BRAIN. Broad Research in Artificial Intelligence and Neuroscience
ISSN: 2068-0473 | e-ISSN: 2067-3957
Covered in: Web of Science (WOS); PubMed.gov; IndexCopernicus; The Linguist List; Google Academic; Ulrichs; getCITED; Genamics JournalSeek; J-Gate; SHERPA/RoMEO; Dayang Journal System; Public Knowledge Project; BIUM; NewJour; ArticleReach Direct; Link+; CSB; CiteSeerX; Socolar; KVK; WorldCat; CrossRef; Ideas RePeC; Econpapers; Socionet.

2023, Volume 14, Issue 3, pages: 254-286 | https://doi.org/10.18662/brain/14.3/474
Submitted: July 7th, 2023 | *Accepted for publication:* August 19th, 2023

Development of a Simulation Environment for the Importance of Histone Deacetylase in Childhood Acute Leukemia with Explainable Artificial Intelligence

Ilhan UYSAL [1],
Utku KOSE [2],

[1] Burdur Mehmet Akif Ersoy University, ORCID ID: https://orcid.org/0000-0002-6091-9110, ilhanuysal@gmail.com, iuysal@mehmetakif.edu.tr
[2] Suleyman Demirel University, Turkey, University of North Dakota, USA, ORCID ID: https://orcid.org/0000-0002-9652-6415, utkukose@gmail.com, utkukose@sdu.edu.tr, utku.kose@und.edu

Abstract: *This study aims to explore new therapeutic opportunities for histone deacetylase (HDAC) inhibitors by leveraging drug repurposing approaches and analyzing their bioactivity and molecular fingerprints. The methodology includes investigating drug repurposing opportunities for HDAC inhibitors, evaluating the bioactivity of repurposing drugs on HDAC enzymes, investigating the role of HDAC genes in therapeutic effects, and analyzing molecular fingerprints with explainable artificial intelligence (XAI) to identify structurally similar compounds with potential HDAC inhibitory activity. In this context, chemical compounds with IC50 (7903 compounds) and Inhibition (1084 compounds) standard types of HDAC genes reported to be associated with childhood acute leukemia were represented by molecular fingerprints. Regression and classification models were applied to the molecular fingerprints, and the results obtained were supported by XAI. All the study results were shared interactively on the website address https://iuysal1905-childhoodacuteleukemia-drug-interacito-arayuz-r89zld.streamlit.app/ by designing a simulation environment. The influence of molecular fingerprints on the models and their impact on potential drug development in childhood acute leukemia were evaluated using XAI techniques, particularly through the analysis of SHAP values. The study contributes to the literature on the use of XAI technology in drug repurposing studies, especially in cancer, the study of molecular properties, and the active use of XAI in drug repurposing studies.*

Keywords: *Explainable artificial intelligence, childhood acute leukaemia, histone deacetylase, regression, classification.*

How to cite: Uysal, I., Kose, U. (2023). Development of a Simulation Environment for the Importance of Histone Deacetylase in Childhood Acute Leukemia with Explainable Artificial Intelligence. *BRAIN. Broad Research in Artificial Intelligence and Neuroscience, 14*(3), 254-286. https://doi.org/10.18662/brain/14.3/474

1. INTRODUCTION

Leukemia is a type of cancer characterized by abnormal cell proliferation. It can be caused by various factors such as genetic predispositions, hereditary diseases, and environmental influences. Among childhood cancers, leukemia accounts for 30% of cases and is the most common malignant tumor. Approximately 98% of childhood leukemia cases consist of acute leukemias, namely Acute Lymphoblastic Leukemia (ALL) and Acute Myeloid Leukemia (AML), which vary based on the type of cells involved. ALL, being the most common form, is characterized by the uncontrolled growth of undeveloped lymphoid cells in the bone marrow, peripheral blood, and various organs. This leads to a reduced capacity of the bone marrow to produce sufficient mature red blood cells, platelets, and neutrophils. Leukemia often presents with a range of typical symptoms such as tiredness, elevated body temperature, excessive perspiration during sleep, reduction in body weight, difficulty breathing, lightheadedness, heightened vulnerability to infections, bluish-purple skin discoloration (known as cyanosis), and indications of bleeding. In pediatric patients, the sole manifestation of discomfort could occasionally be limited to the extremities or joints (Akalın & Yumuşak, 2023; Bordbar et al., 2023; Carroll & Bhatla, 2016; Demir, 2023).

HDACs, also known as histone deacetylases, play a crucial role in gene regulation through the modification of histone proteins. By influencing chromatin structure and gene transcription, they participate in essential cellular functions, including cell cycle advancement, cellular differentiation, and programmed cell death (apoptosis). Abnormalities in HDAC function have been associated with various human diseases, including cancer. In the case of childhood acute leukemia, there is evidence of dysregulated HDAC activity and expression, suggesting their role in the development of leukemia. Understanding the specific mechanisms through which HDACs contribute to leukemia can provide valuable insights into potential targets for therapeutic interventions (Eyal et al., 2005; Thotala et al. 2015).

Artificial intelligence (AI) has made significant advancements in various fields, including healthcare and cancer research. However, traditional machine learning models often lack interpretability, which hinders our understanding of the underlying biological mechanisms and limits their practical use. Explainable AI (XAI) techniques have emerged as a solution to this problem, providing interpretable models that offer transparency and comprehensibility. In the field of cancer research, XAI plays a crucial role in unraveling the intricate interactions between genetic and epigenetic factors.

It facilitates the identification of new biomarkers and therapeutic targets, shedding light on the complex nature of cancer. XAI is also instrumental in drug discovery and repurposing efforts within the field of medicinal chemistry, as it enhances the interpretation and reliability of drug effects. By leveraging XAI, researchers can establish meaningful connections between biological effects and physicochemical factors, enabling the development of accurate and relevant models. The ultimate goal of XAI is to uncover the inner workings of the drug discovery process, shed light on how it is executed, and provide valuable insights related to this information (Christoph, 2020).

Drug repurposing involves investigating the potential use of existing drugs for new therapeutic purposes, specifically in the case of HDAC inhibitors, which have shown promise as anticancer agents. The goal is to determine whether drugs approved for other conditions can also exhibit HDAC inhibitory activity. Assessing the bioactivity of repurposing drugs is crucial to understand their inhibitory potential on HDAC enzymes and their ability to modulate cancer-related processes like cell proliferation, differentiation, and apoptosis. Molecular fingerprints represent unique patterns or representations of chemical compounds based on their structural features and properties. Analyzing the molecular fingerprints of repurposed drugs allows researchers to compare them to known HDAC inhibitors and evaluate their structural similarities and potential interactions with HDAC enzymes. This analysis helps identify drugs that have the potential to target HDACs and exhibit the desired bioactivity. This study focuses on histone deacetylase genes associated with childhood acute leukemia and utilizes explainable artificial intelligence (XAI) to support molecular calculations. The study selects relevant genes from the Chembl database, filters molecules according to Lipinski rules, and determines the most suitable molecules using various artificial intelligence algorithms. By comparing the properties of the molecules and employing XAI technology, the study calculates the performance of models based on molecular fingerprints. Notably, this study stands out by conducting both regression and classification analyses on the molecules, contributing to the literature on molecular properties and the active use of XAI in cancer-focused drug repurposing studies.

2. LITERATURE REVIEW

In the literature review conducted considering the scope of this study, different publications under the titles of histone deacetylase, childhood acute leukaemia, chemical structure and bioactivity and drug repurposing were examined. In the studies in the literature, HDACs are an interesting target in cancer treatment and show an altered expression in many cancers including haematological cancers, some HDAC genes are highly expressed in cancers such as childhood acute lymphoblastic leukaemia (ALL), HDAC inhibitors have emerged as promising drugs in cancer treatment, However, it has been stated that their use is limited due to their toxicity, that redesigning HDAC inhibitors has the potential to discover new treatment options, that they can be discovered by computational methods such as molecular docking and QSAR, and that redesigning existing drugs can be an effective way to discover new inhibitors. In this context, important studies that have had an impact on the establishment of the foundations and comparison processes of the thesis study can be summarised as follows:

In a study conducted by Moreno et al. (2010), the mRNA expression patterns of HDAC genes were investigated in 94 samples of childhood acute lymphoblastic leukemia (ALL). It was observed that certain HDAC genes, namely HDAC2, HDAC3, HDAC8, HDAC6, and HDAC7, exhibited higher levels of expression in ALL samples compared to normal bone marrow samples. Moreover, specific subtypes of ALL, such as T-ALL or B-cell ALL, displayed elevated expression levels of particular HDAC genes. Notably, increased expression of HDAC3 was found to be correlated with decreased overall survival in the patient group as a whole, as well as in T-ALL patients specifically. Similarly, high expression of HDAC7 and HDAC9 was also associated with lower survival rates. These findings indicate that HDAC7 and HDAC9 may serve as potential therapeutic targets and are indicative of poor prognosis in childhood ALL.

Liu et al. (2020), investigated the use of drug repurposing, pharmacophore modelling, 3D-QSAR and docking studies to identify novel HDAC inhibitors. They aimed to discover new inhibitors that could be used to treat cancer and other diseases, and through their research they identified several potential HDAC inhibitors that had not been previously studied. They also noted that drug redesign can be an effective way to discover new inhibitors, as it allows the use of existing drugs that have already been tested for safety and efficacy. It therefore indicates the potential to use a combination of computational methods and drug redesign to identify new HDAC inhibitors, which could ultimately lead to the development of new treatments for cancer and other diseases.

