1, SIRT2, SIRT3, SIRT4, SIRT5, SIRT6, SIRT7
Class IV	HDAC11

Table 1: HDAC family classification

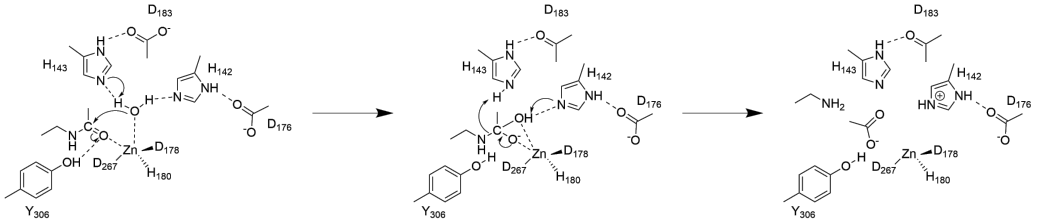
3.3 HDAC8

HDAC8 was first identified in the year 2000 [7]. Even though it is classified into class I HDAC, HDAC8 relatively have shorter C-terminal compared to the other HDACs in the same class [7]. In the other class I HDACs, this C-terminal domains are used for complex formation, which suggests that HDAC8 is either not recruited to form protein complex or that the complex recruitment happens in other part of the surface [8]. The opening to the active site of HDAC8 is able to change to accomodate different ligands, which suggests that HDAC8 might be able to deacetylate lysines on different structures [8].

In the center of HDAC8 catalytic site there is zinc ion which acts as electrophile. There are a couple of deacetylation mechanisms which have been suggested [9, 10]. For the first mechanism, it was proposed that the H142-D176 acts as the base to pull the proton from water molecule which then let the oxygen from water molecule to do nucleophilic attack on the carbonyl carbon of the acetyl lysine substrate [9]. While in the other proposed mechanism, it was the H143-D183 that acts as the general base [10].



(a)



(b)

Figure 2: Proposed mechanism of deacetylation by HDACs

3.3.1 HDAC8 role

Figure 3 shows the summary of HDAC8 substrates. HDAC8 could deacetylates both histones and non-histone proteins in vitro but in vivo, it is still debated whether histone is a bona fide HDAC8 substrate. This may happen because since other HDACs also act on histones, any change in HDAC8 activity may give rise to a counter regulation and hide the effect of the change [11]. Even so, there are some evidence that histone hyperacetylation can be observed upon cell treatment with HDAC8 inhibitor [12].

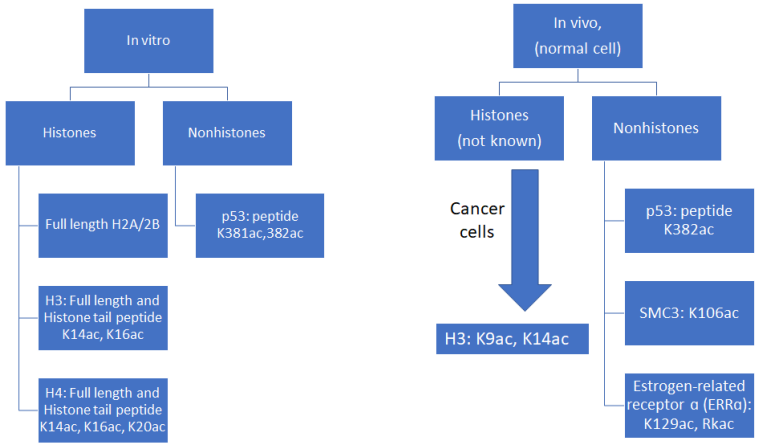


Figure 3: HDAC8 substrate *in vitro* and *in vivo*

As for non-histone substrate, a few proteins have been found to be in this category, such as p53, structural maintenance of chromosome 3 (SMC3), estrogen-related receptor α (ERRα) [13, 14, 15]. Suppression of HDAC8 resulted in higher expression of p53 as well as p53 acetylation in Lys382 leading to apoptosis in hepatocellular carcinoma cells [13]. Besides apoptosis, HDAC8 also plays a role in cell division as deacetylation of SMC3 by HDAC8 is important for dissociation of cohesin and also for facilitating cohesin renewal after its removal from chromatin [14, 16].

3.3.2 HDAC8 in cancers

In cancers, HDAC8 was found to be overexpressed in breast cancer, gastric cancer, esophagus cancer, colon cancer, and prostate cancer [17, 18] and knockdown of HDAC8 was found to be able to inhibit the proliferation of lung, colon, and cervical cancer cells [11] which clues in that HDAC8 is important in the growth of different cancer cells. Different mechanisms of how HDAC8 plays role in cancer cells have been found in different cancer cell lines, suggesting that HDAC8 may provide a scaffolding platform for signaling complexes to act on specific loci of DNA, thus affecting the expression of genes in cancer cells [11]. All this makes HDAC8 to be an interesting drug target.

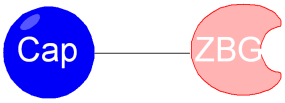
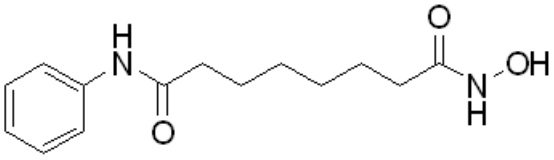


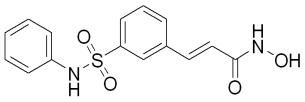
Figure 4: General structure of HDAC inhibitors

3.3.3 HDAC8 inhibitors

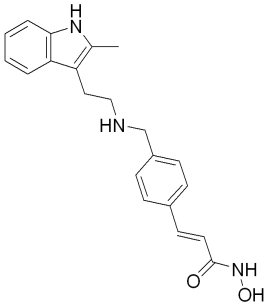
Since HDAC8 contains zinc ion in its active site, the presence of zinc binding group (ZBG) is crucial in the development for its inhibitors. The general structure of HDAC inhibitor consists of ZBG that interacts with active site, linker which occupies the channel, and hydrophobic cap as surface recognition domain (Fig 4) [19]. The most widely used ZBG is hydroxamic acid, since it has high affinity to the zinc ion due to its ability to form bidentate zinc coordination. Vorinostat, also known as SAHA, which is the first non-selective HDAC inhibitor approved by FDA, is bearing this functional group (Fig 5a) with IC50 $<$ 86 nM for HDAC class I and II [20]. Since then, there are more HDAC inhibitors bearing hydroxamic acid as ZBG that were approved by FDA, such as belinostat (Fig 5b) and panobinostat (Fig 5c) with IC50 against HDAC8 to be around 22 nM [21]. There is a reason why hydroxamic acid is the most common ZBG, other than its zinc-binding ability, it also has good solubility and stable in vitro [22]. It is far from perfect however, since there are evidence that it has poor bioavailability in humans (43%) and in other animals such as dog and rats (1.8% and 11% respectively) along with short life in human serum, around 1.74 hours when administered orally [23, 24, 25]. Moreover, hydroxamic acid moeity also has mutagenecity potential [26, 27], as well as its poor selectivity towards the HDAC isoforms [28].



(a) Vorinostat (SAHA)



(b) Belinostat

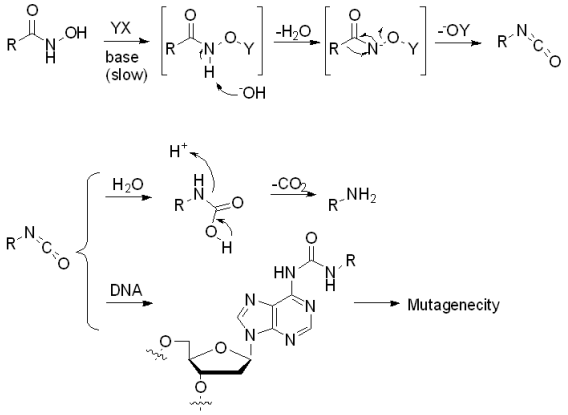


(c) Panobinostat

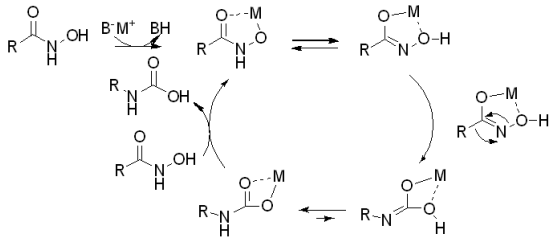
Figure 5: Structure of HDAC inhibitors bearing hydroxamic acid moeity as zinc binding group

The mutagenecity of hydroxamic acid group was shown from the result of Ames test where all hydroxamic acid HDAC inhibitors were found to be postive. The proposed mechanism for this mutagenecity was through the Lossen rearrangement [27]. Lossen reangement is a reaction that transforms an activated hydroxamate into isocyanate, which is unstable and can undergo hydrolysis or react with other nucleophile species, including DNA (Fig 6a) [29, 27]. Lossen rearrangement was observed to happen under physiological condition, where 2-naphthohydroxamic acid was converted into its O-acetyl derivative by acetyl-CoA in bacteria and that both 2-naphthohydroxamic acid and the O-acetyl derivative could form an adduct with bacterial DNA [29]. Moreover, there is also an evidence that the interaction between Zinc(II) and hydroxamic acid could trigger this Lossen rearrangement [30, 31].

Because of the reasons stated above, non-hydroxamic acid inhibitors have been developed. One of them is inhibitors bearing azetidine-2-one, which was found to be selective on HDAC8 in the micromolar range [32]. Other HDAC8 selective inhibitor that has been developed is bearing imidazole thione moeity [33]. This compound, called SB-379278A, was also found to have no activity on class II HDACs [33].

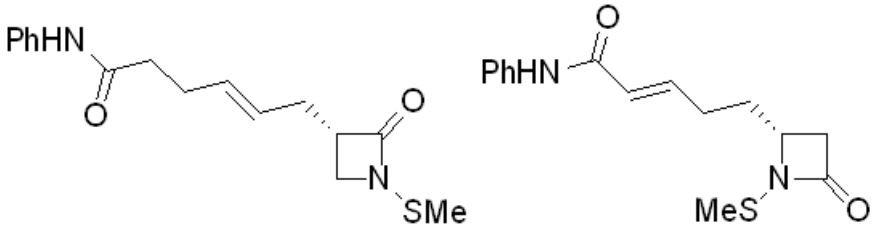


(a) Classical mechanism of Lossen rearrangement. Y could be Ac or $ArSO_2$and X is carboxyl.

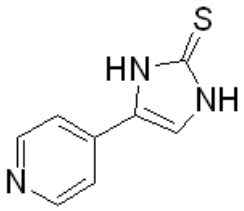


(b) Metal-assisted Lossen rearrangement

Figure 6: Lossen rearrangement



(a) HDAC8 inhibitors bearing azetidine-2-one moeity



(b) HDAC8 inhibitors bearing imidazole thione moeity

Figure 7: Examples of HDAC8 inhibitors bearing non-hydroxamic acid as zinc binding group

3.4 Machine Learning

Machine learning (ML) is the science of programming computers so they can learn from data [34]. The first ML application that became mainstream and took over the world in the 1990s is the spam filter. It was then followed by many more ML applications that we use regularly, such as shopping recommendations to voice search.

ML systems can be classified into two broad categories, based on the need of human supervision at training stage: supervised, unsupervised, semisupervised, and reinforcement learning. Supervised learning, as the name suggests, needs to be fed with training data that includes the desired solutions, called labels. Unsupervised learning, on the other hand, the training data is unlabeled. In semisupervised learning, the algorithm can deal with data that is partially labeled. Meanwhile, reinforcement learning is rather different. The learning system can choose and perform actions and get rewards or penalties in return based on the action they choose. The system has to learn the best strategy to get the most reward by itself.

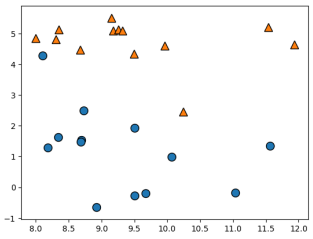
3.4.1 Supervised learning

The spam filter is an example of supervised learning, the system is trainined with many example of emails along with their class (spam or not spam) and it must learn how to differentiate new emails. This kind of task, letting the system learn how to predict a class label from a predefined labels is called classification [35]. Classification task can be differentiate by the number of classes they have, if there are only two classes, it is called binary classification and if there are more than two classes, it is called multiclass classification.

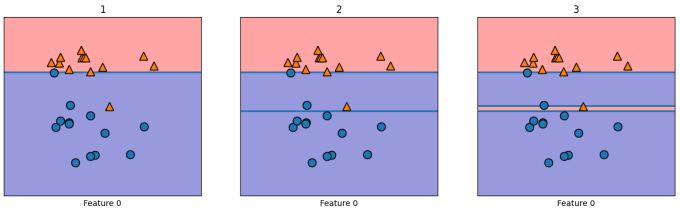
There are many supervised learning algorithms, but the ones used in this study is Decision Tree, Random Forest, SVC, kNN, and GNB.

3.4.1.1 Decision Tree

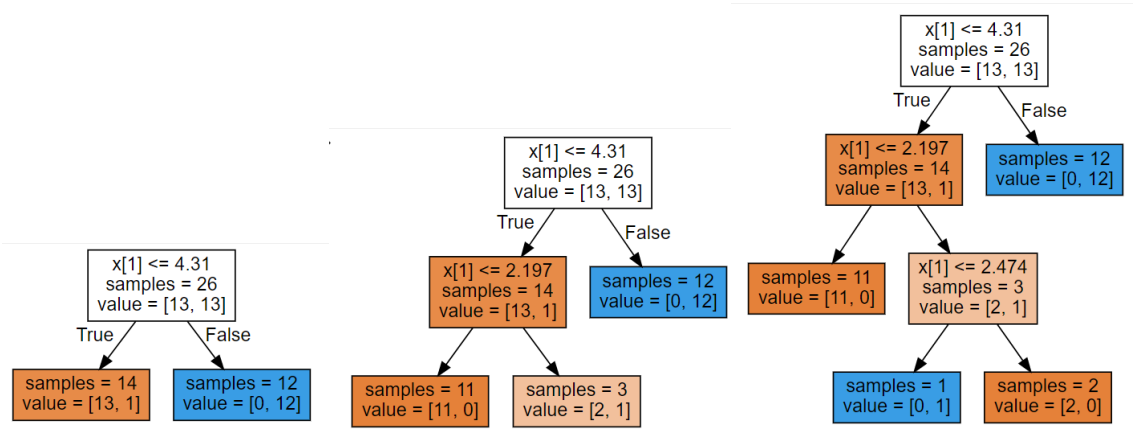
Decision Tree essentially is a model in which the algorithm learns a hierarchy of if/else questions (otherwise knowns as tests), leading to a decision [35]. The algorithm will find the sequence of tests that gets to the true answer the quickest. To build a tree, the algorithm searches over all possible tests and finds the one that is the most informative about the target variable. An example of Decision Tree vizualizations are provided in Figure 8. Figure 8b and 8c shows how the data points are classified and the hierarchy of the tests done, depending on the depth of the tree, respectively. In the tree depth = 3, the separation was done only for 3 data points.This may caused a problem, called overfitting, where the model, in this case Decision Tree, is really good at classifying the test data set but not general enough to classify new and unseen data set.



