A machine learning approach for virtual screening of histone deacetylase inhibitor compounds using aromatic signatures

Alessandra Gomes Cioletti[1], Diego Mariano[1], Raquel Cardoso de Melo-Minardi[1]*

[1] Laboratory of Bioinformatics and Systems, Universidade Federal de Minas Gerais (UFMG), Belo Horizonte, Minas Gerais, Brazil.

raquelcm@dcc.ufmg.br

Abstract. *This work presents a machine learning approach for the virtual screening of histone deacetylase (HDAC) inhibitors, with a focus on the HDAC1 enzyme. Proposing a new structural signature based on the aromaticity of ligand atoms, the method models the protein-ligand binding region as a graph. Machine learning models were trained using this signature to distinguish ligands from decoys, achieving high accuracy. The methodology was applied to identify potential HDAC1 inhibitors from the T3DB database, resulting in the selection of compounds with binding potential. The results suggest that structural signatures are more efficient and computationally less expensive than molecular docking, thereby paving the way for the identification of HDAC inhibitors relevant to autism etiology studies.*

1. Introduction

Histone deacetylases (HDACs) are crucial enzymes that regulate gene expression. Their primary role is to catalyze the removal of acetyl groups from histone proteins, a modification that leads to chromatin condensation and altered accessibility of DNA to transcriptional machinery [Millard et al., 2013]. HDACs typically repress transcription but can also indirectly promote activation, depending on the cellular context and specific cofactors. They play a significant role not only in normal cellular physiology but also in the pathogenesis of various diseases. As a result, they have become an important focus of research across both clinical and physiological domains [Millard et al., 2013].

Recent research suggests a potential link between HDAC inhibition and autism spectrum disorders (ASD), sparking increased interest in how modulating HDAC activity may influence neurodevelopmental processes tied to autism [Sun et al., 2016]. Autism and ASD comprise a group of complex disorders that usually appear in early childhood, marked by deficits in social interaction, communication, and repetitive behaviors. These conditions are more common in boys and exhibit significant clinical and etiological variability, which complicates diagnosis and treatment [Waye; Cheng, 2018].

The inhibition of HDACs through interaction with toxins, pollution, pesticides, among others, may be related to several diseases, such as autism spectrum disorder, mainly affecting pregnant women. Previous studies demonstrated the use of molecular docking approaches to verify the binding between HDAC and small ligands [Cioletti et al., 2024]. Additionally, the authors demonstrated that graph-based structural signatures can guide machine learning in predicting novel inhibitors. However, the use of atomic structural signatures presented limitations, as atomic types cannot be easily attributed to all ligands.

In this paper, we propose the use of a new structural signature based on the atomic elements and the aromaticity of ligand atoms. Our signature models the binding

region of the protein-ligand complex as a graph, considering the counting of atomic element pairs and atoms present in aromatic rings. We present an algorithm for generating this signature and evaluate it through a case study involving the detection of toxins with inhibitory potential against Histone Deacetylase 1 (HDAC1).

2. Materials and methods

In this work, classical classification models were trained to obtain binding compounds that inhibit the histone deacetylase 1, based on a graph-based structural signature. After selecting the model, a virtual screening was performed using the T3DB database. Details of the methodology are described below.

2.1 Data Collection

We collected 184 ligands and 7,023 decoy molecules (we randomly selected 184 decoys to balance the database) from the MUDB-HDACs dataset [Xia et al., 2015]. The MUDB-HDACs dataset provides molecular properties, including logP, molecular weight, number of hydrogen acceptors, number of hydrogen donors, formal charge, and number of rotational bonds [Xia et al., 2015]. Protein-ligand complexes for this dataset generated by molecular docking were previously made available at [Cioletti et al., 2024].

Additionally, we selected the HDAC1 protein structure (PDB ID: 4BKX) [Millard et al., 2013], which is complexed with the ligand ACT (the original ligand has been removed). The remaining structure was used as the receptor template.

Lastly, we collected toxin data from the Toxin and Toxin Target Database (T3DB) [Lim et al., 2010]. This database contains 3,678 toxins, including clinically relevant organic and inorganic compounds, such as pollutants, pesticides, food toxins, and drugs. To use the database, the toxins were filtered based on the presence of a carbon atom in the molecule and the sum of the molecular formula indices greater than 4, to remove small molecules such as methane and inorganic compounds. A total of 2,878 toxin structures were obtained and saved in SMILES format.