Gruhn et al. (2013), investigated histone deacetylase 4 (HDAC4) expression in childhood acute lymphoblastic leukaemia (ALL) and its potential association with clinical and biological features, and aimed to identify relevant HDAC isoforms for childhood ALL and determine their effects on response to treatment and prognosis. The study reported that

HDAC1, HDAC2 and HDAC8 showed significantly higher expression in ALL samples. In addition, high HDAC4 levels were associated with unfavourable prognostic factors and they suggested that HDAC4 may play a role in poor response to prednisone in childhood ALL.

According to a recent study by Pacaud et al. (2023), the clinical applications of HDAC inhibitors have been extensively discussed. The authors highlighted the development of various HDAC inhibitors with distinct structures and functions, aiming to reverse abnormal epigenetic changes observed in cancer cells. The growing body of literature provides sufficient preclinical evidence to explore these drugs in different cancer stages and contexts, either as standalone therapies or in combination with other agents, to effectively target haematological and solid tumor malignancies. Vorinostat (Zolinza), the first approved HDAC inhibitor, has demonstrated success in treating patients with cutaneous T-cell lymphoma and epilepsy. Moreover, HDAC inhibitors have shown promising potential in the treatment of non-cancerous conditions such as cystic fibrosis, spinal muscular atrophy, and human immunodeficiency virus infection.

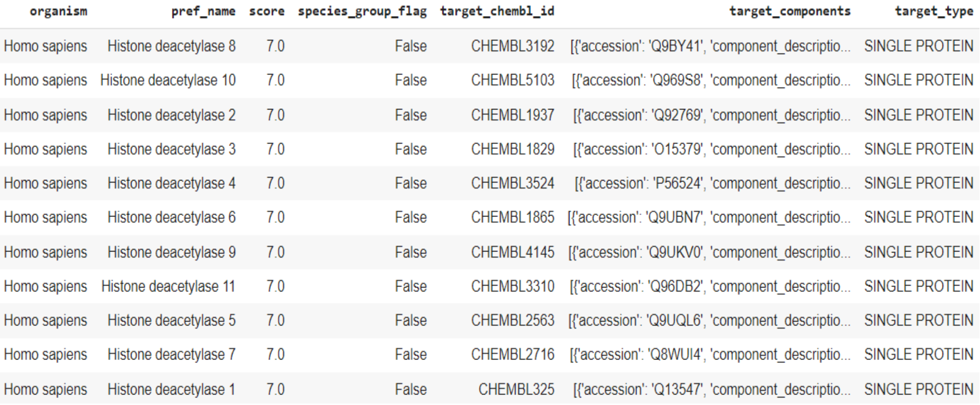
In a study conducted by Cortés-Ciriano et al. (2020), the effectiveness of QSAR-derived affinity fingerprints (QAFFP) in predicting potency was investigated. They proposed a method for calculating QAFFP that allows compounds to be encoded and compared based on their similarity in bioactivity. The researchers compared the predictive ability of QAFFP using IC50 data from the ChEMBL database for different cancer cell lines and protein target datasets. The results showed that QAFFP was able to generate highly accurate models, with RMSE values ranging from approximately 0.6 to 0.9 pIC50 units. These values were similar to the uncertainty observed in the heterogeneous IC50 data in ChEMBL. Moreover, QAFFP performed comparably to Morgan2 fingerprints and physicochemical descriptors in terms of predictive power. These findings demonstrate that QAFFP is highly effective in tasks such as similarity search, composite classification, and scaffold jumping.

Kirboga et al. (2022) focused on the application of artificial intelligence techniques to investigate potential therapeutic options for hereditary Friedreich's Ataxia (FA). The study specifically investigated iron chelation molecules and HDAC inhibitors as potential treatments for FA. A quantitative structure-activity relationship (QSAR) analysis was performed using compounds from the Chembl database. The bioactivity of 436 compounds for Fe chelation and 1,163 compounds for HDAC inhibition was measured using IC50 units. Random Forest technique was used for model building. The models built using PubChem fingerprinting outperformed the others, demonstrating its suitability for interpretation. The study highlighted the importance of nitrogen-containing functional groups and aromatic rings in the compounds analysed by XAI.

3. METHODOLOGY

Various studies have demonstrated the frequent occurrence of HDAC mutations and abnormal gene expressions in different cancer types and haematological malignancies. Specifically, in bone marrow samples of childhood acute lymphoblastic leukemia (ALL), elevated expressions of HDAC2, HDAC3, HDAC4, HDAC6, HDAC7, and HDAC8 genes have been observed compared to healthy children's bone marrow samples (Klimek et al., 2008; Cress & Seto, 2000; Bali et al., 2018).

In this particular study, genes ranging from HDAC1 to HDAC11 were included, and the findings retrieved from the Chembl database are presented in Figure 3.1. This data provides insights into the expression levels of HDAC genes relevant to childhood ALL, highlighting their potential involvement in the disease.



organism	pref_name	score	species_group_flag	target_chembl_id	target_components	target_type
Homo sapiens	Histone deacetylase 8	7.0	False	CHEMBL3192	[{'accession': 'Q9BY41', 'component_descriptio...	SINGLE PROTEIN
Homo sapiens	Histone deacetylase 10	7.0	False	CHEMBL5103	[{'accession': 'Q969S8', 'component_descriptio...	SINGLE PROTEIN
Homo sapiens	Histone deacetylase 2	7.0	False	CHEMBL1937	[{'accession': 'Q92769', 'component_descriptio...	SINGLE PROTEIN
Homo sapiens	Histone deacetylase 3	7.0	False	CHEMBL1829	[{'accession': 'O15379', 'component_descriptio...	SINGLE PROTEIN
Homo sapiens	Histone deacetylase 4	7.0	False	CHEMBL3524	[{'accession': 'P56524', 'component_descriptio...	SINGLE PROTEIN
Homo sapiens	Histone deacetylase 6	7.0	False	CHEMBL1865	[{'accession': 'Q9UBN7', 'component_descriptio...	SINGLE PROTEIN
Homo sapiens	Histone deacetylase 9	7.0	False	CHEMBL4145	[{'accession': 'Q9UKV0', 'component_descriptio...	SINGLE PROTEIN
Homo sapiens	Histone deacetylase 11	7.0	False	CHEMBL3310	[{'accession': 'Q96DB2', 'component_descriptio...	SINGLE PROTEIN
Homo sapiens	Histone deacetylase 5	7.0	False	CHEMBL2563	[{'accession': 'Q9UQL6', 'component_descriptio...	SINGLE PROTEIN
Homo sapiens	Histone deacetylase 7	7.0	False	CHEMBL2716	[{'accession': 'Q8WUI4', 'component_descriptio...	SINGLE PROTEIN
Homo sapiens	Histone deacetylase 1	7.0	False	CHEMBL325	[{'accession': 'Q13547', 'component_descriptio...	SINGLE PROTEIN

Figure 3.1. HDAC genes
Author's own conception

The dataset utilized in this study was sourced from the ChEMBL version 29 database, which specifically caters to HDAC-related information. To enhance the quality of the data, compounds exhibiting IC50 and Inhibition values were selected as the bioactivity units of interest. Consequently, a dataset comprising 7,903 compounds for IC50 and 1,084 compounds for Inhibition was compiled. In order to construct a classification model for HDAC, specific thresholds of <1 µM and >10 µM were established to distinguish between active and inactive compounds, aligning with the intended target. To uniquely represent the compounds, molecular fingerprint identifiers were generated using SMILES indicators, which encode the structural information of the molecules.

3.1. Lipinski Rules

In drug discovery, three commonly used rules are employed to assess the characteristics and similarities of potential drug candidates: Lipinski's rule, Veber's rule, and Ghose's rule. Among these, Lipinski's rule is widely recognized and preferred. It takes into account key physicochemical properties, such as lipophilicity and water solubility, to evaluate the potential similarity of a compound to a drug. Lipinski's rule serves as a valuable guideline for assessing drug candidates by considering factors such as cellular permeability and pharmacokinetic properties.

Lipinski's rule of five outlines specific criteria that a molecule must meet to be considered a potential drug candidate (Lipinski, 2004):

- Molecular weight ≤ 500 g/mol
- Lipophilicity coefficient (LogP) ≤ 5
- No more than five hydrogen bond donors
- No more than ten hydrogen bond acceptors
- Molar refractive values between 4 and 130.

The molecular weight of a compound is important in determining its permeability, as lower molecular weight compounds tend to exhibit higher oral activity. Lipophilicity, as measured by the logP value, is closely associated with drug absorption. An increased number of hydrogen bond donor groups in a compound can impede its penetration through cell membranes. Similarly, a higher number of hydrogen bond acceptors also impacts permeability. These properties play a significant role in drug discovery and the bioavailability of potential drug candidates. Adhering to these rules aids in evaluating the properties of potential drug candidates and optimizing their pharmacokinetic characteristics to enhance their biological activity.

3.2. Explainable Artificial Intelligence

Explainable AI (XAI) plays a vital role in accurately interpreting artificial intelligence models. Its purpose is to enhance transparency in the decision-making processes and ensure reliable predictions by providing explanations for the models' outcomes (Murdoch et al. 2019; Doshi-Velez & Kim, 2017; Lapuschkin et al., 2019; Miller, 2019).

Shapley values are utilized to determine the relative importance of features in predictive models, comprehend feature interactions, and explain model predictions. These values serve as a valuable tool in identifying which attributes are associated with changes in model outputs. Furthermore, Shapley values can be applied in various contexts, including assessing model fairness and detecting attribute-based unfairness or discrimination (Lundberg & Lee, 2017; Lundberg et al. 2020; Kumar et al., 2020).