(a)



(b)

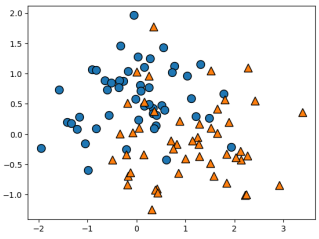


(c)

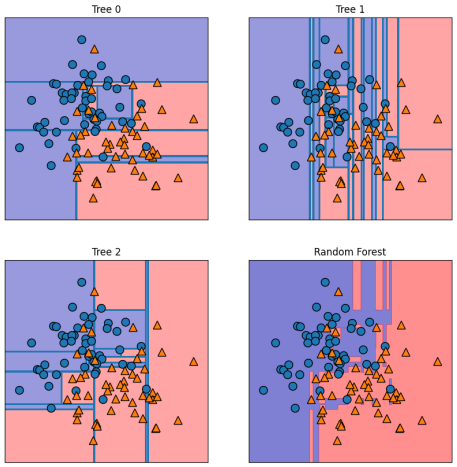
Figure 8: Visualization of Decision Tree model. (a) shows the dataset distribution, (b) shows the separating line for classification with tree depth of 1, 2, and 3, (c) shows the tree split with tree depth of 1, 2, and 3

3.4.1.2 Random Forest

Random forest is a collection of different decision trees. The idea is that each tree might do a relatively good job of predicting but will overfit on the train data. So, if many trees were built, the amount of overfitting can be reduced by averaging their results [35]. This is vizualized on Figure 9, where a random forest model to classify a data set was built by averaging three decision trees models. The result is a more generalized and less overfitted model.



(a)



(b)

Figure 9: Visualization of Random Forest model. (a) shows an example dataset for classification, (b) shows three different trees constructured and also the averaged one used as final model

3.4.1.3 Support Vector Classification

A visualization of support vector classification (SVC) can be seen in Figure 10. Essentially, SVC model will try to fit the widest possible street, which are represented by the dashed lines, between the classes. The decision boundary, represented by solid line, is fully supported by the instances located near the dashed lines (black circled on plot). These instances are called support vectors [34].

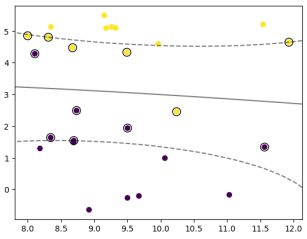
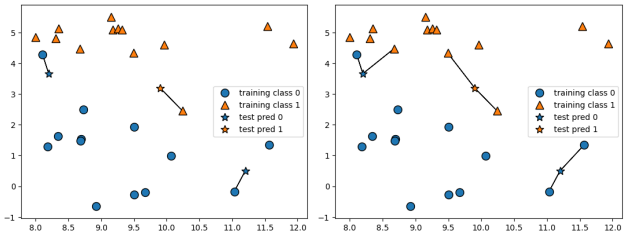


Figure 10: Visualization of SVC model. Data points with black circle around it are the support vectors

3.4.1.4 k-Nearest Neighbors

In k-nearest neighbors model, the algorithm finds the point in training set that is the closest to the new data point and assigns the label of the training point to the new data point [35]. The k means that any fixed number k of neighbors can be chosen in the training, then the prediction will be done based on the majority class of the neighbors (Fig 11).



(a) (b)

Figure 11: Visualization of kNN model with (a) 1 nearest neighbor and (b) 2 nearest neighbors.

3.4.1.5 Naive Bayes

Naive Bayes is a model which was based on Bayes theorem which describe the probability of an event, based on prior knowledge of conditions that might be related to the event [36, 37]. This can be summarized in the mathematic formula (1) below:

$$P_E(H) = \frac{P(H)P_H(E)}{P(E)} \tag{3.1}$$

Where $P_E(H)$ is probability of a hypothesis conditional on a given body of data, $P_H(E)$ is a probability of the data conditional on the hypothesis, and $P(H)$ and $P(E)$ are probability of observing hypothesis and data respectively without any given conditions [36]. When used as classifier, data E with attribute values of $(\chi_1, \chi_2, ..., \chi_n)$ is classified as class H = +, with H only has two classes, - and +, if and only if

$$f_b(E) = \frac{P_E(H = +)}{P_E(H = -)} \geq 1, \tag{3.2}$$

$f_b(E)$ is called Bayesian classifier [38].

In Naive Bayes classifier, or usually called Naive Bayes, all the attributes, $(\chi_1, \chi_2, ..., \chi_n)$ are independent given the value of the class variable, which in mathematical formula is describe as

$$P_H(E) = P_H(\chi_1, \chi_2, ..., \chi_n) = \prod_{i=1}^{n} P_H(\chi_i), \tag{3.3}$$

so that Naive Bayes can be described as:

$$f_{nb}(E) = \frac{P(H = +)}{P(H = -)} \prod_{i=1}^{n} \frac{P_{H=+}(\chi_i)}{P_{H=-}(\chi_i)} \tag{3.4}$$

Naive Bayes is the most simple form of Bayesian network [38].

3.4.2 Performance Measurement

3.4.2.1 Confusion matrix

Example of confusion matrix can be seen on Figure 12. The rows represent the actual classes, while the columns represent predicted classes. Based on this, the general idea of confusion matrix is to count the number of times instances of positive class is classified as negative class [34]. Back to Figure 12, the first row considers the negative class, if instances from negative class is correctly classified as negative class, they are called true negative (TN), while if instances that are falsely classified as positive class, they are called false positive (FP). The second row considers the positive class, if instances from positive class are wrongly classified as negative class, they are called false negative, while if they are correctly classified as positive class, they are called true positive. Based on this confusion matrix, there are a few metric that could be calculated, which will be explained below.

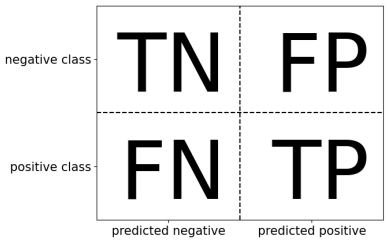


Figure 12: Confusion matrix. TN is true negative, FP is false positive, FN is false negative, and TP is true positive

3.4.2.2 Accuracy

Accuracy is a ratio of correct prediction. In binary classification this means accuracy is a measure of whether the data are correctly predicted into their actual class in both classes. Accuracy can be represented with mathematical formula below:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.5}$$

While at first glance, this can be a metric of comparison. but in imbalanced dataset, where there are class that appears more frequent than the other, accuracy is not the preferred performance measurement. For example, with imbalanced data ratio of 1:99 for positive and negative class respectively, if the model always predict a data as negative class, it would be right for 99% of the time. This beats the purpose of building the model to differentiate two classes and makes the prediction to be meaningless.

3.4.2.3 Precision

Another metric that can be used for performance measurement is called precision. Precision is looking at the accuracy of the postive predictions, represented in the formula below:

$$Precision = \frac{TP}{TP + FP} \tag{3.6}$$

Precision is used when we want to limit the number of false positive.

3.4.2.4 Recall

Recall, also called sensitivity, is a ratio of positive instances that are correctly detected by the model.

$$Recall = \frac{TP}{TP + FN} \tag{3.7}$$

This is used when we want to limit the number of false negative. For example, in cancer diagnosis, we want to correctly predict if someone really have cancers and to not correctly diagnose them as to not have cancer.

3.4.2.5 F1 score

F1 score is a metric from combining both precision and recall. It is defined as the harmonic mean of precision and recall [34]. The equation to calculate F1 score is described below:

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} = \frac{TP}{TP + 0.5(FP + FN)} \tag{3.8}$$

F1 being a harmonic mean means that it gives much more weight to low values. A classifier model will only get a high F1 score if both precision and recall are high [34].

3.4.2.6 ROC curve and ROC AUC score

Receiver operating characteristic (ROC) curve is widely used for binary classifiers. This curve plots recall (also known as true positive rate or TPR) against false positive rate (FPR).

One of the ways to compare different classifier models is by comparing the area under the curve (AUC) score from the ROC curve. A perfect classifier will have a ROC AUC score of 1, whereas a purely random classifier will have ROC AUC score of 0.5 (Fig 13 and Table 2). Based on Fig 13 and Table 2 for example, Classifier 1 is the better model from Classifier 2 because Classifier one has higher ROC AUC score than Classifier 2.

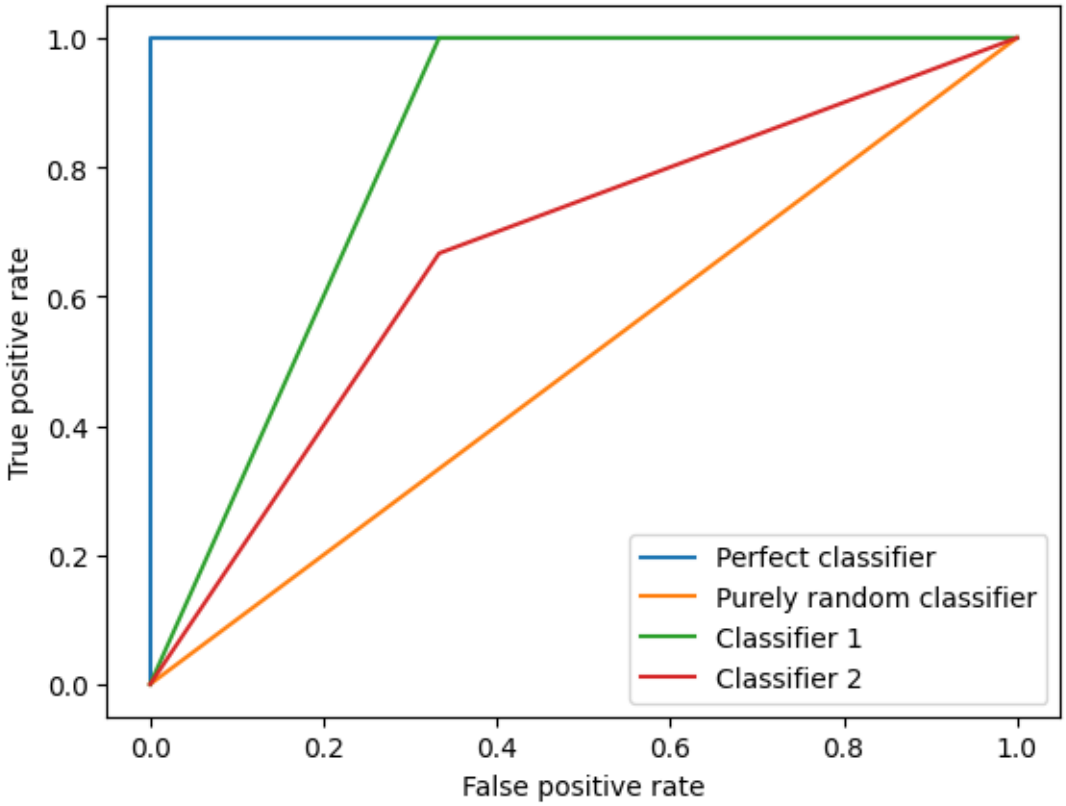


Figure 13: ROC curve example

Model	ROC AUC score
Perfect classifier	1
Purely random classifier	0.5
Classifier 1	0.83
Classifier 2	0.67

Table 2: ROC AUC scores of models in Fig 13

3.4.3 Imbalanced dataset

Imbalanced dataset, in a binary classification, is a dataset where one class has much more instances than the other class. This leads in difficulty of the models to learn from the class with less instances, which is also called minor class, compared from the instances from the class with more instances, called majority class. In general, it is a problem, but if the class we interested in is the minority class, for example, spam detection or even cancer diagnostic kit, where the focus is to detect the minority class (the amount of spam emails is less compared to non-spam and people diagnosed with cancer is not as big as the whole population itself) the models that describe these minority class have to be specialized and can not be easily simplified into more general rules with broader data coverage [39]. Beside this, data with high noise makes it difficult for the model to learn from minority class. The model would not be able to differentiate between minority class and noise-induced casses [40]. And techniques that try to minimize noise usually perform at the expense of the minority class, as they tend to remove both the noise and minority instances [39]. So there is the need to tackle this problem.

There are ways to build models with imbalanced data, which could be divided by 2 groups, which are solutions developed at data and algorithm levels. At data level, the goal is to rebalance the class distribution by resampling the data. Meanwhile, at the algorithm level, the existing classifier algorithm is adapted to strengthen the learning of minority class [39]. In this project, the focus is on the data level. In this level, the way to deal with imbalanced data is by resampling the data. There are different kinds of resampling that can be introduced, such as oversampling the minority class and undersampling the majority class [40]. There are a few different techniques to do oversampling of minority class and undersampling majority class, the one used in this project is oversampling minority class by generating new synthetic data, called Synthetic Minority Oversampling Technique (SMOTE).

3.4.3.1 SMOTE

SMOTE generates synthetic examples by operating in "feature space". The minority class is over-sampled by taking each minority class sample and introducing synthetic examples and is randomly chosen depending on the amount of over-sampling required [41].

Synthetic samples are generated following these steps:

1. Take the difference between the feature vector (sample) under consideration and its nearest neighbor
2. The difference was multiplied by a random number between 0 and 1
3. Add the result to the feature vector under consideration

These steps cause the selection of a random point along the line segment between two specific features and force the decision region of the minority class to be more general [41]. When compared to minority over-sampling by replication, SMOTE could make a broader decision regions, leading to generalization of minority class, rather than becoming more specific [41].

3.5 Drug Discovery

Drug discovery and development, which include lead screening and optimizing, pre-clinical clinical phases tests, to approval and market launch, are a time-consuming process which could take more than 10 years from start to finish (Fig 14) [1]. Other than that, it is also costly, with the average cost for developing a new drug was estimated to be more than 1 billion US dollars [42]. And even though the success rate of drugs that pass clinical trial phase III to launch has increased from less than 50% to more than 60%, the success rate of a drug to pass from phase I to launch has remained stagnant over the years, with only less than 10 percent success rate (Fig 15) [2].