2.2 Structural signatures

Before generating the structural signatures, the methodology described in the [Cioletti et al., 2024] study was initially performed. Each compound underwent molecular docking using the Autodock4Zn Vina protocol, specific for zinc metalloproteins. This protocol extends the Autodock force field by incorporating a potential that describes the energetic and geometric components of ligand interactions with zinc. Finally, a receptor complex with the first pose of each docking was obtained and saved in PDB format. In the case of toxins, 1,141 compounds were selected for virtual screening.

For each complex obtained in the previous step, graph-based structural signatures were generated using an algorithm derived from [Pires et al., 2013], which considers the atomic elements and the presence of carbons in aromatic rings. Hereafter, we refer to this novel signature as the Aromatic Signature (AS). Details of the algorithm will be presented in the results section. In the case study presented here, structural signature vectors were generated cumulatively using cutoff thresholds of 6 and 12 Å,

with a cutoff step of 1.0 Å. The signatures were generated for the complexes with their respective cavity cutoff, with atoms extracted at a distance of 10 Å from the zinc atom, and for the receptor and the first pose of each docking separately. The signatures were saved in a dataset in CSV format, with a label indicating the atom pair involved and the distance of the cutoff step.

2.3 Model training

For training, the structural signatures of the target protein were horizontally concatenated with the ligands/decoys/toxins, then a vector with the target categories (label encoder) was created. The remaining data were normalized using Robustscaler normalization or StandardScaler. The data were then partitioned into training and testing data (Split 80:20 for training and Split 70:30 for the search for the best hyperparameter).

We built machine learning models using the following algorithms: Naive Bayes (GaussianNB), Random Forest (RF), ExtraTrees, GradientBoosting (GradBoost), Adaboost, SVM, MLP, KNN, and Logistic Regression (LogitR). The models were trained using the Scikit-learn library, using k-fold cross-validation with k = 5. The method was evaluated based on accuracy.

Initially, to select the atoms involved in the aromatic signature and the cutoff threshold, the models were trained using the default model configuration (Supplementary Table S1). Then, a hyperparameter search was performed using grid search, including several additional models. The grid of tested hyperparameters is shown in Supplementary Table S2. After analyzing the results, the best model and hyperparameters were chosen for training and application in virtual screening.

3. Results and discussion

In this study, we propose a novel structural signature to represent protein-ligand interactions, specifically designed for the active site interactions of the HDAC protein. Structural signatures are vector representations of macromolecule features that can be used in machine learning tasks [Mariano et al., 2019]. The use of the aCSM algorithm for virtual screening of HDAC toxins has been previously proposed in the literature [Cioletti et al., 2024]. However, the results presented indicate possible overfitting of models created using the simple aCSM signature (training accuracy of ~90%, while test accuracy is ~70%). Therefore, proposing new signatures specifically refined for this protein could be beneficial for the rapid and scalable virtual screening of toxins that may affect HDAC.

Our signature proposal was based on the aCSM algorithm proposed by [Pires et al., 2013]. However, unlike the original aCSM algorithm, we only consider the atomic elements in the close bonding region and the presence of carbon atoms in aromatic rings. This signature runs through the entire structure and calculates the pairwise distribution of the CNOSA+X atoms, where CNOS corresponds to aliphatic carbon, nitrogen, oxygen, and sulfur atoms, respectively. Additionally, “A” corresponds to aromatic carbons, and “X” (any other atom). Using the "X" wildcard allows modeling the interaction of atoms not commonly found in proteins, such as zinc (Zn), found in the

HDAC site. The following pseudocode algorithm presents a representation of the structural signature:

Algorithm 1. Calculation of the aromatic signature

```
1. function aromatic_signature(protein, dmax, cutoff):
2.    for i in protein:
3.      for j in protein:
4.        dist_matrix[] <= calculate_dist(i, j)
5.        atom_class[] <= get_atom_pair_type(i, j)
6.
7.    for d in range(0, dmax, cutoff):
9.      ar_signa[] <= get_frequency(dist_matrix, atom_class, d)
10.
11.   return ar_signa
```

This code returns a numerical vector representing the complex's structure. This signature considers the distance distribution for each combination of CNOS+A+X atoms in the complex's target site. Next, we evaluate this signature by comparing the results with previous works.