3.3. Chembl Database

Chembl is a comprehensive database that houses a wide range of chemical and biological information, including bioactivity data, chemical structures, drug targets, drug discovery projects, and pharmacological profiles of drug molecules. It encompasses data on compound activities in bioassays, their interactions with target molecules, and quantitative assessments of these interactions (Gaulton et al., 2017). Notably, the data in Chembl is sourced from scientific literature, which ensures its reliability and can be cross-referenced for verification. The database also incorporates drug efficacy data that is linked to published articles, further bolstering its credibility. Given these advantageous features, the Chembl database was selected as the data source for this study.

3.4. Artificial Intelligence Models and Python Libraries

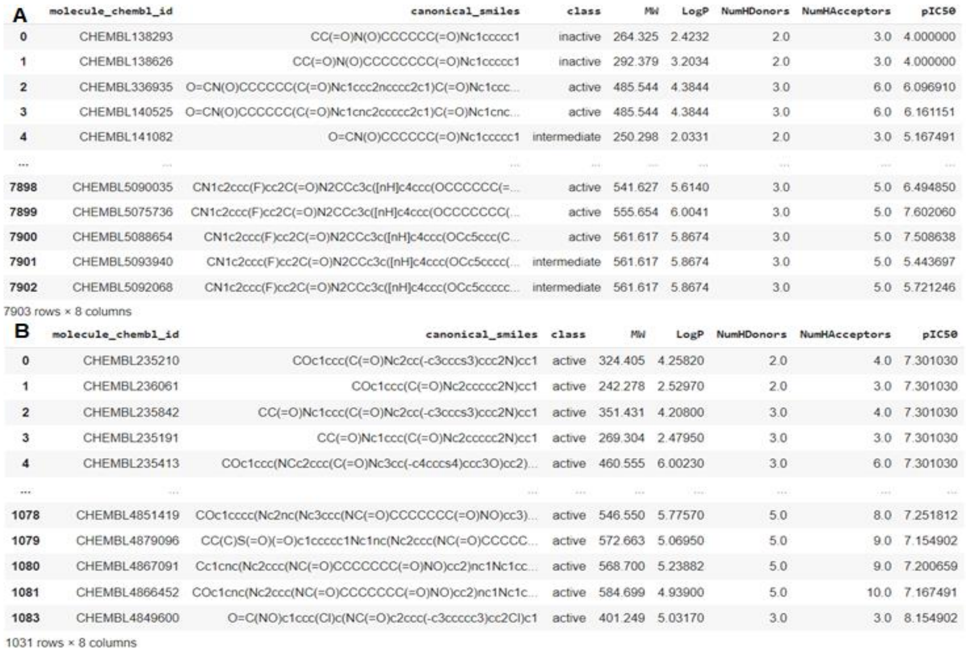
Regression is a supervised learning technique employed in machine learning to predict continuous numerical values. It involves establishing a relationship between input variables (features) and the corresponding output variable (target) by fitting a mathematical function to the data. The objective is to create a model that can accurately forecast the continuous value of the target variable based on the provided input characteristics. There are several regression algorithms commonly used in data analysis, including linear regression, polynomial regression, support vector regression (SVR), decision tree regression, and random forest regression. To evaluate the effectiveness of these regression models, various metrics are employed, such as root mean squared error (RMSE), R-square (coefficient of determination), and adjusted R-square. These metrics provide insights into how well the models fit the data and predict the target variable (Bishop & Nasrabadi, 2006; Géron, 2022; Hastie et al., 2009).

Classification, on the other hand, is a supervised learning technique used to predict the categorical class or label of an input data point. It involves mapping input features to predefined output classes. The goal is to construct a model that can accurately assign new data points to their respective classes based on patterns and relationships learned from training data. Classification models often utilize various algorithms such as logistic regression, support vector machines (SVM), decision trees, random forests, and neural networks (including deep learning models). When evaluating the performance of these models, metrics such as accuracy, F1-score, and the area under the receiver operating characteristic curve (ROC-AUC) are commonly employed (Murphy, 2012; Bishop & Nasrabadi, 2006; Géron, 2022; Hastie, 2009).

The study employed several Python libraries, including pandas for data analysis, numpy for calculations, rdkit for molecular modeling, matplotlib and seaborn for visualization, eli5 for interpreting models, lazy predict for assessing the performance metrics of models collectively, and stream lit packages for developing an interactive web application.

4. RESULTS

Figure 4.1 presents the calculated activity states, SMILES representations, pIC50 values, and Lipinski properties (molecular weight, lipophilicity, number of hydrogen bond donors, and number of hydrogen bond acceptors) of the molecules under investigation. The activity states indicate whether the molecules are active, inactive, or have an intermediate value. The pIC50 values represent the negative logarithmic scale of the IC50 data, providing a more uniform distribution for the IC50 values. Additionally, the Lipinski properties provide important information about the molecular characteristics of the compounds. The molecular weight (MW) reflects the size of the molecule, while lipophilicity (LogP) indicates the compound's affinity for lipid or water environments. The number of hydrogen bond donors (NumHDonors) and hydrogen bond acceptors (NumHAcceptors) provide insights into the compound's potential for forming hydrogen bonds. By examining these properties, researchers can gain a better understanding of the molecular features and potential drug-like properties of the compounds.



A	molecule_chembl_id	canonical_smiles	class	MW	LogP	NumHDonors	NumHAcceptors	pIC50
0	CHEMBL138293	CC(=O)N(O)CCCCCC(=O)Nc1ccccc1	inactive	264.325	2.4232	2.0	3.0	4.000000
1	CHEMBL138626	CC(=O)N(O)CCCCCCCC(=O)Nc1ccccc1	inactive	292.379	3.2034	2.0	3.0	4.000000
2	CHEMBL336935	O=CN(O)CCCCCC(C(=O)Nc1ccc2ncccc2c1)C(=O)Nc1ccc...	active	485.544	4.3844	3.0	6.0	6.096910
3	CHEMBL140525	O=CN(O)CCCCCC(C(=O)Nc1cnc2ccccc2c1)C(=O)Nc1cnc...	active	485.544	4.3844	3.0	6.0	6.161151
4	CHEMBL141082	O=CN(O)CCCCCC(=O)Nc1ccccc1	intermediate	250.298	2.0331	2.0	3.0	5.167491
...	...	...	...	...	...	...	...	...
7898	CHEMBL5090035	CN1c2ccc(F)cc2C(=O)N2CCc3c([nH]c4ccc(OCCCCCC(=...	active	541.627	5.6140	3.0	5.0	6.494850
7899	CHEMBL5075736	CN1c2ccc(F)cc2C(=O)N2CCc3c([nH]c4ccc(OCCCCCCC(...	active	555.654	6.0041	3.0	5.0	7.602060
7900	CHEMBL5088654	CN1c2ccc(F)cc2C(=O)N2CCc3c([nH]c4ccc(OCc5ccc(C...	active	561.617	5.8674	3.0	5.0	7.508638
7901	CHEMBL5093940	CN1c2ccc(F)cc2C(=O)N2CCc3c([nH]c4ccc(OCc5cccc(...	intermediate	561.617	5.8674	3.0	5.0	5.443697
7902	CHEMBL5092068	CN1c2ccc(F)cc2C(=O)N2CCc3c([nH]c4ccc(OCc5ccccc...	intermediate	561.617	5.8674	3.0	5.0	5.721246

7903 rows × 8 columns

B	molecule_chembl_id	canonical_smiles	class	MW	LogP	NumHDonors	NumHAcceptors	pIC50
0	CHEMBL235210	COc1ccc(C(=O)Nc2cc(-c3cccs3)ccc2N)cc1	active	324.405	4.25820	2.0	4.0	7.301030
1	CHEMBL236061	COc1ccc(C(=O)Nc2ccccc2N)cc1	active	242.278	2.52970	2.0	3.0	7.301030
2	CHEMBL235842	CC(=O)Nc1ccc(C(=O)Nc2cc(-c3cccs3)ccc2N)cc1	active	351.431	4.20800	3.0	4.0	7.301030
3	CHEMBL235191	CC(=O)Nc1ccc(C(=O)Nc2ccccc2N)cc1	active	269.304	2.47950	3.0	3.0	7.301030
4	CHEMBL235413	COc1ccc(NCc2ccc(C(=O)Nc3cc(-c4cccs4)ccc3O)cc2)...	active	460.555	6.00230	3.0	6.0	7.301030
...	...	...	...	...	...	...	...	...
1078	CHEMBL4851419	COc1cccc(Nc2nc(Nc3ccc(NC(=O)CCCCCCC(=O)NO)cc3)...	active	546.550	5.77570	5.0	8.0	7.251812
1079	CHEMBL4879096	CC(C)S(=O)(=O)c1ccccc1Nc1nc(Nc2ccc(NC(=O)CCCCC...	active	572.663	5.06950	5.0	9.0	7.154902
1080	CHEMBL4867091	Cc1cnc(Nc2ccc(NC(=O)CCCCCCC(=O)NO)cc2)nc1Nc1cc...	active	568.700	5.23882	5.0	9.0	7.200659
1081	CHEMBL4866452	COc1cnc(Nc2ccc(NC(=O)CCCCCCC(=O)NO)cc2)nc1Nc1c...	active	584.699	4.93900	5.0	10.0	7.167491
1083	CHEMBL4849600	O=C(NO)c1ccc(Cl)c(NC(=O)c2ccc(-c3ccccc3)cc2Cl)c1	active	401.249	5.03170	3.0	3.0	8.154902