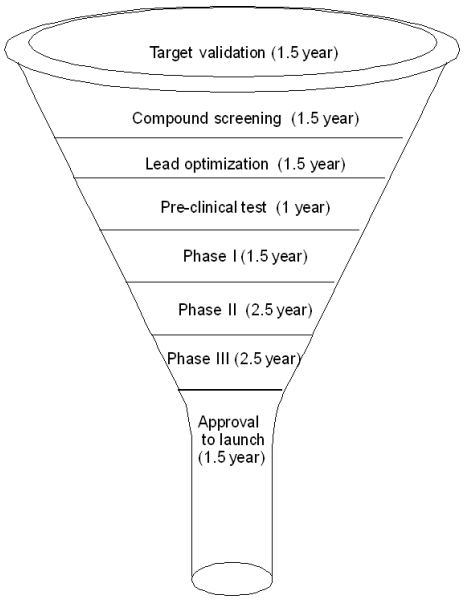


Figure 14: Drug discovery pipeline and approximation of time need to finish each step [1]

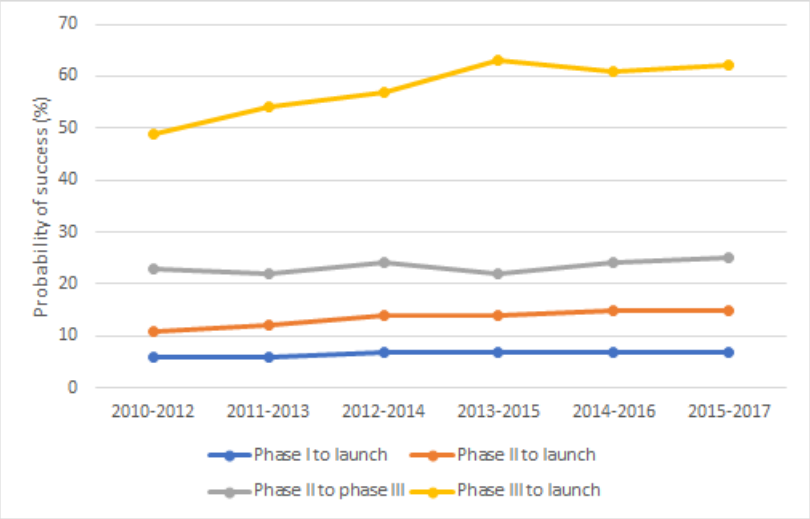


Figure 15: Probability of success from phase I, II, and III to launch [2]

This low success result drives the pharmaceutical companies to use machine learning technique in their drug development to lower the overall time and costs needed. The popular use machine learning in drug discovery can be seen on the interest of pharmaceutical companies collaborating and developing their own artificial intelligence/machine learning technologies.

Machine learning algorithms and software have been developed and used in all stages of drug discovery and development [43]. Moreover, in 2020, Excentia reported the first AI-design drug candidate to enter clinical trial [44] and in June 2023 it was reported by Insilico INS018_55, an inhibitor treatment for idiophatic pulmonary fibrosis, that was discovered and designed by AI had entered clinical Phase II trial [45]. Due to the nature of this study, the focus will be on the usage of machine learning in compound screening and lead optimization part.

3.5.1 The use of machine learning in compound screening and lead discovery

In 2015 Chemical Abstracts Service (CAS), who authorized chemical information, registered the 100 millionth chemical substance in its registry with more than 650 million chemicals are predicted to be added in the next 50 years [46]. With this, alongside with the common use of high-throughput screening in pharmaceutical research [47], the amount of data related to drug discovery is also growing fast and there is a demand to explore these growing data [48]. The use of computational methods, machine learning in particular, thus is seen as the tool to keep up with this growing data.

One of the ways machine learning could be used in drug discovery is model generations for properties predictions. Using the results from high-throughput data from chemical libraries as training data, property of a new, unseen molecules could be predicted, pre-

venting researchers to work on compounds that are unlikely to be effective.

Many models of property predictions have been developed using many different algorithm and infrastructures, from traditional machine learning to deep learning and some are publicly available, such as DeepChem [49], OpenChem [50], etc.

3.5.2 Chemical databases

With the growing amount of data produced by chemists, comes the demand to have databases to pool those chemical data for easy access and transparency. One of the these databases is called ChEMBL. ChEMBL is an open-access database which contains information about small molecules and their biological activity. The data is extracted from full text articles of Medinical Chemistry journals [51]. They also integrate it with data on approved drugs and clinical development candidates [51]. Searching for the target protein could also be done through ChEMBL and information about the associated bioactivities with the associated compounds and assays could be viewed and retrieved.

Retrieval of data from ChEMBL could be done by its web service which can be called via programming language or workflow tool [52]. The web service can also be used to filter and sorting the ChEMBL database to get specific entry, whether it is molecules, target proteins, bioactivity data points, etc. This makes the ChEMBL database to have variety of applications such as training machine learning models for target prediction.

3.6 PubChem Fingerprints

PubChem generates a binary substructure fingerprints for chemical structures which are then used for similarity neighboring and similarity searching [53]. The size of this fingerprint is 881 bits long and each bit represents a test for the presence of, an element count, a type of ring system, atom pairing, etc in a structure [53].

3.7 Chemical Space and Compound Similarity

Chemical space is a concept to illustrate the distribution of molecules as well as their properties in the form of geographical map [54]. This concept is used in medicinal chemistry to describe the ensemble of all organic molecules to be considered when searching for new drugs [55]. This geographical map can be obtained by assigning dimensions to the molecular descriptors of each molecule. Then dimension reduction method is done so the map can be projected in 2D or 3D map [54].

One of the ways to reduce the dimensions of molecule descriptors is by using a method called t-distributed stochastic neighbor embedding (t-SNE). t-SNE visualized high dimensional data by giving the datapoints a location in a two or three dimensional map in such a way that similar datapoints are modeled by nearby points and dissimilar datapoints are modeled by distant points with high probability [56].

Using t-SNE, the projected chemical space thus then can group compounds that are similar to each other. This similarity concept is used in many drug design with the basis

of a principle that states that structurally similar compounds are more likely to show similar properties [57]. A method that is often used to measure compound similarity is Tanimoto coefficient which is deemed to be one of the best metric for compound similarity calculations [58]. Tanimoto coefficient is defined as the ratio of the intersection of two sets over the union of the two sets. The calculation is using molecular descriptors for each molecules and compared to each other's.

3.8 Absorption, distribution, metabolism, excretion, and toxicity

The terms absorption, distribution, metabolism, and excretion in drug discovery was first brought up in English by Eino Nelson in his 1961 paper where he mentioned that detailed kinetic study on drugs is needed to help understand the mechanism of how the drugs act, how the body modifies and eliminates drugs, and the relationship between drug concentration and activity [59]. Since then, the absorption, distribution, metabolism, and excretion (ADME) has been widely used and became a standard term in drug testing and clinical practices [60].

Traditionally, in the drug discovery process, after promising compounds were developed, their ADME and toxicity (making it into ADMET) properties were then investigated. In this stage, usually, adverse findings were discovered which resulted in the project to be paused or even restarted from the very begining. In a study done in 2016, safety of the drug is one of the major reasons why drugs fail to pass clinical trials (Fig 16) [3]. So there is a need for ADMET properties to be considered on the earlier stage of drug development.

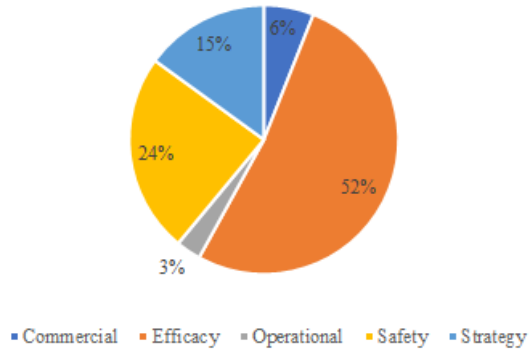


Figure 16: Reasons for clinical failures in 2013-2015 [3]

The need for ADMET properties data early in the stage of drug discovery are done in three ways: automated *in vitro* assays, *in silico* models to assist the assays and selection of compounds, and the development of predictive models for ADMET properties [61]. The use of predictive models to determine ADMET properties is especially useful in the compound design stage, to determine whether a compound could proceed to the next

stage or not. There are many predictive models for ADMET properties developed to carry this purpose, the one used in this research is called ADMETlab2.0.

ADMETlab2.0 was introduced in 2021 and is an upgrade from ADMETlab, which was released in 2018. It is capable of doing single-molecule evaluation and batch screening to calculate 88 ADMET related properties (Table **??**) [62]. Compared to other web-based tool for ADMET properties prediction, ADMETlab2.0 could predict more properties, especially toxicity properties, and done it in a shorter amount of time [62]. It is also publicly available and can be access freely through web server, which makes it convinient, fast and reliable way to predict ADMET properties for this project.

3.9 Molecular docking

Molecular docking is a method to predict ligand-receptor complex structure using computation methods [63]. Conformations of the ligand in the active site of protein are sampled and then ranked by a scoring function. Molecular docking is widely used as a part of in silico workflow, combined with other computational techniques and experimental data, of drug discovery. There are a lot of docking programs that are widely used in pharmaceutical industries, but the one used in this study and will be explained further is Glide.

Glide uses a series of hierarchical filters to search for possible locations of the ligand in the active-site of the receptor. GlideScore, a scoring system for predicting binding affinity and rank-ordering ligands, is produced as one of the outputs, along with the pose of the ligand with the protein [64].

3.10 Purpose of this work

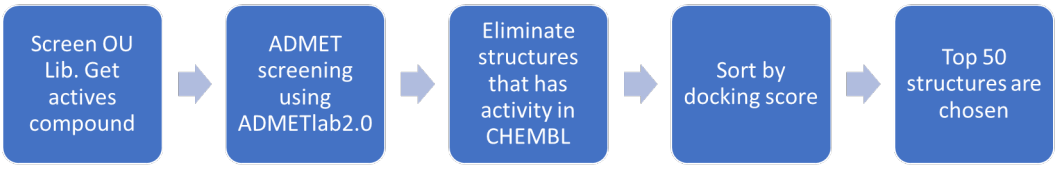
Since there is a need to find HDAC8 inhibitor with non-hydroxamic acid as its zinc binding group, I would like to use machine learning to find new molecules with HDAC8 inhibitory activity. Also, since drug discovery dataset could be prone to data imbalance, I would like to use SMOTE to solve this issue. I would argue that the generalization capability of SMOTE could lead to a discovery of different moeity capable to inhibit HDAC8.

4 Material and Methods

4.1 General methodology

Data for training the models were obtained from ChEMBL. After conversion to PubChem fingerprints and labeling the compounds as active or inactive, the dataset were split into training and testing dataset. Training dataset were then used to train all five models (Decision Tree, Random Forest, SVC. kNN, and Naive Bayes). Then the ROC AUC score of each models were compared (Fig 17a). After obtaining the Osaka University Library dataset, PubChem fingerprints were made based on the SMILES sequences and screened to the chosen model. ADMET screening was then done and compounds that passed this screen were then checked if it has similarity to known active compounds in ChEMBL database. Compounds that do not have similarities with known active compounds were then sorted by its docking score to HDAC8. The top 50 compounds were then tested for its HDAC8 inhibition capability (Fig 17b).

(a) Data retrieval and model screening workflow



(b) Screening and filtering of Osaka University Library

Figure 17: Workflows of models evaluation (a) and screening lead structures (b)

Data retrieval, model building, and screening of Osaka University Library compounds were done in Python 3.9, using Jupyter notebook. pandas module [65] was used to do data analysis. The packages used in the model building were listed in Table 3.

Packages	Version
pandas	1.5.3
numpy	1.22.4
matplotlib	3.7.1
sklearn	1.2.2
imblearn	0.10.1
seaborn	0.12.2
padelpy	0.1.11
pickle	4.0

Table 3: List of packages version used in this study

4.2 Data retrieval from ChEMBL and PubChem fingerprint conversion

ChEMBL data retrieval was done using ChEMBL websource client [52] in Python 3.9. HDACs from *Homo sapiens* were chosen. Then the dataset is filtered as follows:

1. Only retrieve structures which have IC50 value
2. Must have canonical SMILES
3. Duplicates were dropped

After that, IC50 value was converted to pIC50 value and structures which has pIC50 $>$ 7 is labeled as active and others as inactive.

Canonical SMILES of each structure was used to generate the PubChem fingerprints using PaDELPy package (https://github.com/ecrl/padelpy) [66].

4.3 Models and prediction

The ChEMBL dataset was randomly divided into train and test set with ratio 8:2 respectively. The models were built in Scikit-learn package [67], all using default parameters. SMOTE was implemented through the imbalanced-learn [68] package. Cross validation was done using RepeatedStratifiedKFold from Scikit-learn [67] with ten-fold validation and 3 repeats and the mean AUC was reported. For the ROC AUC score from test set RocCurveDisplay [67] module from Scikit-learn was used. For prediction, the canonical SMILES of the Osaka University compound library were converted into PubChem fingerprints and the fingerprints were used to predict with the output as a label of 'active' or 'inactive' for each compound. Compounds labeled as 'active' were continued to be filtered to the next step.

4.4 Molecule comparison to known ChEMBL active compounds

To eliminate re-discovery of known compounds, compounds that were predicted as active were compared to ChEMBL database. This was possible by using DataWarrior's "Get Similar Compounds from ChEMBL Actives" option [69]. Using this option, DataWarrior compared the predicted compounds with ChEMBL database based on their SkelSphere descriptors and their Tanimoto similarity score was calculated [**?**]. Compounds that had Tanimoto similarity score of 0.8 or higher compared to any known active ChEMBL compounds were filtered out.

4.5 ADMET screening using ADMETlab2.0

ADMET screening is done using ADMETlab2.0 web resource (https://admetmesh.scbdd.com/) [62]. Compound was deemed to pass this screening if they pass all of these criterias as below:

Category	Threshold	Notes
Lipinski	Accepted	
Pfizer	Accepted	
GSK	Accepted	
Golden Triangle	Accepted	
PAINS	0	Pan assay interference compound
QED	> 0.67	A measure of druglikeness based on the concept of desirebility
Ames	< 0.7	Ames test for mutagenecity
hERG	< 0.7	

Table 4: ADMETlab2.0 criterias thresholds

4.6 Datawarrior's Druglikeness score

Datawarrior is a program built for data analysis in chemistry which can be used for visualization and data analysis [69]. One of the analysis that DataWarrior can do is the calculation of Druglikeness score. Their approach on Druglikeness score is based on a list of around 5300 distinct substructures with their own scores. The list was created by shredding 3300 traded drugs as well as 15000 commercially available (Fluka) chemicals. The occurence frequence of every fragments was determined within the collection of traded drugs and the supposedly non-drug-like collection of Fluka compounds. All fragments with an overall frequence above a threshold were inverse clustered to remove highly redundant fragments. As for the remaining fragment, the druglikeness score was determined as the logarithm of the quotent of frequencies in traded drugs versus Fluka chemicals. A positive druglikeness score means that the molecule contrains predominantly fragments which are frequently present in commercial drugs [**?** 69].