3.1 Signature evaluation

To evaluate the signature, we collected the docked protein-ligand and protein-decoy complexes proposed by the study [Cioletti et al., 2024]. These complexes were derived from the MUDB-HDACs dataset [Xia et al., 2015]. We then extracted the atoms in the region near the active site. Finally, we generated the structural signatures using the algorithm proposed in this work.

To assess whether the signature could accurately represent protein-ligand complexes and differentiate them from protein-decoy complexes, we performed several machine learning experiments. Table 1 summarizes the results obtained.

Table 1 - Results and comparison to other works (AS: aromatic signature)

Model	Signature	Accuracy Train	Accuracy Test	Δtrain vs. test	Precision	ROC AUC	Source
RF	aCSM	0.918	0.716	0.202	0.719	0.786	[Cioletti et al., 2024]
MLP	aCSM	0.719	0.770	-0.051	0.781	0.771	[Cioletti et al., 2024]
ExtraTree	AS	0.976	0.851	0.125	0.875	0.946	This study.
SVM	AS	0.976	0.878	0.098	0.861	0.955	This study.
GaussianNB	AS	0.714	0.635	0.079	0.722	0.729	This study.
RF	AS	0.855	0.864	-0.009	0.864	0.865	This study.
Adaboost	AS	0.969	0.864	0.105	0.878	0.928	This study.
GradBoost	AS	0.983	0.838	0.145	0.871	0.953	This study.
MLP	AS	0.976	0.824	0.152	0.867	0.903	This study.
KNN	AS	0.966	0.810	0.156	0.839	0.861	This study.
LogitR	AS	0.959	0.851	0.108	0.833	0.881	This study.

The results presented in Table 1 indicate that the models produced with AS (Aromatic Signature) achieved superior accuracy in both training with cross-validation and testing for most algorithms tested. It is noteworthy that the new signature led to models with minor differences in accuracy, as observed in both training and testing, compared to the results of [Cioletti et al., 2024]. This indicates a lower possibility of overfitting. Indeed, these results suggest that the carbon analysis in aromatics and the atomic element analysis were good representatives of the HDAC active site, allowing the proposed signature to depict the characteristics of these structures accurately. Thus, we decided to apply the new proposed models for virtual screening.

3.2 Virtual screening

Finally, we applied the proposed signature with the ExtraTree model for virtual screening of toxins that could interact with HDAC. The model returned 31 possible ligands with reasonable probability. After applying the model and obtaining the top results, we reproduced the protein-toxin docking using the same docking protocols as those used in [Cioletti et al., 2024]. The docking results corroborate the model results, showing that the model built using our novel signature was able to detect potential ligands for HDAC. Table 2 summarizes the top five main results obtained.

Table 2. The top five toxins with a higher chance of HDAC inhibition returned by our approach (the complete list is available in the Supplementary Material)

toxinID	Probability - ligand	Docking Score	Name
T3DB_T2639	65.00%	-28.09	Aminopterin
T3DB_T946	63.00%	-29.38	Dibrompropamidine
T3DB_T1560	62.67%	-23.58	Pteroyl-D-glutamic acid
T3DB_T1454	62.00%	-28.22	Folic acid
T3DB_T2452	61.67%	-24.13	S-Adenosylhomocysteine

Table 2 shows the likelihood that the modeled substance is a ligand, the affinity score values obtained through docking, and the ranking of the best affinity scores. For example, the toxin Aminopterin showed the highest probability of being a true HDAC ligand. Figure 1 shows the structure of Aminopterin docked against HDAC (PDB: 4bkx). These results highlight the potential use of the proposed structural signature for the detection of possible toxins that may be related to autism-causing agents.

4. Conclusion

Using structural signatures can be a more efficient and computationally less costly process than molecular docking for separating ligands and decoys and predicting histone deacetylase inhibitors. Validating the results will require biological testing; however, many of the compounds identified by our method are in some way described or related to some of the multifactorial pathways involved in autism. Once proven, HDAC inhibition by these compounds could be a tool to aid in the study of autism etiology.