1031 rows × 8 columns

Figure 4.1. SMILES, pIC50 and Lipinksi Values of Molecules. A) IC50, B) Inhibition
Author's own conception

Figure 4.2 presents the calculated base values of the molecules based on Lipinski's rules. Lipinski's rules are widely used in drug discovery to assess the drug-likeness and potential for oral bioavailability of compounds. These rules consider several physicochemical properties of the molecules, including molecular weight, lipophilicity (LogP), number of hydrogen bond donors, and number of hydrogen bond acceptors. Comparing the base values of the molecules according to Lipinski's rules allows researchers to evaluate their compliance with the criteria for drug-likeness. For example, if the molecular weight is below 500 g/mol, LogP is less than or equal to 5, and the number of hydrogen bond donors and acceptors is within the specified limits, the molecule is considered more likely to have favourable pharmacokinetic properties. Analyzing the structure and properties of the compounds using chemo informatics methods, such as those provided by the Rdkit library in Python, enables researchers to gain insights into their drug-like characteristics and potential for effective and safe use. Additionally, considering the ADME processes, which include absorption, distribution, metabolism, and elimination, helps evaluate the bioavailability of the drug candidate. Molecular flexibility and hydrogen bond count are important factors in assessing the ADME properties and overall drug efficacy.

(Landrum, 2016; Lipinski et al., 2012; Veber et al., 2002). By examining the base values of the molecules and assessing their compliance with Lipinski's rules, researchers can make informed decisions about the potential drug-likeness and bioavailability of the compounds.

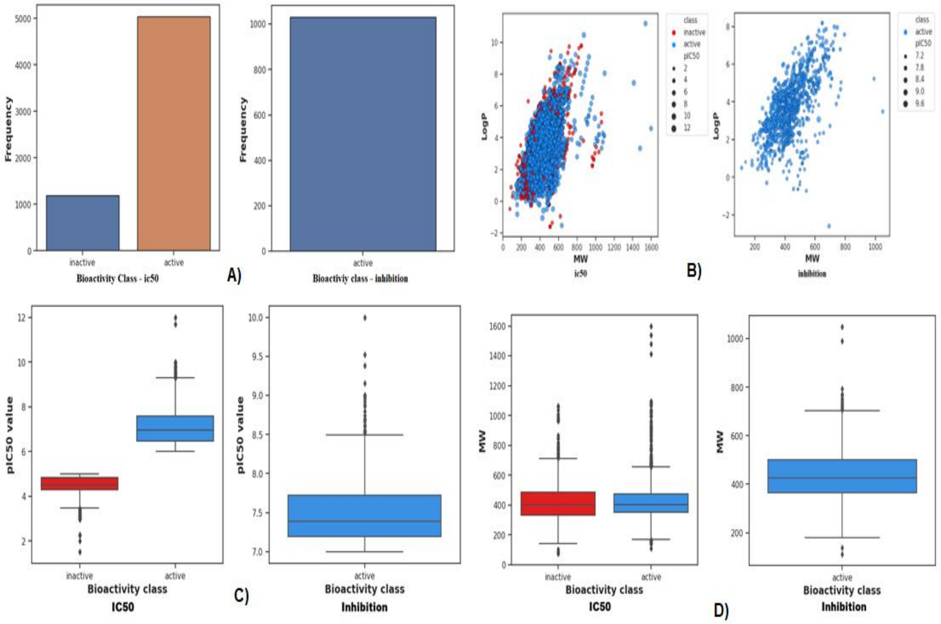


Figure 4.2. Molecule analysis A) Activity status, B) LogP and Molecule Weight (MW), C) PIC50 and Bioactivity, D) MW and Bioactivity
Author's own conception

Figure 4.3 presents the model performance metrics obtained using the Lazypredict package for the binary matrix created based on the molecular fingerprints of the molecules. The metrics used to evaluate the models include R2 (coefficient of determination), Adjusted R2, and Root Mean Squared Error (RMSE). These metrics provide insights into the accuracy and reliability of the models in predicting the target variable. According to the results in Figure 4.3, the decision tree regressor model has demonstrated the highest performance among the evaluated models for both standard types (IC50 and Inhibition). The decision tree algorithm is a non-parametric supervised learning technique that can capture complex interactions between features and provide interpretable results. Its ability to handle both numerical and categorical data, as well as its capacity to capture non-linear relationships, makes it well-suited for modeling the relationship between molecular fingerprints and drug-like properties. The choice of the decision tree regressor model as the most successful model suggests that it is capable of accurately predicting the target variable based on the molecular fingerprints and capturing the underlying patterns and relationships within the data (Hochreiter et al., 2018; David et al., 2020).

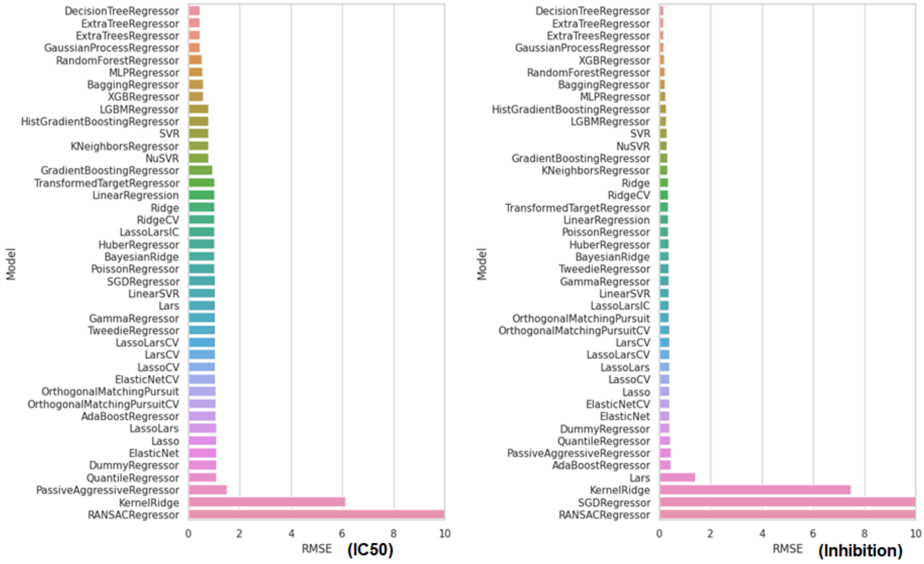


Figure 4.3. Comparison of RMSE values of the models
Author’s own conception

Figure 4.4 provides a comparison of the R-squared (R2) values obtained from the different models for both standard types (IC50 and Inhibition). R-squared is a statistical metric that quantifies the amount of the target variable's variation that can be explained by the independent variables. It is commonly used to assess the goodness of fit of regression models, indicating how well the model captures the variability in the data. According to the results in Figure 4.4, the decision tree regressor model has consistently achieved the highest R-squared values among all the evaluated models for both standard types. This indicates that the high R-squared values obtained by the decision tree regressor model suggest that it can effectively capture the underlying relationships between the molecular fingerprints and the drug-like properties.

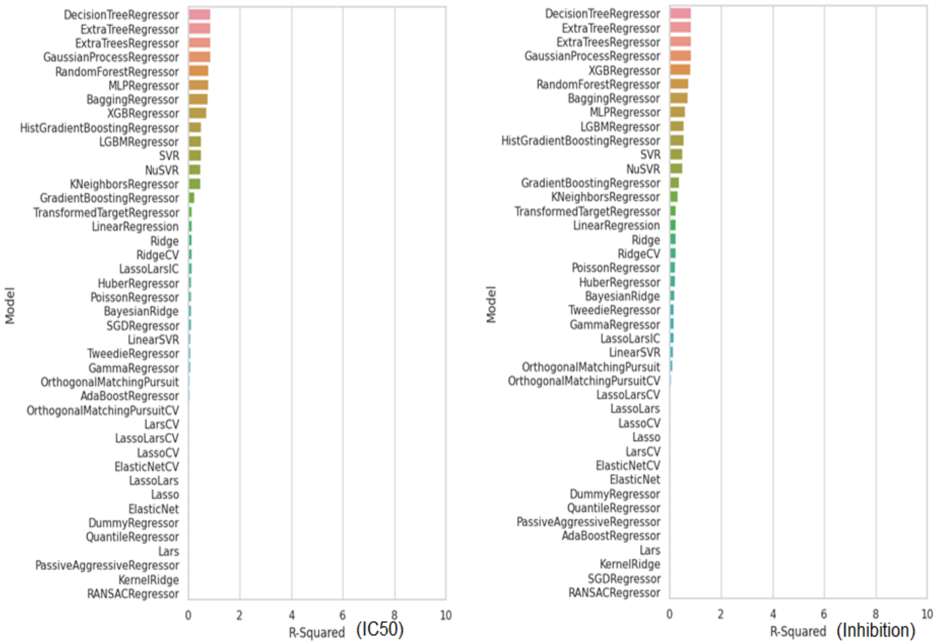


Figure 4.4. Comparison of R-Squared values of the models
Author's own conception

The comparison of the Adjusted R-Squared values of the models is given in Figure 4.5. According to this, the most successful model in both standard types has been the decision tree regressor.