4.7 Docking score calculation and benchmarking using Glide

Docking was done using Glide [64]. Protein crystal structure of HDAC8 was imported from PDB (ID: 1T69). Grid generation was done excluding water molecules and not allowing any reside to rotate. Ligand preparation was done in LigPrep with OPLS4 force field with ionization at pH 7.4 using Epik and metal binding site was added. The structures were also desalted and tautomers were generated. For sorting and benchmarking, docking was done in SP mode, while for confirmation, docking was done in XP mode.

For benchmarking Glide program, compounds with similar properties as ligands, or in this case is called decoys, for HDAC8 as well as HDAC8 ligands were obtained from DUD-E database [70]. Total ligands docked for benchmark was 10170 and there were 170 active compounds among them. The enrichment was then calculated and the ROC plot were constructed.

4.8 HDAC8 inhibitory assay

HDAC8 inihibitory assay was done using CycLex® HDAC8 Deacetylase Fluorometric Assay Kit Ver.2 from MBL Life Science. This assay is based on mechanism shown in Fig 18.

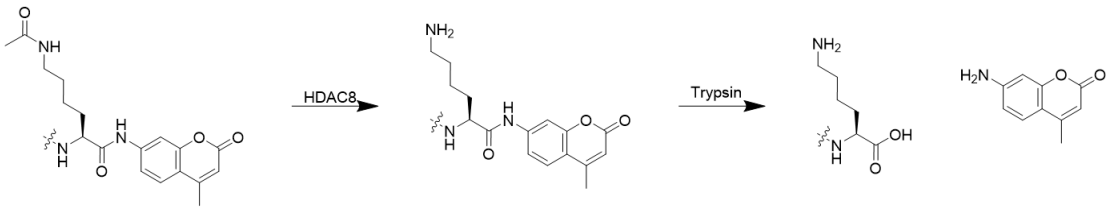


Figure 18: Mechanism of HDAC8 inhibitory assay

To screen the top 50 compounds, the final concentration of compounds was set to 10 μM. Following their protocol, reaction was done in black 96 wells plate and reaction mix was incubated in 30 °C for 30 minutes in a shaker with 500 rpm. After stop solution was added, fluorescence measurement was done using EnSight Fluoresence plate reader with excitation wavelength of 280 nm, emission wavelength of 460 nm, and 100 flashes. Experiment was done in triplicates. IC50 graph was made using GraphPad software.

4.9 HDAC inhibitory assay

Following similar mechanism from HDAC8 inhibitory assay, inhibitory assay for HDAC1 and HDAC3 for compound 12 was done. Reaction mix consists of HDAC assay buffer, compound 12, recombinant HDAC were mixed in 96 wells plate and incubated in 30 °C for 30 minites in a shaker with 500 rpm. Then reaction substrate which is a peptide with sequence of RHKK(Ac)AMC was added and mixture was incubated again with the same condition as the first one. After that, developer solution was added and mixture was

once again incubated with the same condition. Fluorescence level was then measured using EnSight Fluoresence plate reader with excitation wavelength of 280 nm, emission wavelength of 460 nm, and 100 flashes. Experiment was done in triplicates. IC50 graph was made using GraphPad software.

5 Results and Discussions

5.1 ChEMBL database preparation

Using the method already described in previous chapter, compounds which have been tested against HDACs were retrieved from ChEMBL database. As ChEMBL stores information of target proteins from many species, only HDACs from humans (*Homo sapiens*) were chosen. Table 5 shows the ChEMBL IDs of HDAC used in this study. After filtering compounds which do not have IC50 values and dropping any duplicates, each compound was labeled as active or inactive based on their pIC50 values, with compounds with pIC50 more than 7 was labeled as active. Figures 19 show the distributions of compounds labeled as active or inactive for each HDAC target and Table 6 shows the ratio of the distributions.

HDAC	ChEMBL ID
HDAC1	CHEMBL325
HDAC2	CHEMBL1937
HDAC3	CHEMBL1829
HDAC4	CHEMBL3524
HDAC5	CHEMBL2563
HDAC6	CHEMBL1865
HDAC7	CHEMBL2716
HDAC8	CHEMBL3192
HDAC9	CHEMBL4145
HDAC10	CHEMBL5103
HDAC11	CHEMBL3310

Table 5: ChEMBL code

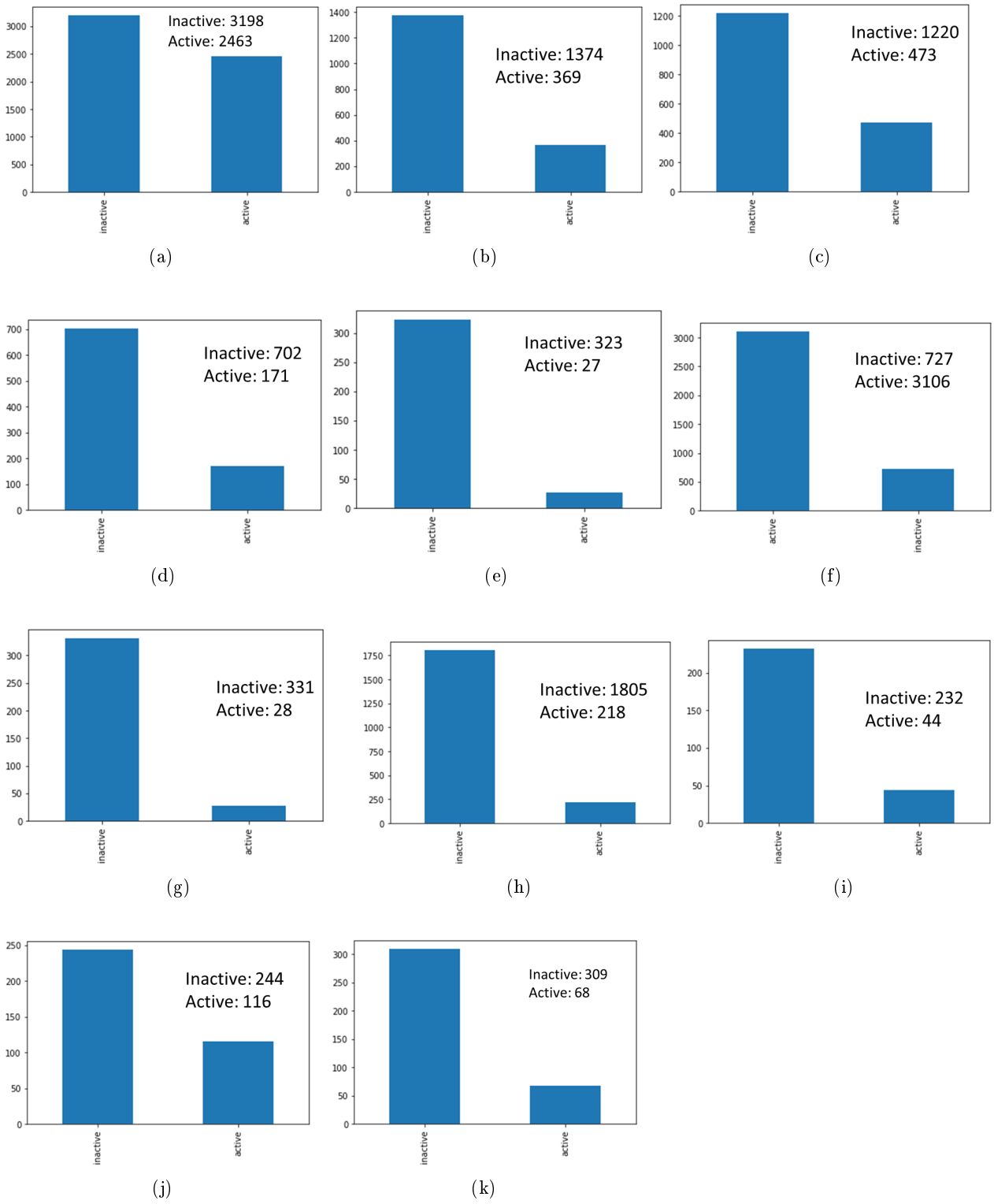


Figure 19: Distribution of compounds labeled active and inactive on ChEMBL library for (a) HDAC1, (b) HDAC2, (c) HDAC3, (d) HDAC4, (e) HDAC5, (f) HDAC6, (g) HDAC7, (h) HDAC8, (i) HDAC9, (j) HDAC10, (k) HDAC11

HDAC	inactive : active ratio
HDAC1	1 : 0.77
HDAC2	1 : 0.27
HDAC3	1 : 0.39
HDAC4	1 : 0.24
HDAC5	1 : 0.08
HDAC6	1 : 4.27
HDAC7	1 : 0.08
HDAC8	1 : 0.12
HDAC9	1 : 0.19
HDAC10	1 : 0.48
HDAC11	1 : 0.22

Table 6: Ratio of compounds labeled inactive and active for each HDAC target protein

All HDAC datasets are imbalanced with different degrees of severity, with HDAC1 being less imbalanced in comparison with other HDAC datasets and HDAC6 dataset is skewed towards active compounds. As for HDAC8, although it is not as severe as, let's say HDAC7, the dataset is still skewed towards inactive compounds. With this in mind, models using these dataset were built.

5.2 First Screening

5.2.1 Model building and comparison

Models were built using PubChem fingerprints as input. PubChem fingerprints are generated based on the substructure of molecules. Usually these fingerprints are used by PubChem for similarity neighboring and similarity searching [71]. This fingerprint was chosen because it covers a wide range of different substructures and features. To generate PubChem fingerprints, SMILES sequence of each molecules were converted using padelpy [**?**].

Using HDAC8 dataset, different models were tested to choose the best one for this study. The area under curve (AUC) of reciever operating characteristic (ROC) scores of the cross validation and test set were used as comparison. The use of SMOTE to solve the imbalanced dataset were also observed.

Table 7 show the ROC AUC scores for each models, both using normal distributions and SMOTE. Random Forest seems to be the most suited model for this work and the usage of SMOTE with Random Forest model (RF-SMOTE) improved the precision, recall, F1 score and accuracy compared to normal distribution (Table 7c and 7d).

The reason that SMOTE could improve those parameters might be because SMOTE helped the model to generalized better [41] along with overfitting that could be reduced by using random forest, the combined model seems to be able to make a more general model, than the others. Based on this, RF-SMOTE model was used to screen Osaka

HDAC	Amount of compounds labeled as active by RF-SMOTE
HDAC8	1556

Table 8: List of Osaka University Library compounds that were labeled as active on HDAC8

University compound library.

HDAC	True distribution					SMOTE				
	Decision Tree	Random Forest	SVC	kNN	Naive Bayes	Decision Tree	Random Forest	SVC	kNN	Naive Bayes
HDAC8	0.69	0.84	0.78	0.79	0.62	0.91	0.98	0.95	0.95	0.76

(a) ROC AUC score from cross validation

HDAC	True distribution					SMOTE				
	Decision Tree	Random Forest	SVC	kNN	Naive Bayes	Decision Tree	Random Forest	SVC	kNN	Naive Bayes
HDAC8	0.66	0.75	0.75	0.71	0.56	0.91	0.99	0.97	0.97	0.67

(b) ROC AUC score from test dataset

	Precision	Recall	F1 score
Active	0.52	0.29	0.38
Inactive	0.92	0.96	0.94
Accuracy	0.89	0.89	0.89

(c) Classification report true distribution Random Forest

	Precision	Recall	F1 score
Active	0.94	0.96	0.95
Inactive	0.96	0.94	0.945
Accuracy	0.95	0.95	0.95

(d) Classification report Random Forest SMOTE

Table 7: ROC AUC scores table

5.2.2 Screening Osaka University Compound Library

Using RF-SMOTE model that has been built, compounds from Osaka University library were screened. Table 8 shows the amount of compounds from Osaka University Library that were labeled as active in HDAC8. All 1556 compounds were then subjected into ADMET screening.

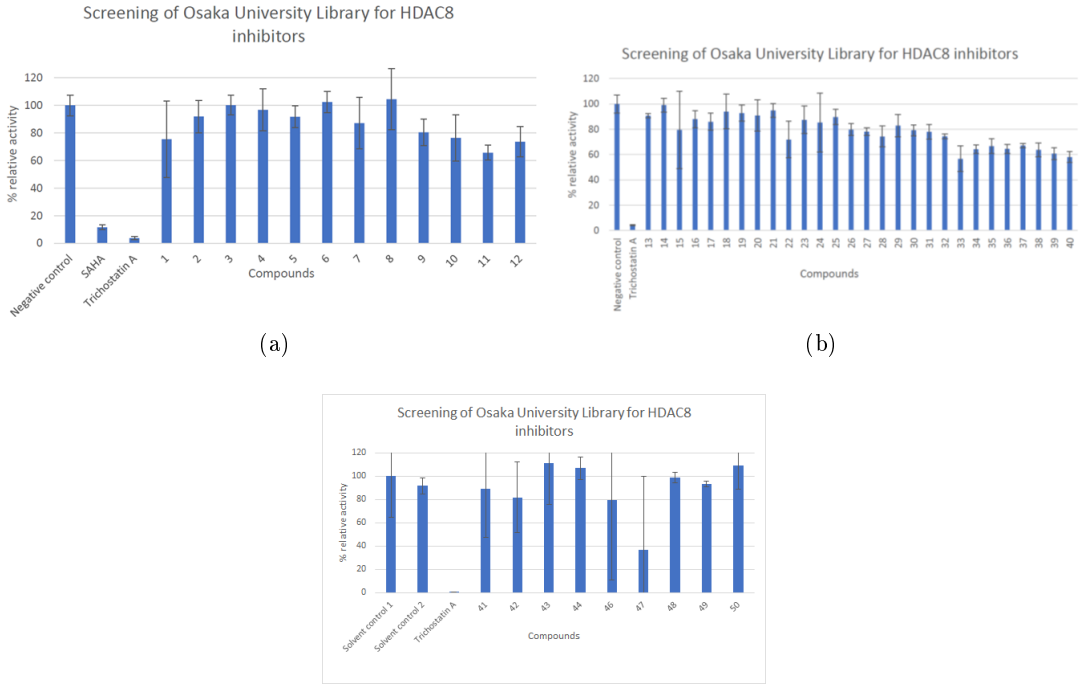
5.2.3 Similarity to ChEMBL active compounds and ADMET screening

Because the purpose of this study is to find new molecules that could inhibit HDAC, compounds were screened based on their similarity to ChEMBL active compounds. This was done through DataWarrior as explained in previous chapter [69]. 68 compounds were found to be similar with known ChEMBL active compounds, making it only 1488 compounds to go through the next screening.

To filter out compounds that do not or have low drug-like properties, ADMET screening was done using ADMETlab2.0 described in earlier chapter [62]. There were 405 compounds that passed this step. These 405 compounds were then sorted by their Druglikeness score from DataWarrior [69] and 50 compounds with the highest Druglikeness score were subjected into HDAC8 inhibitory assay.