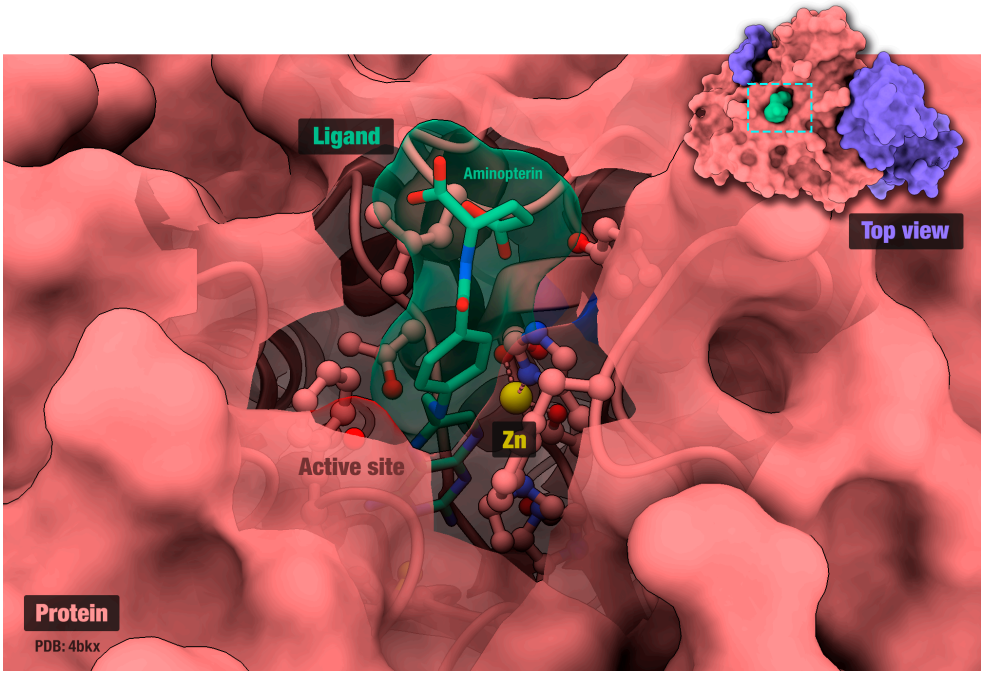

Figure 1. Complex formed by HDAC docked with Aminopterin

Acknowledgments: The authors thank the funding agencies: CAPES, FAPEMIG, and CNPq. This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

Supplementary material: Supplementary tables and figures used in this study are available at https://github.com/LBS-UFMG/HDAC-docking/tree/main/aromatic_signature.

Data availability: The dataset is accessible at https://github.com/jwxia2014/MUBD-HDACs.

Conflict of Interest: none declared.

5. References

CIOLETTI, Alessandra *et al.* Using graph-based structural signatures and machine learning algorithms for molecular docking assessment of histone deacetylases and small ligands. *In*: BRAZILIAN SYMPOSIUM ON BIOINFORMATICS. **Anais do XVII BSB**. Vitória, Brazil: 2 dec. 2024.

LIM, Emilia *et al.* T3DB: a comprehensively annotated database of common toxins and their targets. **Nucleic Acids Research**, v. 38, n. Database issue, p. D781-786, jan. 2010.

MARIANO, Diego *et al.* A Computational Method to Propose Mutations in Enzymes Based on Structural Signature Variation (SSV). **International Journal of Molecular Sciences**, v. 20, n. 2, 15 jan. 2019.

MILLARD, Christopher J. *et al.* Class I HDACs share a common mechanism of regulation by inositol phosphates. **Molecular Cell**, v. 51, n. 1, p. 57–67, 11 jul. 2013.

PIRES, Douglas E. V. *et al.* aCSM: noise-free graph-based signatures to large-scale receptor-based ligand prediction. **Bioinformatics (Oxford, England)**, v. 29, n. 7, p. 855–861, 1 apr. 2013.

SUN, Wenjie *et al.* Histone Acetylome-wide Association Study of Autism Spectrum Disorder. **Cell**, v. 167, n. 5, p. 1385- 1397.e11, 17 nov. 2016.

WAYE, Mary M. Y.; CHENG, Ho Yu. Genetics and epigenetics of autism: A Review. **Psychiatry and Clinical Neurosciences**, v. 72, n. 4, p. 228–244, apr. 2018.

XIA, Jie *et al.* Comparative modeling and benchmarking data sets for human histone deacetylases and sirtuin families. **Journal of Chemical Information and Modeling**, v. 55, n. 2, p. 374–388, 23 feb. 2015.