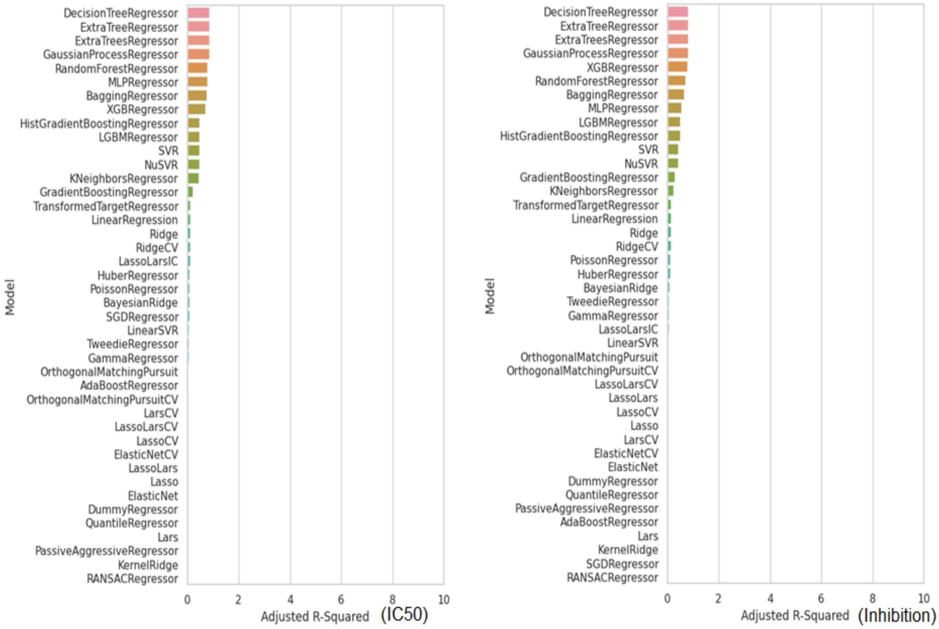


Figure 4.5. Comparison of Adjusted R-Squared values of the models
Author's own conception

Comparison of Absolute Error (AE), Relative Error (RE), Squared Error (SE) and Root Mean Squared Error (RMSE) metrics of the models was performed with Rapidminer software and given in Figure 4.6. Accordingly, the most successful model was the Gradient Boosted Trees model.

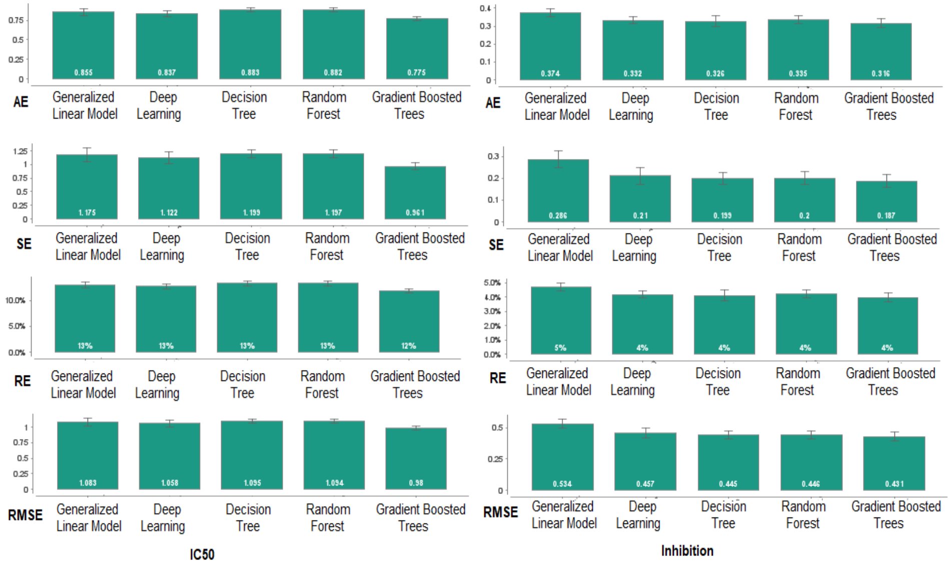


Figure 4.6. Comparison of AE, SE, RE, RMSE values of the models with Rapidminer (IC50 on the left, Inhibition on the right)

A visualisation of the time taken to calculate the performance metrics of the models is given in Figure 4.7. It is noteworthy that the Quantile Regressor model is the most time-consuming model when calculating performance metrics for both IC50 and Inhibition. It is important that Decision Tree Regressor and Extra Tree Regressor, which are the most successful models in performance metrics, are also among the most successful models in time taken.

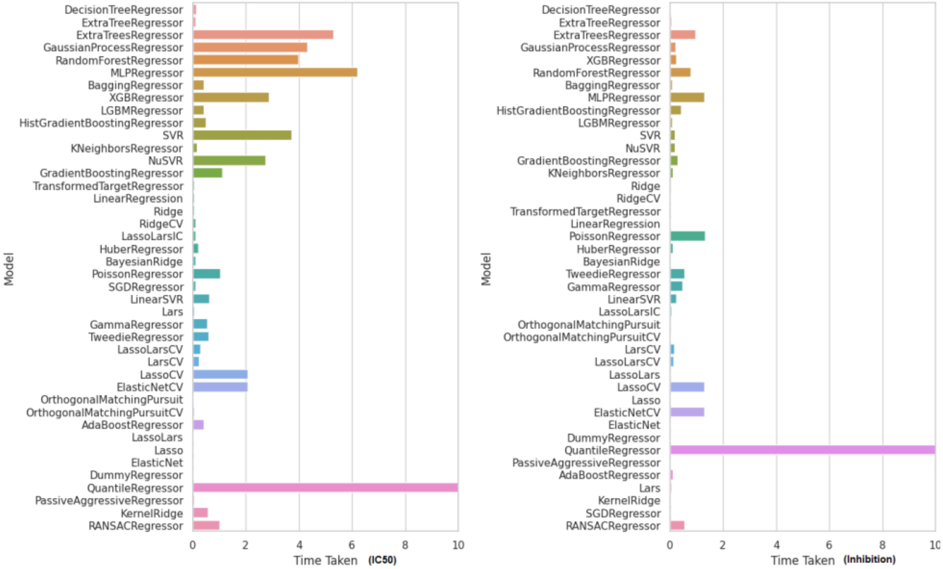


Figure 4.7. Time Taken of models
Author's own conception

By selecting the Decision Tree Regressor model, a training set regression visualisation was created as shown in Figure 4.8.

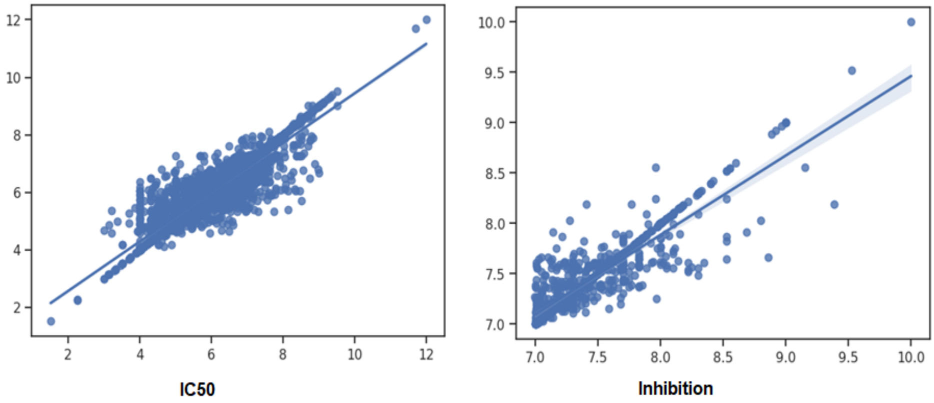


Figure 4.8. Regression graph of the Decision Tree Regressor model
Author's own conception

The values in the Smiles column were converted into molecule objects using the rdkit.Chem and rdkit.Chem.Draw modules and the Chem.MolFromSmiles function. The first 20 molecule structures were visualised with the MolsToGridImage function and given in Appendix 1. Then, the calculated morgan fingerprints values of the molecules were combined with the values labelled as 1 and 0 according to the activity status in the label column and the resulting data set was divided into training and test sets. A model was created on the training set with Random Forest Classifier and the predictions of the model were calculated with the test data. The performance of the classification model was calculated with the ROC curve and a high success was obtained with a value of 0.9725.

The compounds in the data set were labelled as "active" and "inactive". Then, according to these labels, the compounds were classified according to their label values. Chemical structures in Smiles format were represented using Morgan Fingerprints, known as chemical fingerprints. A chemical fingerprint is a numerical vector used to represent the properties of chemical compounds. In this way, they are converted into a form that can be processed by machine learning algorithms. In order to evaluate the true and predicted classes of the model, the Confusion Matrix class from the Pycm library was used to calculate the complexity matrix between the true labels and the predicted labels and visualised with a heatmap in Figure 4.9. Accordingly, the number of instances where the true positive class was correctly predicted as positive was 3196, the number of instances where the true negative class was incorrectly predicted as positive was 32, the number of instances where the true positive class was incorrectly predicted as negative was 14 and the number of instances where the true negative class was correctly predicted as negative was 600.