5.2.4 HDAC8 inhibitory activity

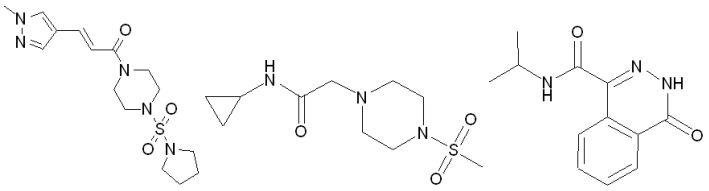
Figure 20 show the result of HDAC8 inhibitory activity of the top 50 compounds. From these results, compound 11 (Fig 21a) was deemed to be promising with around 30% inhibition on HDAC8. Compound 36 (Fig 21b) and compound 46 (Fig 21c) were randomly chosen as comparison.



(a) (b)

(c)

Figure 20: HDAC inhibitory activity of the top 50 compounds



(a) Compound 11 (b) Compound 36 (c) Compound 46

Figure 21: Structures of compound 11, 36, and 46

All of the hit compounds in the 1st screening were found to be inactive (Fig 22).

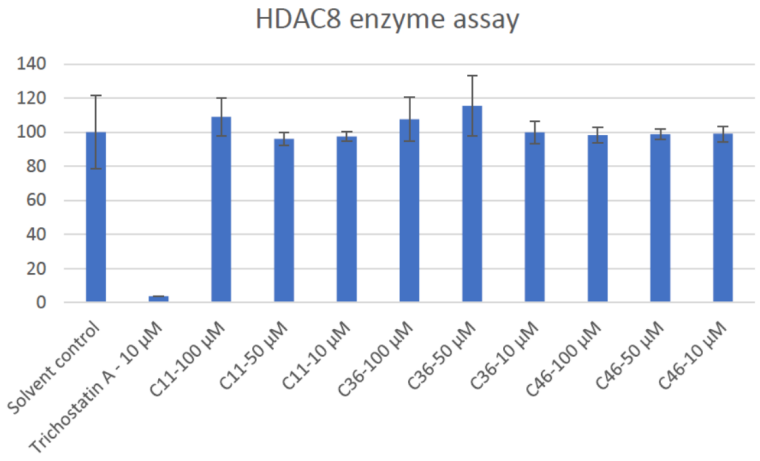


Figure 22: Reproducibility result

5.3 III.3 Second screening

5.3.1 Model building and comparison

Since the top 50 compounds were found to be inactive, those compounds were labeled as inactive and put into the training dataset alongside ChEMBL dataset then RF-SMOTE model were rebuilt. Similar to the first screening, the use of SMOTE improved the precision, recall, F1 score, and accuracy of random forest model in this work.

	Precision	Recall	F1 score
Active	0.56	0.34	0.42
Inactive	0.92	0.97	0.94
Accuracy	0.90	0.90	0.90

(a) Classification report true distribution Random Forest

	Precision	Recall	F1 score
Active	0.94	0.98	0.96
Inactive	0.98	0.94	0.96
Accuracy	0.96	0.96	0.96

(b) Classification report Random Forest SMOTE

Table 9: Classification report for true distribution of Random Forest and Random Forest with SMOTE

5.3.2 Screening Osaka University Compound Library

Using RF-SMOTE, models for each HDAC dataset were built and Osaka University Compound Library were screened using these models. Models were built for all HDACs, not just HDAC8 to check the selectivity of the compounds done later in this work. Tabel 10 showed the amount of Osaka University Compound Library that were labeled active for each HDACs.

HDAC	Amount of compounds labeled as active by RF-SMOTE
HDAC1	1181
HDAC2	1239
HDAC3	2172
HDAC4	1407
HDAC5	708
HDAC6	294
HDAC7	248
HDAC8	235
HDAC9	3101
HDAC10	5559
HDAC11	4792

Table 10: List of Osaka University Library compounds that were labeled as active against each HDACs

For HDAC8, compared to the first screening, the amount of compound labeled as active were reduced, this might be because the addition of 50 compounds were able to help reducing the amount of false positive results. Figure 23 shows the chemical space of ChEMBL active compound (in black), compounds labeled active in the first screening (in blue), and the top 50 compounds from the first screening (in green) and compounds labeled active from the second screening (in purple). The addition of 50 compounds as false positive managed to change the distribution of the active labeled compounds in the second screening (in purple).

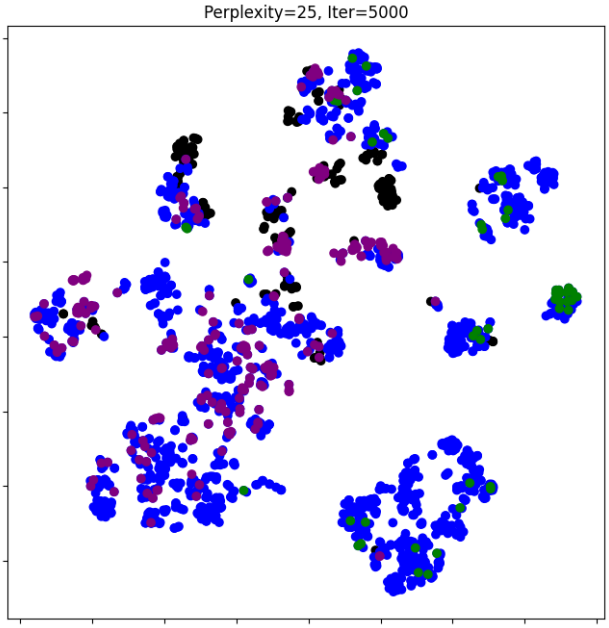


Figure 23: Chemical space of ChEMBL active compound (black), compounds labeled active in the first screening (blue), and the top 50 compounds from the first screening (green) and compounds labeled active from the second screening (purple).

The similarity of active-labeled compound from ChEMBL and from this screening was also measured according to their Tanimoto similarity score (Fig 24). From this result, the similarity between these two compound groups are rather low, which means there is high variability in the screened compound and new structures for HDAC8 inhibitor could be identified.

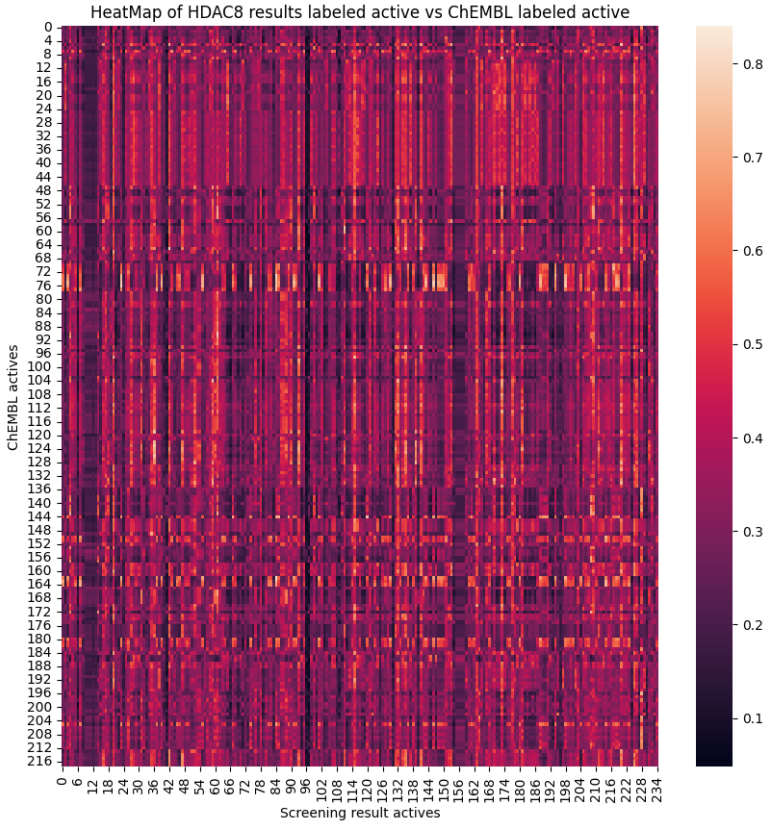


Figure 24: Heatmap of tanimoto score of active-labeled ChEMBL compounds and from the screening resul

5.3.3 Similarity to ChEMBL active compounds and ADMET screening

Similar to the first screening, similarity search to ChEMBL active compounds [69] and ADMET screening using ADMETlab2.0 [62] were done. 62 compounds from the active-labeled HDAC8 dataset were passed and they would be sorted using their docking score next.

5.3.4 Docking benchmark and sorting with docking score

This time, docking score was used to sort the compounds. Druglikeness was not used this time because screening through ADMET was decided to be enough to filter compounds that have low drug-like properties. Moreover, Druglikeness score is not a target-specific property, which I think is important in this stage, since the purpose is to find a new inhibitor for a specific target. Before sorting, benchmarking needed to be done to the docking program, in this case Glide, as described in previous chapter. The decoys and active structures were obtained from DUD-E website, which is a database for decoys for many protein target, including HDAC8 [70].

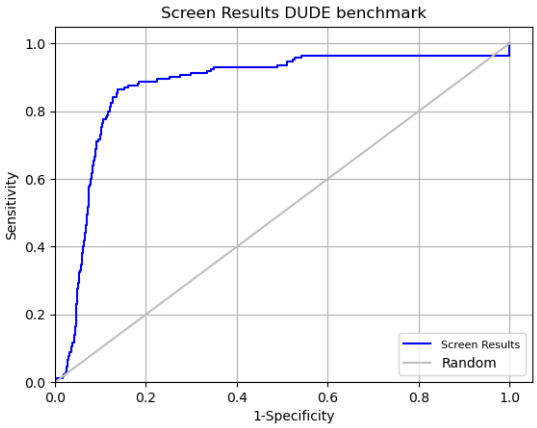


Figure 25: ROC AUC score of docking benchmark. ROC AUC value was calculated to be 0.87

Figure 25 shows the ROC AUC curve and score of 0.87 for the benchmark based on their docking score in Glide. This score was deemed to be acceptable so docking score was used as sorting to get the top 50 compounds. Table 11 shows the list of 62 compounds from Osaka University library sorted by their docking score in HDAC8 crystal structure.

No	OUID	Docking score	No	OUID	Docking score
1	OU0020016	-10.83	32	OU0036739	-4.95
2	OU0053108	-9.75	33	OU0067441	-4.63
3	OU0040430	-7.89	34	OU0030193	-4.53
4	OU0057063	-7.72	35	OU0019458	-4.41
5	OU0015622	-7.68	36	OU0049240	-4.18
6	OU0056840	-7.37	37	OU0052119	-4.16
7	OU0027295	-7.02	38	OU0016021	-4.08
8	OU0029040	-6.99	39	OU0026938	-4.08
9	OU0067490	-6.93	40	OU0029105	-4.03
10	OU0022158	-6.93	41	OU0059356	-4.03
11	OU0036595	-6.71	42	OU0042656	-3.93
12	OU0025878	-6.56	43	OU0028831	-3.87
13	OU0054562	-6.39	44	OU0039851	-3.87
14	OU00035329	-6.27	45	OU0069101	-3.80
15	OU0043654	-6.27	46	OU0024378	-3.80
16	OU0031774	-6.26	47	OU0019528	-3.78
17	OU0028633	-6.21	48	OU0051274	-3.78
18	OU0031802	-6.03	49	OU0038447	-3.66
19	OU0054020	-5.96	50	OU0021493	-3.55
20	OU0027253	-5.90	51	OU0051823	-3.52
21	OU0024597	-5.86	52	OU0042499	-3.40
22	OU0016126	-5.80	53	OU0015547	-3.39
23	OU0026909	-5.69	54	OU0037759	-3.37
24	OU0046217	-5.62	55	OU0069078	-3.27
25	OU0067411	-5.40	56	OU0051813	-3.25
26	OU0067626	-5.40	57	OU0068718	-3.18
27	OU0043908	-5.34	58	OU0053557	-3.09
28	OU0061261	-5.34	59	OU0064484	-2.97
29	OU0025848	-5.27	60	OU0021393	-2.95
30	OU0036709	-5.18	61	OU0023736	-2.67
31	OU0024945	-5.13	62	OU0065772	-1.56

Table 11: Docking score

Docking score was used to support the machine learning and did not become the main method for screening is because the best ranked solution according to the scoring system is not always the pose that was produced in experimental method [72]. So it was thought that docking would be better to be used after other filtering methods were done.

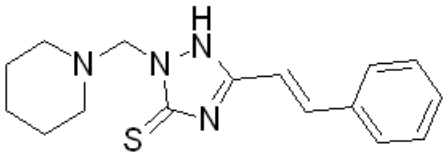
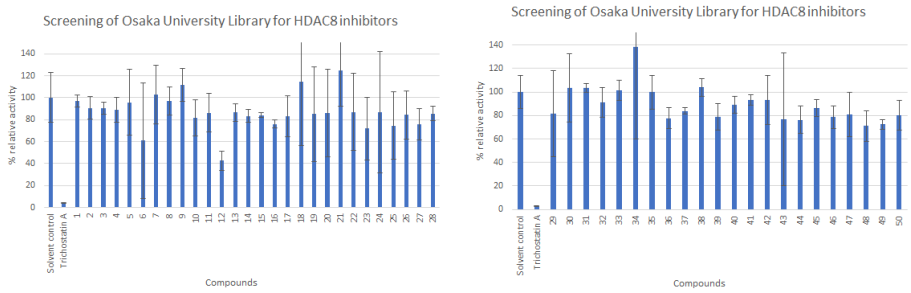


Figure 27: Structure of compound 12

5.3.5 HDAC8 inhibitory activity screening

The top 50 compounds were then subjected into HDAC8 inhibitory activity measurement. The result is shown on Fig 26. Compound 12 (Fig 27) was found to have the best inhibitory activity, with it being able to reduce the activity of HDAC8 to around 40% in 10 μM.



(a) (b)

Figure 26: HDAC8 inhibitory activity of top 50 compounds from the second screening. Concentration of all compounds was set to 10 μM

IC50 measurement was done for both compounds. Compound 12 IC50 on HDAC8 was around 842 $\pm$ 165 nM (Fig 28).

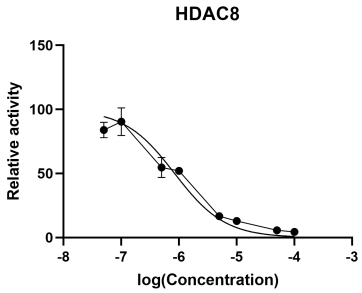


Figure 28: HDAC8 inhibition curve of compound 12

5.3.6 Docking Pose Study of Compound 6 and 12 on HDAC8

Docking pose study was done to see the interaction of compound 12 with HDAC8. Based on the pose, compound 12 could enter HDAC8 active pocket (Fig. 29a) and the thioltriazole was able to interacted with zinc ion in the active pocket and this interaction is also stabilized by the pi-pi interaction of the triazole with histidine residues (Fig 29c).

