



|              |                                                                             |
|--------------|-----------------------------------------------------------------------------|
| Title        | Drug discovery study on HDAC8 inhibitors using machine learning             |
| Author(s)    | Nurani, Atika                                                               |
| Citation     | 大阪大学, 2023, 博士論文                                                            |
| Version Type | VoR                                                                         |
| URL          | <a href="https://doi.org/10.18910/93025">https://doi.org/10.18910/93025</a> |
| rights       | © 2024 The Pharmaceutical Society of Japan                                  |
| Note         |                                                                             |

*The University of Osaka Institutional Knowledge Archive : OUKA*

<https://ir.library.osaka-u.ac.jp/>

The University of Osaka

## Abstract of Thesis

Name ( Atika Nurani )

Title

Drug discovery study on HDAC8 inhibitors using machine learning  
(機械学習を用いたHDAC8阻害薬の探索)

## Abstract of Thesis

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 (Figure 1), compound **12**

(Figure 2) was found to have inhibition activity on HDAC8

with IC<sub>50</sub> of around 842 nM. Compound **12** was also found to have inhibitory activity against HDAC1 and HDAC3 with IC<sub>50</sub> 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 (Figure 3).

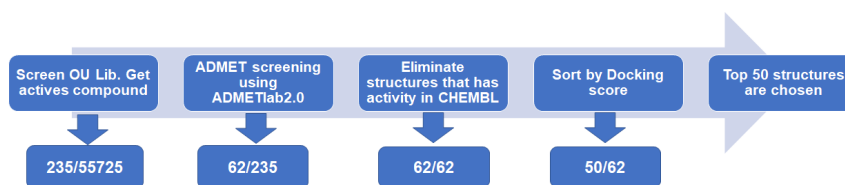


Figure 1. Flowchart of screening new HDAC8 inhibitors

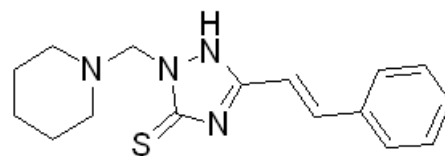
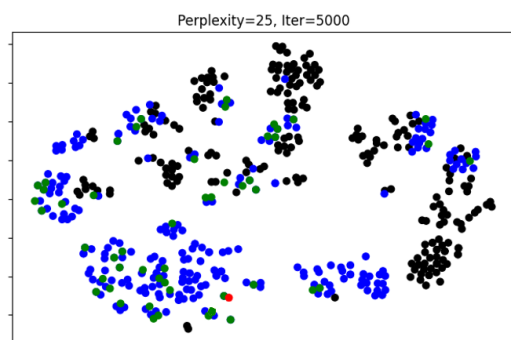
Figure 2. Structure of compound **12**

Figure 3. Chemical space from known compound (black), screening results (blue and green), and compound **12** (red). The use of SMOTE could expand chemical space beyond known active compounds

## 論文審査の結果の要旨及び担当者

| 氏 名 ( Atika Nurani ) |     |     |       |
|----------------------|-----|-----|-------|
| 論文審査担当者              |     | (職) | 氏 名   |
|                      | 主 査 | 教授  | 鈴木 孝禎 |
|                      | 副 査 | 教授  | 中谷 和彦 |
|                      | 副 査 | 教授  | 深瀬 浩一 |

## 論文審査の結果の要旨

HDAC8 阻害薬は、腫瘍細胞増殖阻害活性を示すことが知られており、ケミカルバイオロジー研究のツール分子および抗がん剤としての応用が期待されている。これまでに報告された HDAC8 阻害薬のほとんどは、ヒドロキサム酸構造を有するが、ヒドロキサム酸系化合物には、変異原性や代謝安定性に懸念がある。変異原性の低さと生物学的利用能の高さの観点から、非ヒドロキサム酸系 HDAC8 阻害薬が望まれている。本研究では、機械学習を利用して、大阪大学の化合物ライブラリーから、非ヒドロキサム酸系 HDAC8 阻害薬を見出すことを目指した。

まず、種々の機械学習モデルを用い、HDAC8 阻害活性化合物の予測精度を検証した。化合物と生物活性が紐づけされた ChEMBL データベースから、HDAC8 阻害活性に関する 2023 個のデータを抽出したのち、 $pIC_{50} = 7$  を基準として活性化合物(218 化合物)と不活性化合物(1805 化合物)をラベル化し、学習データセットを作成した。化合物の活性予測には、活性の有無を二値分類として出力する判別モデルを採用した。作成した学習データセットに対し、代表的な判別モデルである決定木、ランダムフォレスト、サポートベクターマシン等種々の機械学習アルゴリズムを用いて検証を行ったが、全てのモデルにおいて、受信者操作特性 (Receiver Operating Characteristic, ROC) 曲線における AUC 値が 0.8 を下回った。これは、学習データセット中の活性化合物と不活性化合物の数のアンバランスが問題であった。そこで、SMOTE(Synthetic Minority Over-sampling Technique)を用いて、少数クラス(活性化合物)のオーバーサンプリングを利用することとした。その結果、複数の機械学習モデルにおいて、AUC 値が 0.9 を上回り、中でも、ランダムフォレストが最も高い精度で予測できることを見出した。さらに、本モデルは、他の HDAC に対しても  $AUC > 0.9$  の精度で化合物活性予測ができることがわかった。続いて、学習済みの活性予測モデルを用いて、阪大ライブラリーから、50 化合物をバーチャルヒット化合物として絞り込んだ。一次スクリーニングで得られた 50 化合物は、すべて不活性であったが、それらの疑陽性データを再学習させた後に行った二次スクリーニングで得られたバーチャルヒット 50 化合物の HDAC8 阻害活性評価を行った結果、 $IC_{50} = 842 \text{ nM}$  の HDAC8 阻害活性を示す非ヒドロキサム酸系化合物 **12** を見出した。非ヒドロキサム酸系化合物 **12** の HDAC1 阻害活性評価、HDAC3 阻害活性評価を行った結果、 $IC_{50}$  値は、それぞれ、 $38 \mu\text{M}$  and  $12 \mu\text{M}$  であり、化合物 **12** は、HDAC1 に対して 45 倍、HDAC3 に対して 14 倍の高い HDAC8 選択性を示した。化合物 **12** は、非ヒドロキサム酸系化合物の中では、最も高活性かつ高選択的な HDAC8 阻害薬であることが分かった。化学計算ソフトウェア Glide を用いて、結合様式解析を行った結果、化合物 **12** のチオケトン部分がチオエノール構造に互変異性化し、アニオン化した硫黄原子が HDAC8 の活性中心にある亜鉛イオンと相互作用することが示唆された。

本研究では、膨大なデータを注意深く扱い、機械学習により、HDAC8 阻害薬のケミカルスペースの拡大に貢献した。他の創薬標的に対するリガンドについても、同様の展開が期待できる。

よって、本論文は博士(理学)の学位論文として十分価値あるものと認める。