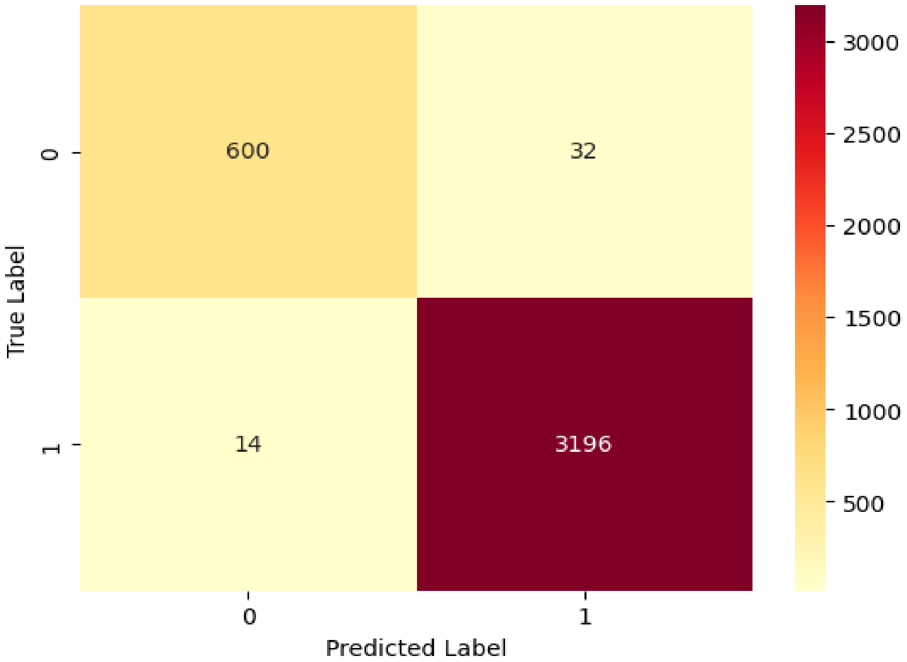


Figure 4.9. Confusion Matrix
Author's own conception

In order to see the performance of the model more clearly as a result of the classification, some metrics were evaluated. The values of these metrics are given in Appendix-2 and Appendix-3. According to the performance metrics, high performance is obtained for both classes (0 and 1). The ACC (Accuracy) value is approximately 98.8% for both class 0 and class 1, i.e., the proportion of correctly classified samples is quite high. The AUC (Area under the ROC curve) and AUPR (Area under the PR curve) values were calculated as approximately 0.97 for both classes. This shows that the classification ability of the model is high. The F1 score, which is the harmonic mean of precision and recall measurements, has high values for both classes. The F1 score was calculated as 0.96 for class 0 and 0.99 for class 1. The false positive rate (FPR) and false negative rate (FNR) are also low for both classes. This shows that the model minimises both false positive classifications and false negative classifications. In general, it is seen that the model has high accuracy, precision and specificity values as a result of the given metrics. Comparison of Accuracy, F1 Score and ROC-AUC values of models is given in Figure 4.10. Accordingly, the most successful models were Extra Trees Classifier, Extra Tree Classifier, Random Forest Classifier and Decision Tree Classifier.

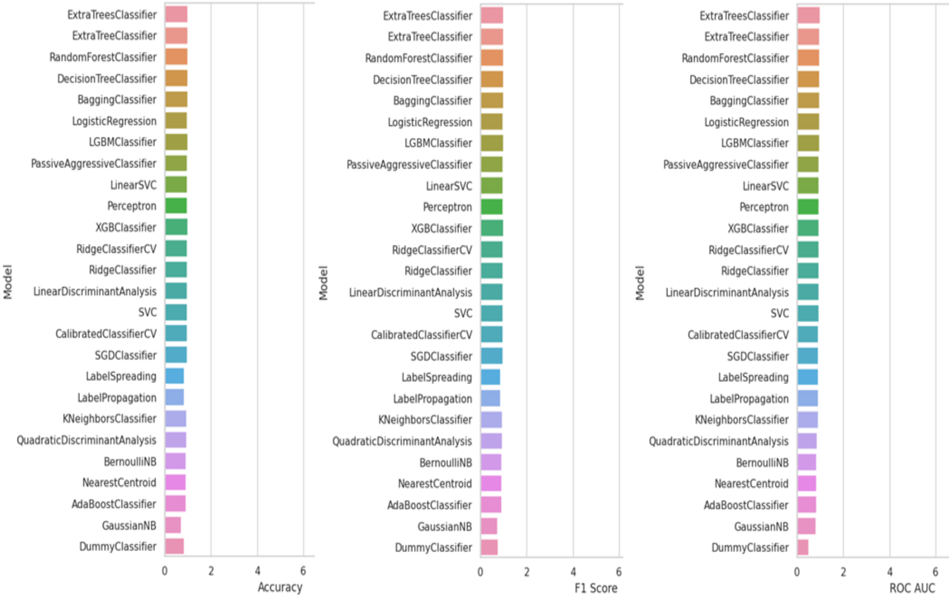


Figure 4.10. Comparison of Accuracy, F1 Score and ROC-AUC values of models
Author's own conception

5. CONCLUSIONS

In this study, virtual screening of histone deacetylase genes for childhood acute leukemia, which molecular fingerprints these genes have and the contribution of these molecular fingerprints to the discovery of histone deacetylase genes IC50 and Inhibition standard types have been determined. After determining the current status of the molecular fingerprints, the explainable artificial intelligence method was applied to determine the importance of the molecular features. After the virtual screening of the molecular properties of the candidate molecules, the most appropriate model over the binary matrix containing the information of the molecular fingerprints has been determined using Decision Tree Regressor. In order to apply an explainable artificial intelligence method on this model, it is necessary to calculate the effect of each molecular fingerprint on the models. For this reason, Shap values of each feature have been calculated and its effect on the models has been observed. In order to see the effects of molecular features on the models, different types of graphics have been used.

Waterfall plots show step-by-step changes in an initial value (usually a baseline or mean value). At each step, a feature contribution is added or subtracted from the total value. In this way, the total contribution of the features is determined. The visualisation designed to determine the effect of each molecular fingerprint on the predictions is given in Figure 5.1.

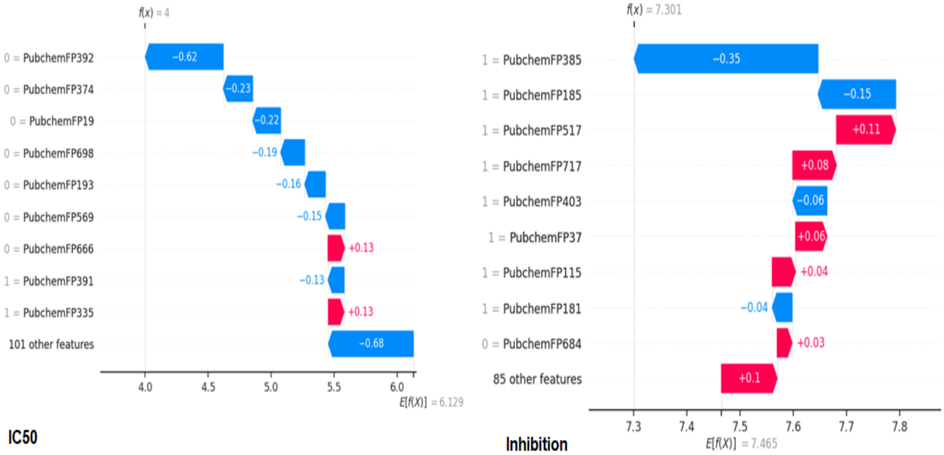


Figure 5.1. SHAP Waterfall graphics
Author’s own conception

The lower section of the waterfall plot begins with the predicted output value of the model, denoted as E[f(X)]. Each row illustrates the directional shift, either positive (indicated by red) or negative (indicated by blue), caused by the contribution of each feature from the expected output value to the model's final output value. According to the results, the expected model output value for IC50 was 6.129 while the current model output value was 4. The expected model output value for inhibition was 7.465 while the current model output value was 7.301. The grey numbers in front of the feature names represent the value of each feature in this example. Cases where the features have a value of 1 are expressed as present and cases where they have a value of 0 are expressed as absent.

In the analysis of molecular properties, certain substructures called PubchemFP335 and PubchemFP666 were found to have a positive effect on the model's output. Specifically, PubchemFP335 represents a substructure indicating the presence of three carbon atoms and one hydrogen atom around a carbon atom. This substructure suggests that the atoms surrounding aromatic bonds have less significance, and their effect ratio on the model's output is +0.13. On the other hand, PubChemFP392, representing a different substructure, contributed -0.62 to the model's prediction, indicating that it is not suitable as a drug candidate for childhood acute leukemia. Other molecules in the dataset also had varying contributions, with some positively affecting the model's prediction and others negatively affecting it. For example, the presence of oxygen atoms, as represented by PubChemFP19, had a negative impact, while a specific carbon-oxygen bond represented by PubChemFP666 had a positive effect on the model's prediction. In the analysis of molecular properties, several substructures represented by PubchemFP385, PubChemFP185, PubChemFP517, PubChemFP717, PubChemFP403, PubChemFP37, PubChemFP115, PubChemFP181, and PubChemFP684 were evaluated for their contribution to the model's output. PubchemFP385, representing a structure with three bonds of one carbon atom, had a negative contribution (-0.35) to the model, suggesting that its presence may reduce drug potential or have a negative effect on the target molecule. Similarly, PubChemFP185, representing a structure with at least two rings of size 6, also had a negative contribution (-0.15) to the model. On the other hand, PubChemFP517, representing a carbon atom bonded with nitrogen atoms, and PubChemFP717, representing a benzene ring with a chlorine atom, had positive contributions (+0.11 and +0.08, respectively), indicating that their presence may increase the drug potential or have a favorable effect on the target molecule. Other substructures, such as PubChemFP403, PubChemFP37, PubChemFP115, PubChemFP181, and PubChemFP684, also had varying contributions to the model, suggesting their potential impact on the drug potential or effect on the target molecule.