5.3.7 Selectivity of Compound 12 on Other HDACs

Compound 12 were screened on the models of other HDACs, specifically HDAC1, -3, and -6 to check if these models can predict the activity of both compounds on the respective HDACs. HDAC1 and HDAC3 were chosen because, along with HDAC8, they are members of the Class I HDAC. Table 12 shows the results of the screening. Compound 12 was predicted to be inactive in all the tested models shown, meaning the models predicted that Compound 12 would have pIC50 more than or equal to 7 on those HDACs.

.

Compound No	OUID	HDAC1	HDAC3
12	OU0024597	inactive	inactive

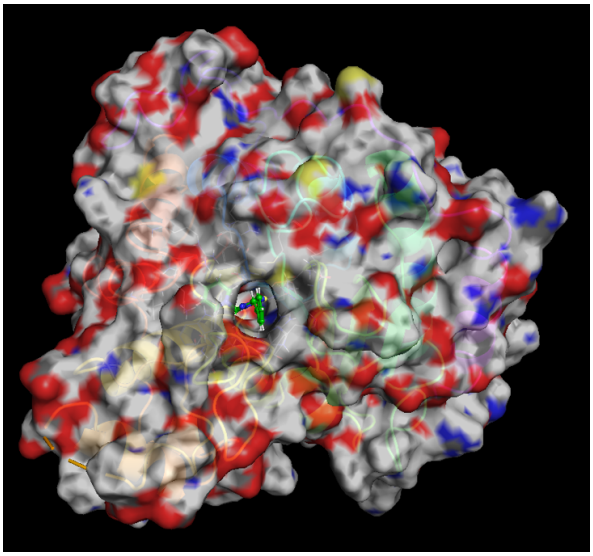
Table 12: Predicted activity of compound 6 and 12 on other HDAC models

To test this prediction, the IC50 of compound 12 on HDAC1, and HDAC3 were measured and compound 12 was found to be active on other HDACs, with IC50 of 42 μM and 24 μM for HDAC1 and HDAC3 respectively, and IC50 of less than 10 nM for HDAC4 (Table 13). The reason why compound 12 was predicted to be inactive on other HDACs will be discussed in the later part.

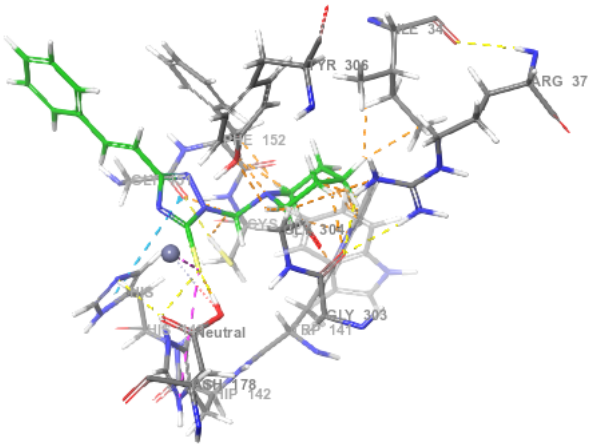
Compound No	OUID	HDAC1	HDAC8:HDAC1	HDAC3	HDAC8:HDAC3
12	OU0024597	38 $\pm$ 17 μM	0.02:1	12 $\pm$ 0.6 μM	0.07:1

Table 13: IC50 values of compound 12 on HDAC1 and HDAC3 with theIC50 ratio of with HDAC8 IC50

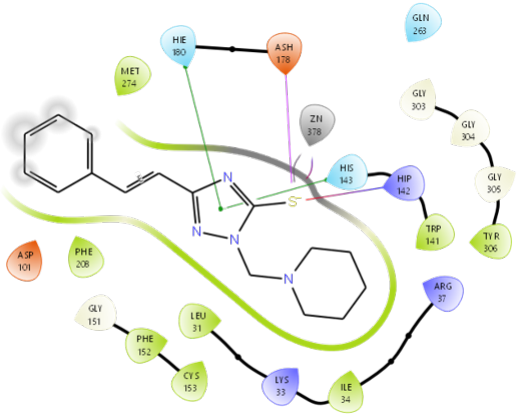
Although compound 12 passed the PAINS screening from ADMETlab2.0, to solidifying this result, the activity of compound 12 on HDAC assay developer, which is most likely trypsin, was done. As it can be seen on Figure 30, compound 12 does not have activity on the developer, which means that compound 12 is selective to HDAC.



(a) Docking pose of compound 12 with HDAC8



(b)



(c) 2D interaction of compound 12 with HDAC8

Figure 29: Docking pose and interaction of compound 12 with HDAC8.

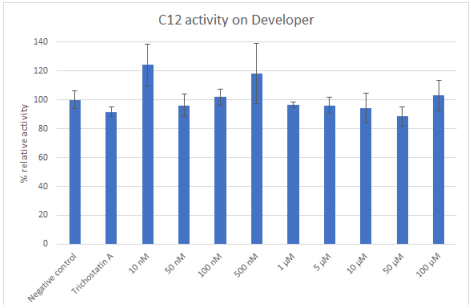


Figure 30: Compound 12 activity of HDAC assay developer

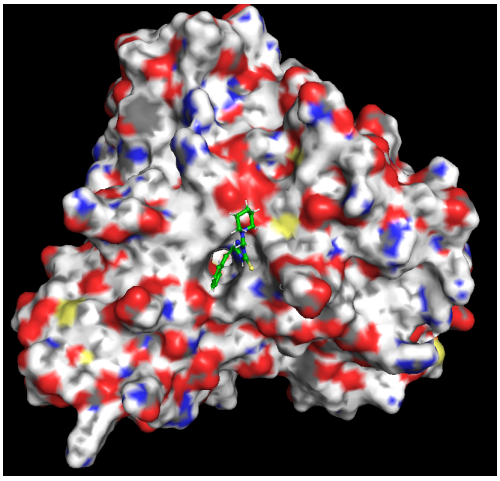
5.3.8 Docking study of Compound 12 on HDAC1 and HDAC3

Docking study was done to observe the interaction between compound 12 and other HDACs. Based on the docking pose, compound 12 does not have interaction with HDAC1 which might explain the high IC50 of compound 12 against HDAC1 (Fig 31). As for HDAC3, compound 12 also sits on the entrance of the active pocket of HDAC3 and this interaction was supported with pi-pi interaction with Phe144 and Asp93 (Fig 32).

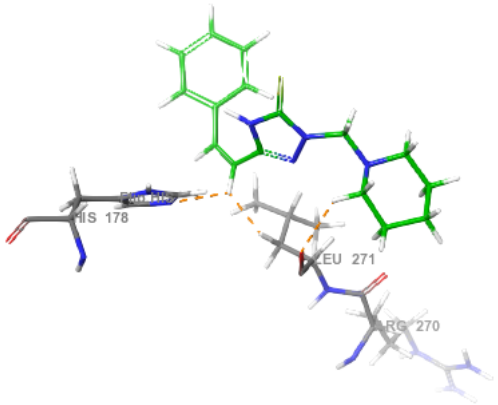
5.3.9 HDAC8 Model Studies

The usage of random forest as model in work allowed me to observe its feature importance. In this case, feature importance shows which variables are the most important to make prediction and tree split. The scoring system used to compare the variables is called Mean Decrease Impurity which calculates each feature importance as the sum over the number of splits across all tres that include the feature [73]. Since each bit of PubChem is corresponding to information about the substructure of a molecule, it is worth to check the feature importance to see which substructure information that are important for HDAC8 inhibition activity of a molecule.

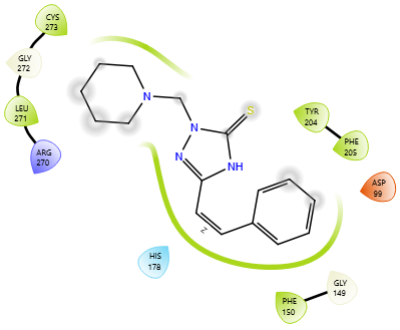
Figure 33a shows the top 10 features from the RF-SMOTE model of HDAC8, with their details explained in Table 14. Based on this, while the top 2 bits clearly are the most important, but the effect of other bits in the top 10 are not negligible. To show their relationship, pair plot of these top 10 features was constructured (Fig 33). From the first glance of the pair plot, the top 10 features could not differentiate between active (orange) and inactive (blue) compounds cleanly, but some combinations of features could differentiate a good amount of inactive compounds from the rest of the dataset. For example, while the presence and/or absence of bit 392 and/or 374 can not differentiate active or inactive molecules, but a molecule could be active if it either has both O=C-N-C-C fragment and N(~C)(~C)(~H) pattern or none of them, which suggest that in this case, the presence of N(~C)(~C)(~H) pattern in active compounds is from the O=C-N-C-C fragment. Other bit worth noting is bit 12, that gives a clue on the size of the inhibitors. in which the majority of the active compounds have carbon atoms equal or more than 16 and only a small amount of active compounds have less than 16 carbon atoms. Other bit which presence can be observed in many of the active molecules is



(a) Docking pose of compound 12 on HDAC1

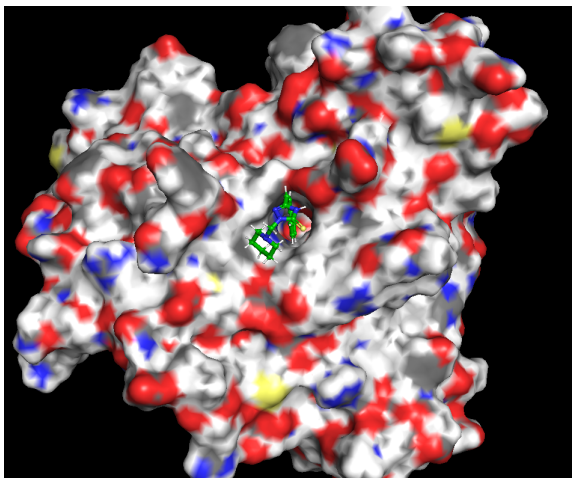


(b)

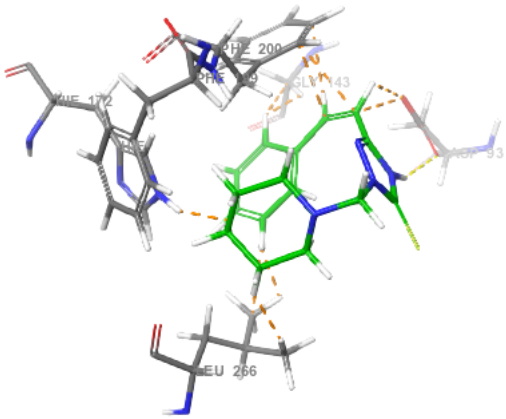


(c) 2D interaction diagram of compound 12 against HDAC1

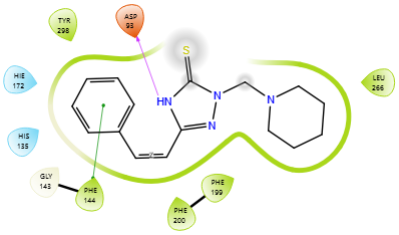
Figure 31: Docking pose and interaction diagram of compound 12 against HDAC1



(a) Docking pose of compound 12 on HDAC3



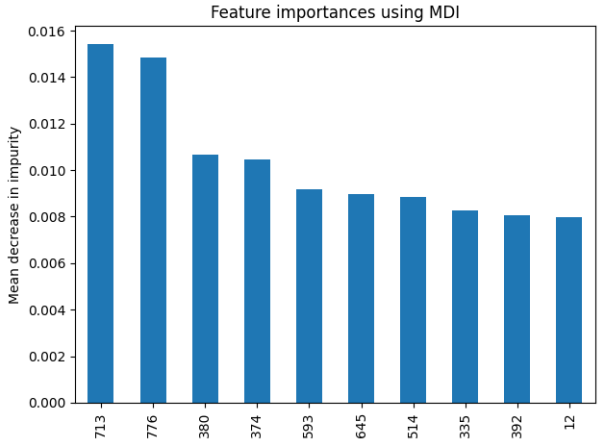
(b)



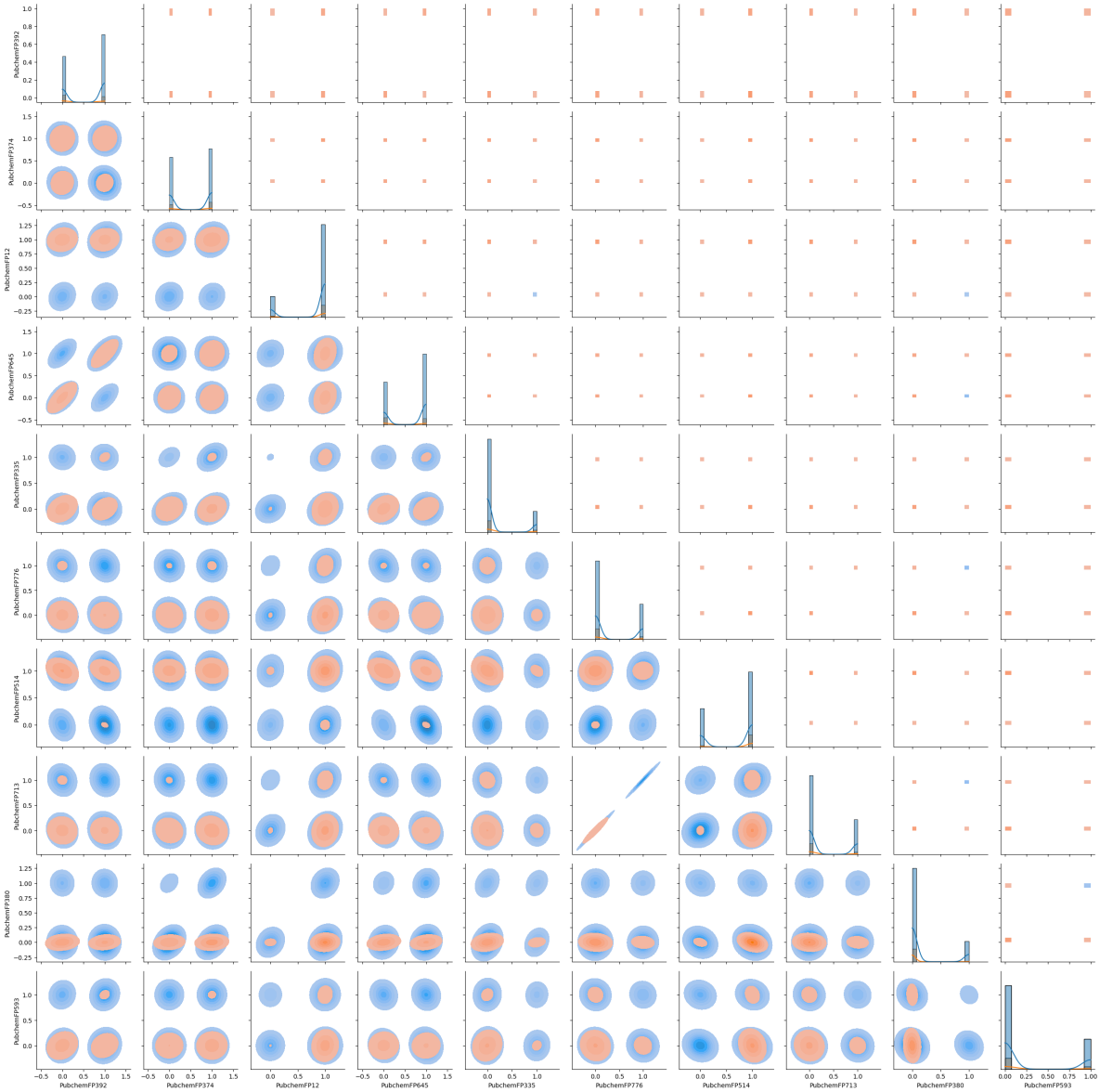
(c) 2D interaction diagram of compound 12 against HDAC3

Figure 32: Docking pose and interaction diagram of compound 12 against HDAC3

bit 514 (SMARTS pattern O-N-C-C). This pattern most likely indicates the presence of hydroxamic acid moeity in the structures. Meanwhile, the top 2 features, bit 713 and bit 776, the reason why they were placed high in this feature importance is that none of the active ChEMBL compounds do not have these two fragments, Cc1ccc(C)cc1 and CC1CCC(C)CC1 respectively. The absence of these fragments in the active molecules might be because the pocket of HDAC8 is L-shaped and so having phenyl and cyclohexane fragment with two substituent in the para positions would not fit into HDAC8 pocket. The presence of bit 645 (SMART pattern of O=C-N-C-C) in active compounds seems to be dependent of other bits, for example, the presence of both bit 645 and 514 in active amolecule might suggest the presence of hydroxamate acid, but since not all active compounds bearing this moeity, the absence of either one of them suggested the other fragment is from a non-hydroaxmic acid molecules.