In conclusion, certain substructures, such as PubChemFP335 and PubChemFP666, were found to have a positive impact on the model's output. These substructures suggest the presence of specific molecular patterns that contribute positively to the drug potential for childhood acute leukemia (ALL). The positive contributions of these molecular fingerprints indicate that they have properties that are desirable for drug candidates targeting ALL.

Bar graph representation, which is another type of graph, is given in Figure 5.2. Molecules that contribute to the model output in a slightly positive direction are shown in the graph. The bar graph also supports the results of the waterfall graph. Other graph types and all the results obtained were presented interactively to the users in the streamlit web application by developing a simulation environment. All results obtained from the study can be accessed at https://iuysal1905-childhoodacuteleukemia-drug-interacito-arayuz-r89zld.streamlit.app/.

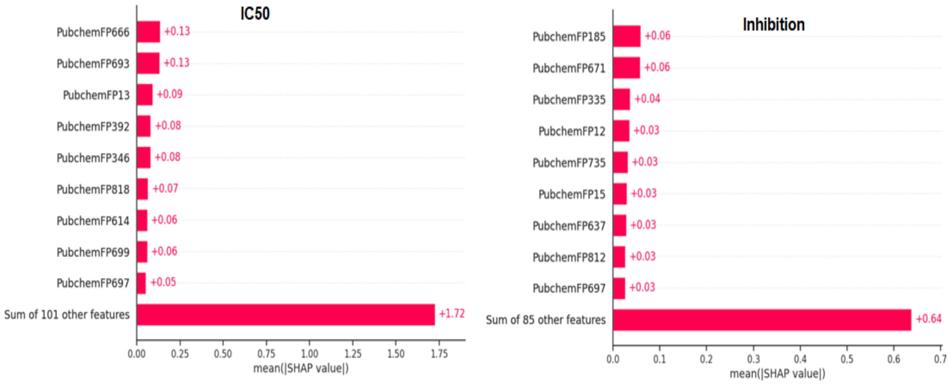


Figure 5.2. SHAP Bar graphics
Author's own conception

The main aim of this research is to utilize artificial intelligence (AI) models that promote transparency and interpretability to broaden the range of therapeutic compounds available for treating childhood acute leukemia. Thus, it will facilitate the drug development and discovery process and ensure confidence in the results obtained. Incorporating explainable algorithms like SHAP is essential for enhancing trust and transparency in the use of machine learning models, enabling data-driven approaches in the development of drugs and chemicals. The study examines the significance of molecular fingerprints identified through the SHAP annotator as predictors for IC50 and Inhibition, which represent standard types of HDAC inhibitors, within the prediction model.

In future investigations, it is advisable to examine SHAP plots to gain insights into long-term predictions and explore the SHAP properties of therapeutic trials and other contributing molecules that may impact HDAC inhibitors relevant to childhood acute leukemia.

Explainable AI (XAI) offers a range of benefits in the field of drug discovery, including improved cost-effectiveness and time efficiency. However, it also presents certain challenges such as reduced model accuracy, conflicting molecular fingerprints, and the need for accurate threshold selection. It is expected that the integration of XAI and drug discovery studies will mutually reinforce each other through the dissemination of computational research, enhanced applicability through in vitro and in vivo experiments, and hold significant promise for future investigations into drug development.

References

Akalın, F., Yumuşak, N. 2023. Classification of gene anomalies of ALL, AML and MLL leukaemia types in microarray dataset with LSTM neural network. *Journal of Gazi University Faculty of Engineering and Architecture, 38*(3), 1299-1306.

Bali, D. A., Kurekci, A. E., Akar, M. N. 2018. The Relationship of HDAC2, HDAC4, HDAC5, HDAC7, HDAC8, HDAC9 Gene expression levels with prognosis in childhood acute leukaemias. *Journal of SDU Faculty of Medicine, 25*(4), 400-406.

Bishop, C. M., and Nasrabadi, N. M. 2006. Pattern recognition and machine learning. 4(4), 738. New York: Springer.

Bordbar, M., Jam, N., Karimi, M., Shahriari, M., Zareifar, S., Zekavat, O. R., Mottaghipisheh, H. 2023. The survival of childhood leukemia: An 8 year single center experience. *Cancer Reports, 6*(4), e1784.

Carroll, W.L., Bhatla, T. 2016. Acute lymphoblastic leukemia. *Lanzkowsky's manual of pediatric hematology and oncology (sixth edition*). San Diego: Academic Press.

Christoph, M. 2020. Interpretable machine learning. *A Guide for Making Black Box Models Explainable.*

Cortés-Ciriano, I., Škuta, C., Bender, A., Svozil, D. 2020. QSAR-derived affinity fingerprints (part 2): modeling performance for potency prediction. *Journal of Cheminformatics, 12*(1), 41.

Cress, W. D., Seto, E. 2000. Histone deacetylases, transcriptional control, and cancer. *Journal of cellular physiology, 184*(1), 1-16.

David, L., Thakkar, A., Mercado, R., Engkvist, O. 2020. Molecular representations in AI-driven drug discovery: a review and practical guide. *Journal of Cheminformatics, 12*(1), 1-22.

Demir, D. 2023. Childhood leukaemias and nursing care. *Health & Science 2023: Evidence-based practices in paediatric nursing.* Efe Akademi Publications, 95.

Doshi-Velez, F., Kim, B. 2017. Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608.

Eyal, S., Yagen, B., Shimshoni, J. Bialer, M., 2005. Histone Deacetylases Inhibition and Tumor Cells Cytotoxicity by CNS-Active VPA Constitutional Isomers and Derivatives. *Biochemical Pharmacology, 69*(10), 1501–1508.

Gaulton, A., Hersey, A., Nowotka, M., Bento, A. P., Chambers, J., Mendez, D., Leach, A. R. 2017. The ChEMBL database in 2017. *Nucleic acids research, 45*(D1), D945-D954.

Géron, A. 2022. Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow. O'Reilly Media, Inc.

Gruhn, B., Naumann, T., Gruner, D., Walther, M., Wittig, S., Becker, S., Sonnemann, J. 2013. The expression of histone deacetylase 4 is associated with prednisone poor-response in childhood acute lymphoblastic leukemia. *Leukemia research, 37*(10), 1200-1207

Hastie, T., Tibshirani, R., Friedman, J. H., Friedman, J. H. 2009. The elements of statistical learning: data mining, inference, and prediction. 2, 1-758. New York: Springer.

Hochreiter, S., Klambauer, G., Rarey, M. 2018. Machine learning in drug discovery. *Journal of Chemical Information and Modeling, 58*(9), 1723-1724.

Kirboga, K. K., Kucuksille, E. U., Kose, U. 2022. Ignition of Small Molecule Inhibitors in Friedreich's Ataxia with Explainable Artificial Intelligence. 10.21203/rs.3.rs-1408745/v1

Klimek, V. M., Fircanis, S., Maslak, P., Guernah, I., Baum, M., Wu, N., Nimer, S. D. 2008. Tolerability, pharmacodynamics, and pharmacokinetics studies of depsipeptide (romidepsin) in patients with acute myelogenous leukemia or advanced myelodysplastic syndromes. *Clinical Cancer Research, 14*(3), 826-832.

Kumar, I. E., Venkatasubramanian, S., Scheidegger, C., Friedler, S. 2020. Problems with Shapley-value-based explanations as feature importance measures. *In International Conference on Machine Learning*, 5491-5500, PMLR.

Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W., Müller, K. R. 2019. Unmasking Clever Hans predictors and assessing what machines really learn. *Nature communications, 10*(1), 1096.

Landrum, G. 2016. RDKit: Open-source cheminformatics

Lipinski, C. A. 2004. Lead-and drug-like compounds: the rule-of-five revolution. *Drug discovery today: Technologies, 1*(4), 337-341.

Lipinski, C. A., Lombardo, F., Dominy, B. W., Feeney, P. J. 2012. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced drug delivery reviews, 64*, 4-17.

Liu, J., Zhu, Y., He, Y., Zhu, H., Gao, Y., Li, Z., Li, W. 2020. Combined pharmacophore modeling, 3D-QSAR and docking studies to identify novel HDAC inhibitors using drug repurposing. *Journal of Biomolecular Structure and Dynamics, 38*(2), 533-547.

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Lee, S. I. 2020. From local explanations to global understanding with explainable AI for trees. *Nature machine intelligence, 2*(1), 56-67.

Lundberg, S. M., Lee, S. I. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems, 30*, 1-10.

Miller, T. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence 267*, 1–38.

Murphy, K. P. 2012. Machine learning: a probabilistic perspective. MIT press.

Moreno, D. A., Scrideli, C. A., Cortez, M. A. A., De Paula Queiroz, R., Valera, E. T., Da Silva Silveira, V., Tone, L. G. 2010. Differential expression of HDAC3, HDAC7 and HDAC9 is associated with prognosis and survival in childhood acute lymphoblastic leukaemia. *British journal of haematology, 150*(6), 665-673.

Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R. Yu, B. 2019. Interpretable machine learning: definitions, methods, and applications. arXiv preprint arXiv:1901.04592.

Pacaud, R., Garcia, J., Thomas, S., Munster, P. N. 2023. Clinical Applications of Histone Deacetylase Inhibitors. *In Handbook of Epigenetics*, 793-819. Academic Press.

Thotala, D., Karwas, R.M.,Engelbach, J.A., Garbow, J.R., Hallahan, A.N., DeWees, T.A., Laszlo, A. and Hallahan, D.E., 2015. Valproic Acid Enhances the Efficacy of Radiation Therapy by Protecting Normal Hippocampal Neurons and Sensitizing Malignant Glioblastoma Cells. *Oncotarget, 6*(33), 35004–35022.

Veber, D. F., Johnson, S. R., Cheng, H. Y., Smith, B. R., Ward, K. W., Kopple, K. D. 2002. Molecular properties that influence the oral bioavailability of drug candidates. *Journal of medicinal chemistry, 45*(12), 2615-2623. https://doi.org/10.1021

İlhan Uysal

Ilhan Uysal is a PhD student at the Department of Computer Engineering at Suleyman Demirel University. He received her undergraduate degrees from Suleyman Demirel University, Department of Computer Teaching in 2007 and Computer Engineering in 2017, and his master's degrees from Antalya Bilim University, Department of Electrical and Computer Engineering in 2019. He works in the fields of artificial intelligence, machine learning and drug repurposing.

He continues to carry out her studies on explainable artificial intelligence in the biomedical field. His personal website is https:// https://abs.mehmetakif.edu.tr/iuysal/PersonelDetay/ and www.ilhanuysal.com.

Utku Köse

Dr. Utku Kose received the B.S. degree in 2008 from computer education of Gazi University, Turkey as a faculty valedictorian. He received M.S. degree in 2010 from Afyon Kocatepe University, Turkey in the field of computer and D.S. / Ph. D. degree in 2017 from Selcuk University, Turkey in the field of computer engineering.

Between 2009 and 2011, he has worked as a Research Assistant in Afyon Kocatepe University. Following, he has also worked as a Lecturer and Vocational School – Vice Director in Afyon Kocatepe University between 2011 and 2012, as a Lecturer and Research Center Director in Usak University between 2012 and 2017, and as an Assistant Professor in Suleyman Demirel University between 2017 and 2019. Currently, he is an Associate Professor in Suleyman Demirel University, Turkey. He has more than 200 publications including articles, authored and edited books, proceedings, and reports. He is also one of the editors of of the Biomedical and Robotics Healthcare (CRC Press) and Computational Modeling Applications for Existential Risks (Elsevier) book series. His research interest includes artificial intelligence, machine ethics, artificial intelligence safety, biomedical applications, optimization, the chaos theory, distance education, e-learning, computer education, and computer science. He has also works regarding the field of Literature.

Author Contributions

IU; researching, methodology, writing, software, editing, methodology, data analysis; UK; revising, methodology, data analysis, software.

Conflict of interest

On behalf of all authors, the corresponding author states that there is no conflict of interest.

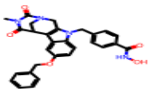
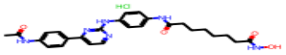
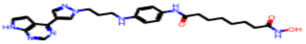
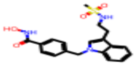
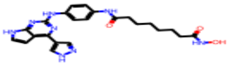
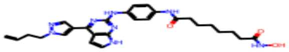
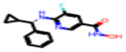
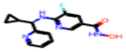
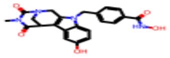
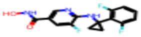
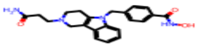
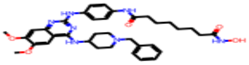
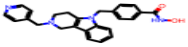
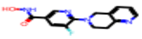
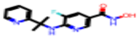
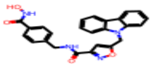
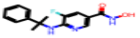
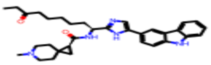
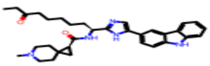
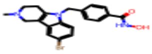
Data Availability

Data supporting the findings of this study are available in the article, in the supplementary material and at https://iuysal1905-childhoodacuteleukemia-drug-interacito-arayuz-r89zld.streamlit.app/

Acknowledging

This research is the product of Ilhan Uysal of PhD thesis work.

APPENDIX 1 - Visualisation of Smiles molecular structures

APPENDIX 2 - Other performance metrics-1

Metric	Value
95% CI	(0.98459,0.99147)
ACC Macro	98.803
ARI	94.022
AUNP	9.725
AUNU	9.725
Bangdiwala B	98.367
Bennett S	97.605
CBA	96.973
CSI	95.615
Chi-Squared	351.193.482
Chi-Squared DF	1
Conditional Entropy	8.137
Cramer V	95.608
Cross Entropy	64.508
F1 Macro	97.797
F1 Micro	98.803
FNR Macro	275
FNR Micro	1.197
FPR Macro	275
FPR Micro	1.197
Gwet AC1	98.356
Hamming Loss	1.197
Joint Entropy	72.633
KL Divergence	12
Kappa	95.594

Kappa 95% CI	(0.94328,0.9686)
Kappa No Prevalence	97.605
Kappa Standard Error	646
Kappa Unbiased	95.594
Krippendorff Alpha	95.594
Lambda A	92.722
Lambda B	92.508
Mutual Information	55.249
NIR	8.355
NPV Macro	98.364
NPV Micro	98.803
Overall ACC	98.803
Overall CEN	8.178
Overall J	(1.9146,0.9573)
Overall MCC	95.608
Overall MCEN	7.019
Overall RACC	72.827
Overall RACCU	72.828
P-Value	None
PPV Macro	98.364
PPV Micro	98.803
Pearson C	69.106
Phi-Squared	91.409
RCI	85.663
RR	1921.0
Reference Entropy	64.496
Response Entropy	63.386

SOA1(Landis & Koch)	Almost Perfect
SOA2(Fleiss)	Excellent
SOA3(Altman)	Very Good
SOA4(Cicchetti)	Excellent
SOA5(Cramer)	Very Strong
SOA6(Matthews)	Very Strong
SOA7(Lambda A)	Very Strong
SOA8(Lambda B)	Very Strong
SOA9(Krippendorff Alpha)	High
SOA10(Pearson C)	Strong
Scott PI	95.594
Standard Error	175
TNR Macro	9.725
TNR Micro	98.803
TPR Macro	9.725
TPR Micro	98.803
Zero-one Loss	46

APPENDIX 3 Other performance metrics-2

Metric	Class 0	Class 1
ACC(Accuracy)	98.803	98.803
AGF(Adjusted F-score)	97.283	98.295
AGM(Adjusted geometric mean)	98.288	969
AM(Difference between automatic and manual classification)	-18	18
AUC(Area under the ROC curve)	9.725	9.725
AUCI(AUC value interpretation)	Excellent	Excellent
AUPR(Area under the PR curve)	96.328	99.286

BB(Braun-Blanquet similarity)	94.937	99.009
BCD(Bray-Curtis dissimilarity)	234	234
BM(Informedness or bookmaker informedness)	94.501	94.501
CEN(Confusion entropy)	20.844	5.727
DOR(Diagnostic odds ratio)	428.035.714	428.035.714
DP(Discriminant power)	200.214	200.214
DPI(Discriminant power interpretation)	Fair	Fair
ERR(Error rate)	1.197	1.197
F0.5(F0.5 score)	9.715	99.119
F1(F1 score - harmonic mean of precision and sensitivity)	96.308	99.285
F2(F2 score)	95.481	99.452
FDR(False discovery rate)	228	991
FN(False negative/miss/type 2 error)	32	14
FNR(Miss rate or false negative rate)	5.063	436
FOR(False omission rate)	991	228
FP(False positive/type 1 error/false alarm)	14	32
FPR(Fall-out or false positive rate)	436	5.063
G(G-measure geometric mean of precision and sensitivity)	96.318	99.286
GI(Gini index)	94.501	94.501
GM(G-mean geometric mean of specificity and sensitivity)	97.223	97.223
HD(Hamming distance)	46	46
IBA(Index of balanced accuracy)	90.149	98.896
ICSI(Individual classification success index)	92.657	98.573
IS(Information score)	257.058	24.491
J(Jaccard index)	92.879	98.581
LS(Lift score)	5.9405	118.502

MCC(Matthews correlation coefficient)	95.608	95.608
MCCI(Matthews correlation coefficient interpretation)	Very Strong	Very Strong
MCEN(Modified confusion entropy)	33.456	9.969
MK(Markedness)	96.729	0.

APPENDIX 4 - Effect of FP molecules on the model and bit substructure

FP	Contribution to the model	Bit Substructure
PubChemFP392	-0.62	N(~C)(~C)(~H)
PubChemFP374	-0.23	C(~H)(~H)(~H)
PubChemFP19	-0.22	>= 2 O
PubChemFP698	-0.19	O-C-C-C-C-C-C-C
PubChemFP193	-0.16	>= 3 saturated or aromatic carbon-only ring size 6
PubChemFP569	-0.15	N-C-C-N
PubChemFP666	+0.13	C=C-C-O-C
PubChemFP391	-0,13	N(~C)(~C)(~C)
PubChemFP335	+0,13	C(~C)(~C)(~C)(~H)
PubChemFP385	-0.35	C(:C)(:C)(:C)
PubChemFP185	-0.15	>= 2 any ring size 6
PubChemFP517	+0.11	N-N-C-N
PubChemFP717	+0.08	Cc1ccc(Cl)cc1
PubChemFP403	-0.06	N(:C)(:C)
PubChemFP37	+0.06	>= 1 Cl
PubChemFP115	+0.04	>= 1 any ring size 3
PubChemFP181	-0.04	>= 1 saturated or aromatic heteroatom-containing ring size 6
PubChemFP684	+0.03	O=C-C-C-C-C