(a)



(b)

Figure 33: Feature importance. (a) shows 10 most important features from the HDAC8 model and (b) shows the pair plot of those 10 features against each other along with the distribution of active (orange) and inactive (blue) compounds from ChEMBL library.

No	Bit position	Bit information
1	713	Cc1ccc(C)cc1 (SMARTS pattern)
2	776	CC1CCC(C)CC1 (SMARTS pattern)
3	380	C(~O)(~O)
4	374	C(~H)(~H)(~H)
5	593	N-C-C-C-N (SMARTS pattern)
6	645	O=C-N-C-C (SMARTS pattern)
7	514	O-N-C-C (SMARTS pattern)
8	335	C(~C)(~C)(~C)(~H)
9	392	N(~C)(~C)(~H)
10	12	>= 16 C

Table 14: Bit information from the top 10 most important features from HDAC8 model

Comparing the pattern these top 10 features of the active compound and compound 12 (Table 15), Other than bit 12 which describes the size of compound 12, compound 12 does not have the other substructures, which makes sense, since many of the top 10 bits decribes the presence of hydroxamic acid moeity. This shows that the model could identify non-hydroxamic acid inhibitors.

Bit position									
713	776	380	374	593	645	514	335	392	12
0	0	0	0	0	0	0	0	0	1

Table 15: Values of compound 12's top 10 important features from HDAC8 model

Another way to visualize the capability of the model to predict the molecule is by looking at the chemical space of the active compounds. Figure 34 shows the chemical space of ChEMBL active (black), compounds from Osaka University library that were labeled active (blue), the top 50 compounds from second screening (green), and compound 6 (yellow), as well as compound 12 (red). The screened compounds, including compound 12, manage to explore chemical space beyond the known active compounds. This shows that the model, using SMOTE could be used to explore the chemical space beyond known active compounds.

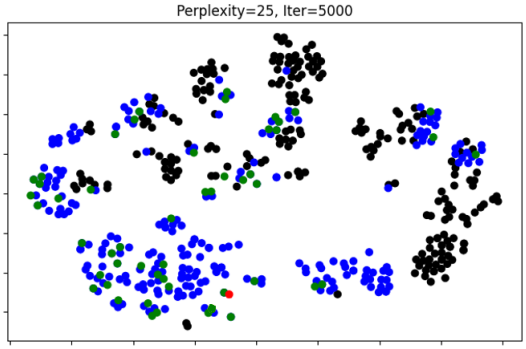
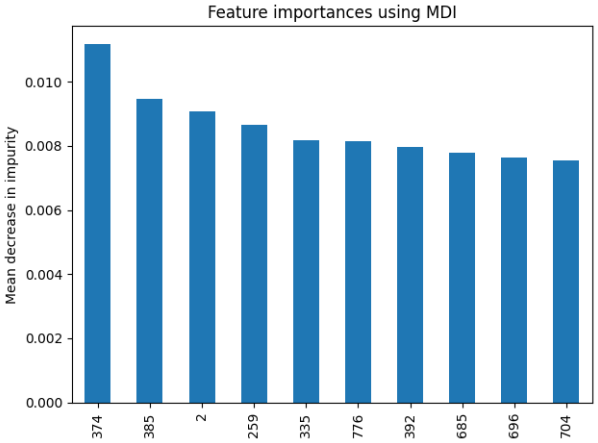


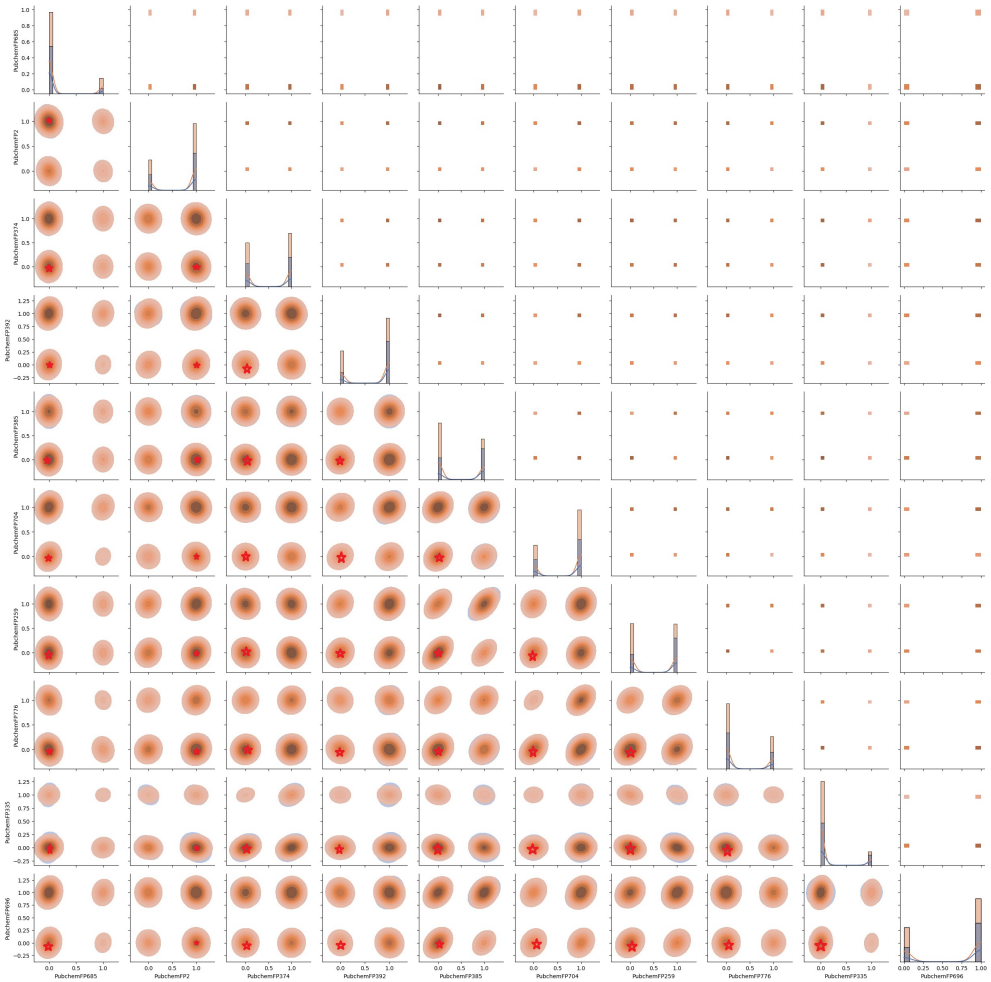
Figure 34: Chemical space of HDAC8 actives

5.3.10 Studies of the Prediction of Compound 12 on other HDAC models

In this section, the possible reasoning of why compound 12 was predicted to be inactive again HDAC1 and HDAC3 will be introduced using the feature importance of each model (Fig 35-36). For HDAC1 model, it seems there is no one feature or a combination of them that could differentiate whether a molecule is active (orange) or inactive (blue) (Fig 35). This could cause the difficulty in tree splitting and also difficulty in predicting new dataset, which might explain why compound 12 was wrongly predicted to be inactive against HDAC1. For HDAC3 model, similar thing happened, because compound 12 does not have C(~H)(~H)(~H) pattern, which was also seem to be important for this model (Fig 36), as well as the absence of bit 643, which is present in the majority of HDAC3 active compounds (Fig 37 and 38).



(a)



(b)

Figure 35: The top 10 important features of HDAC1 model and their pair plot. Red stars show the placement of compound 12 based on its fingerprints.

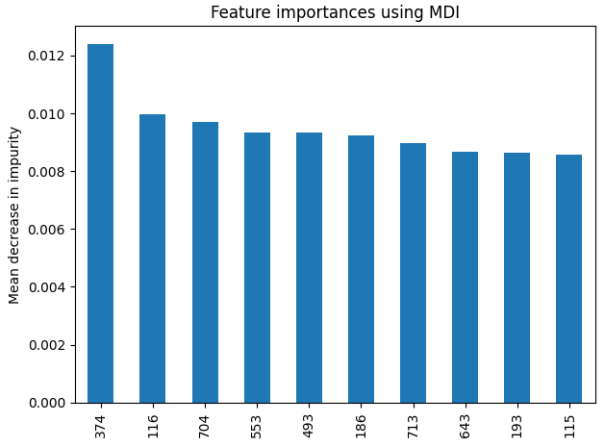


Figure 36: Top 10 feature importance of HDAC3 model

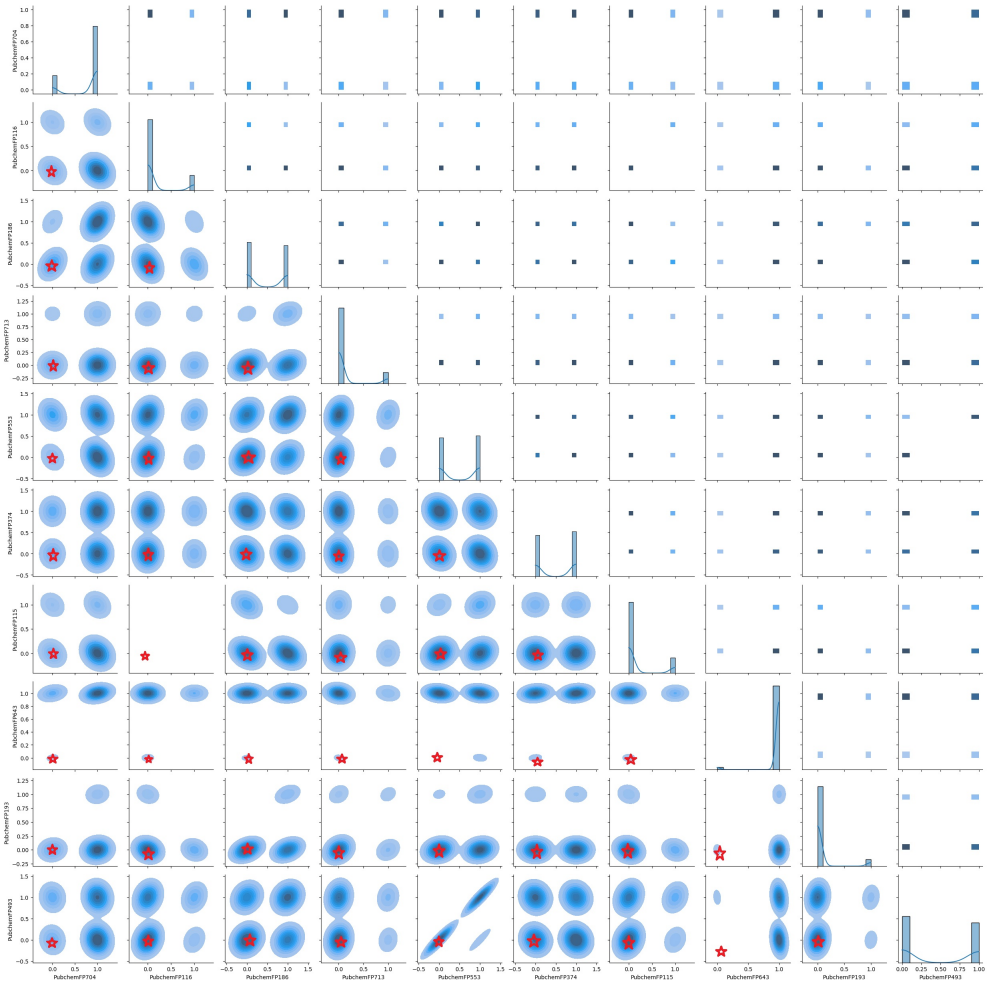


Figure 37: Pairplot of the top 10 feature importance from HDAC3 model on HDAC3 active-labeled compounds from ChEMBL dataset. Red stars show the placement of compound 12 based on its fingerprints.

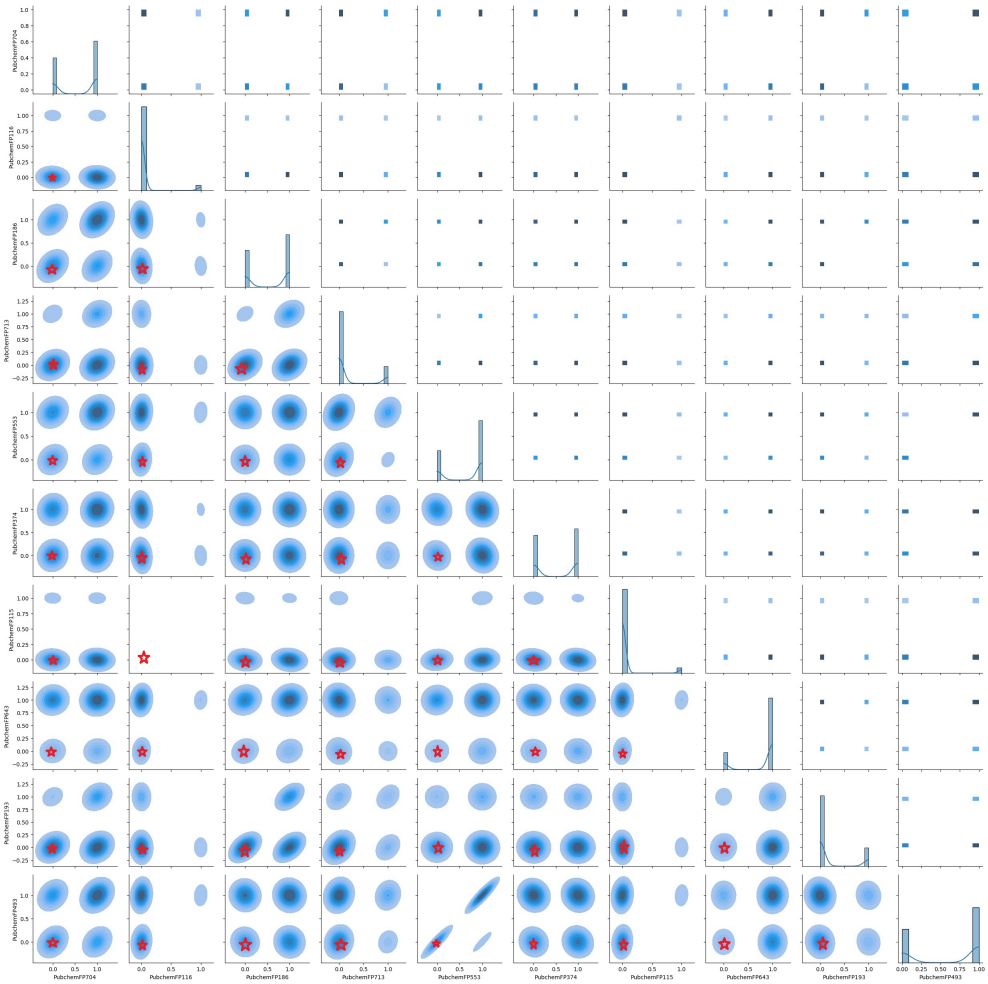


Figure 38: Pairplot of the top 10 feature importance from HDAC3 model on HDAC3 inactive-labeled compounds from ChEMBL dataset. Red stars show the placement of compound 12 based on its fingerprints.

6 Conclusion

In conclusion, using SMOTE to solve the data imbalance problem, random forest model was built to find new HDAC8 inhibitor. After screening and filtering, compound 12 was found to have HDAC8 inhibitory activity with IC50 of 842 nM. Moreover, compound 12 was also found to have inhibitory effect on HDAC1 and HDAC3 as well, with IC50 of 38 μM and 12 μM respectively, making compound 12 to be selective toward HDAC8 among Class I HDAC. After examining each models for their feature importance and the chemical space, it was shown that the using SMOTE, the model could explore a wider chemical space to find new inhibitor candidate for HDACs.

References

[1] D. Sun et al. Why 90% of clinical drug development fails and how to improve it? *Acta Pharmaceutica Sinica B*, 12(7):3049–3062, jul 2022.

[2] H. Dowden and J. Munro. Trends in clinical success rates and therapeutic focus. *Nature Reviews Drug Discovery*, 18(7):495–496, may 2019.

[3] R. K. Harrison. Phase II and phase III failures: 2013–2015. *Nature Reviews Drug Discovery*, 15(12):817–818, nov 2016.

[4] A. D. Goldberg et al. Epigenetics: A landscape takes shape. *Cell*, 128(4):635–638, feb 2007.

[5] C. Waddington. *The Strategy of the Genes*. Routledge, apr 1957.

[6] E. Seto and M. Yoshida. Erasers of histone acetylation: The histone deacetylase enzymes. *Cold Spring Harbor Perspectives in Biology*, 6(4):a018713–a018713, apr 2014.

[7] J. J. Buggy et al. Cloning and characterization of a novel human histone deacetylase, HDAC8. *Biochemical Journal*, 350(1):199–205, aug 2000.

[8] J. R. Somoza et al. Structural snapshots of human HDAC8 provide insights into the class i histone deacetylases. *Structure*, 12(7):1325–1334, jul 2004.

[9] M. S. Finnin et al. Structures of a histone deacetylase homologue bound to the TSA and SAHA inhibitors. *Nature*, 401(6749):188–193, sep 1999.

[10] R. Wu et al. A proton-shuttle reaction mechanism for histone deacetylase 8 and the catalytic role of metal ions. *Journal of the American Chemical Society*, 132(27):9471–9479, jun 2010.

[11] A. Chakrabarti et al. HDAC8: a multifaceted target for therapeutic interventions. *Trends in Pharmacological Sciences*, 36(7):481–492, jul 2015.

[12] C. Schölz et al. Acetylation site specificities of lysine deacetylase inhibitors in human cells. *Nature Biotechnology*, 33(4):415–423, mar 2015.

[13] J. Wu et al. The up-regulation of histone deacetylase 8 promotes proliferation and inhibits apoptosis in hepatocellular carcinoma. *Digestive Diseases and Sciences*, 58(12):3545–3553, sep 2013.

[14] M. A. Deardorff et al. HDAC8 mutations in cornelia de lange syndrome affect the cohesin acetylation cycle. *Nature*, 489(7415):313–317, aug 2012.

[15] B. J. Wilson et al. An acetylation switch modulates the transcriptional activity of estrogen-related receptor . *Molecular Endocrinology*, 24(7):1349–1358, jul 2010.

[16] M. A. Deardorff et al. Structural aspects of HDAC8 mechanism and dysfunction in cornelia de lange syndrome spectrum disorders. *Protein Science*, 25(11):1965–1976, sep 2016.

[17] C. Park. Histone deacetylases 1, 6 and 8 are critical for invasion in breast cancer. *Oncology Reports*, mar 2011.

[18] M. Nakagawa et al. Expression profile of class i histone deacetylases in human cancer tissues. *Oncology Reports*, oct 2007.

[19] T. A. Miller et al. Histone deacetylase inhibitors. *Journal of Medicinal Chemistry*, 46(24):5097–5116, oct 2003.

[20] V. M. Richon. Cancer biology: mechanism of antitumour action of vorinostat (suberoylanilide hydroxamic acid), a novel histone deacetylase inhibitor. *British Journal of Cancer*, 95(S1):S2–S6, dec 2006.

[21] A. D. Bondarev et al. Recent developments of HDAC inhibitors: Emerging indications and novel molecules. *British Journal of Clinical Pharmacology*, 87(12):4577–4597, may 2021.

[22] L. Zhang et al. Zinc binding groups for histone deacetylase inhibitors. *Journal of Enzyme Inhibition and Medicinal Chemistry*, 33(1):714–721, jan 2018.

[23] S. A. Kavanaugh et al. Vorinostat: A novel therapy for the treatment of cutaneous t-cell lymphoma. *American Journal of Health-System Pharmacy*, 67(10):793–797, may 2010.

[24] P. Sandhu et al. Disposition of vorinostat, a novel histone deacetylase inhibitor and anticancer agent, in preclinical species. *Drug Metabolism Letters*, 1(2):153–161, apr 2007.

[25] M. Iwamoto et al. Clinical pharmacology profile of vorinostat, a histone deacetylase inhibitor. *Cancer Chemotherapy and Pharmacology*, 72(3):493–508, jul 2013.

[26] A. G. Kazantsev and L. M. Thompson. Therapeutic application of histone deacetylase inhibitors for central nervous system disorders. *Nature Reviews Drug Discovery*, 7(10):854–868, oct 2008.

[27] S. Shen and A. P. Kozikowski. Why hydroxamates may not be the best histone deacetylase inhibitors-what some may have forgotten or would rather forget? *ChemMedChem*, 11(1):15–21, nov 2015.

[28] M. Yoshida et al. Potent and specific inhibition of mammalian histone deacetylase both in vivo and in vitro by trichostatin a. *Journal of Biological Chemistry*, 265(28):17174–17179, oct 1990.

[29] M.-S. Lee and M. Isobe. metabolic activation of the potent mutagen, 2-naphthohydroxamic acid, in salmonella typhimurium ta98. *Cancer research*, 50(14):4300–4307, 1990.

[30] L. Ducháčková and J. Roithová. The interaction of zinc(II) and hydroxamic acids and a metal-triggered lossen rearrangement. *Chemistry - A European Journal*, 15(48):13399–13405, dec 2009.

[31] L. Jašíková et al. Metal-assisted lossen rearrangement. *The Journal of Organic Chemistry*, 77(6):2829–2836, mar 2012.

[32] P. Galletti et al. Azetidinones as zinc-binding groups to design selective HDAC8 inhibitors. *ChemMedChem*, 4(12):1991–2001, oct 2009.

[33] E. Hu et al. Identification of novel isoform-selective inhibitors within class i histone deacetylases. *Journal of Pharmacology and Experimental Therapeutics*, 307(2):720–728, sep 2003.

[34] A. Géron. *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow Concepts, Tools, and Techniques to Build Intelligent Systems*. O'Reilly Media, second edition, 2019.

[35] A. C. Muller and S. Guido. *Introduction to machine learning with Python : a guide for data scientists*. O'Reilly Media, 2017.

[36] J. Joyce. Bayes' Theorem. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Fall 2021 edition, 2021.

[37] F. R. S. Bayes. LII. an essay towards solving a problem in the doctrine of chances. by the late rev. mr. bayes, f. r. s. communicated by mr. price, in a letter to john canton, a. m. f. r. s. *Philosophical Transactions of the Royal Society of London*, 53:370–418, dec 1763.

[38] H. Zhang. The optimality of naive bayes. *Aa*, 1(2):3, 2004.

[39] Y. SUN et al. CLASSIFICATION OF IMBALANCED DATA: A REVIEW. *International Journal of Pattern Recognition and Artificial Intelligence*, 23(04):687–719, jun 2009.

[40] G. M. Weiss. Mining with rarity. *ACM SIGKDD Explorations Newsletter*, 6(1):7–19, jun 2004.

[41] N. V. Chawla et al. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, jun 2002.

[42] O. J. Wouters et al. Estimated research and development investment needed to bring a new medicine to market, 2009-2018. *JAMA*, 323(9):844, mar 2020.

[43] J. Vamathevan et al. Applications of machine learning in drug discovery and development. *Nature Reviews Drug Discovery*, 18(6):463–477, apr 2019.

[44] T. Wills. Ai drug discovery: assessing the first ai-designed drug candidates to go into human clinical trials, September 2022.

[45] A. Philippidis. Insilico's ai candidate for ipf doses first patient in phase ii, June 2023.

[46] Cas assigns the 100 millionth cas registry numbero a substance designed to treat acute myeloid leukemia, June 2015.

[47] R. Macarron et al. Impact of high-throughput screening in biomedical research. *Nature Reviews Drug Discovery*, 10(3):188–195, mar 2011.

[48] S. J. Lusher et al. Data-driven medicinal chemistry in the era of big data. *Drug Discovery Today*, 19(7):859–868, jul 2014.

[49] B. Ramsundar et al. *Deep Learning for the Life Sciences*. O'Reilly Media, 2019. https://www.amazon.com/Deep-Learning-Life-Sciences-Microscopy/dp/1492039837.

[50] M. Korshunova et al. OpenChem: A deep learning toolkit for computational chemistry and drug design. *Journal of Chemical Information and Modeling*, 61(1):7–13, jan 2021.

[51] D. Mendez et al. ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Research*, 47(D1):D930–D940, nov 2018.

[52] M. Davies et al. ChEMBL web services: streamlining access to drug discovery data and utilities. *Nucleic Acids Research*, 43(W1):W612–W620, apr 2015.

[53] National Center For Biotechnology Information. Pubchem subgraph fingerprint, May 2009.

[54] J.-L. Reymond and M. Awale. Exploring chemical space for drug discovery using the chemical universe database. *ACS Chemical Neuroscience*, 3(9):649–657, may 2012.

[55] J.-L. Reymond et al. The enumeration of chemical space. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 2(5):717–733, apr 2012.

[56] L. Van der Maaten and G. Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[57] A. Bender and R. C. Glen. Molecular similarity: a key technique in molecular informatics. *Organic &amp Biomolecular Chemistry*, 2(22):3204, 2004.

[58] D. Bajusz et al. Why is tanimoto index an appropriate choice for fingerprint-based similarity calculations? *Journal of Cheminformatics*, 7(1), may 2015.

[59] E. Nelson. Kinetics of drug absorption, distribution, metabolism, and excretion. *Journal of Pharmaceutical Sciences*, 50(3):181–192, mar 1961.

[60] M. P. Doogue and T. M. Polasek. The ABCD of clinical pharmacokinetics. *Therapeutic Advances in Drug Safety*, 4(1):5–7, jan 2013.

[61] H. van de Waterbeemd and E. Gifford. ADMET in silico modelling: towards prediction paradise? *Nature Reviews Drug Discovery*, 2(3):192–204, mar 2003.

[62] G. Xiong et al. ADMETlab 2.0: an integrated online platform for accurate and comprehensive predictions of ADMET properties. *Nucleic Acids Research*, 49(W1):W5–W14, apr 2021.

[63] X.-Y. Meng et al. Molecular docking: A powerful approach for structure-based drug discovery. *Current Computer Aided-Drug Design*, 7(2):146–157, jun 2011.

[64] R. A. Friesner et al. Glide: a new approach for rapid, accurate docking and scoring. 1. method and assessment of docking accuracy. *Journal of Medicinal Chemistry*, 47(7):1739–1749, feb 2004.

[65] W. McKinney. Data structures for statistical computing in python. 2010.

[66] C. W. Yap. PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. *Journal of Computational Chemistry*, 32(7):1466–1474, dec 2010.

[67] F. Pedregosa et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[68] G. Lemaître et al. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research*, 18(17):1–5, 2017.

[69] T. Sander et al. DataWarrior: An open-source program for chemistry aware data visualization and analysis. *Journal of Chemical Information and Modeling*, 55(2):460–473, feb 2015.

[70] M. M. Mysinger et al. Directory of useful decoys, enhanced (DUD-e): Better ligands and decoys for better benchmarking. *Journal of Medicinal Chemistry*, 55(14):6582–6594, jul 2012.

[71] E. E. Bolton et al. PubChem: Integrated platform of small molecules and biological activities. pp. 217–241, 2008.

[72] D. Ramírez and J. Caballero. Is it reliable to take the molecular docking top scoring position as the best solution without considering available structural data? *Molecules*, 23(5):1038, apr 2018.

[73] G. Louppe et al. Understanding variable importances in forests of randomized trees. In C. Burges et al., editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.