International Journal of Computing and Digital Systems
ISSN (2210-142X)
Int. J. Com. Dig. Sys. #, No.# (Mon-20..)

Interpretable Machine Learning in Drug Discovery: QSAR Modeling of Molecular Properties for Alzheimer's Disease Using Random Forest

Author 1 Alyssa Imani[1], Author 2 Alexander Agung Santoso Gunawan[2], and Author 3 Derwin Suhartono[2]

[1] Mathematics Department, School of Computer Science, Bina Nusantara University, Jakarta 11480, Indonesia
[2] Computer Science Department, School of Computer Science, Bina Nusantara University, Jakarta 11480, Indonesia

E-mail address: alyssa.imani@binus.ac.id, aagung@binus.edu, dsuhartono@binus.edu

Received ## Mon. 20##, Revised ## Mon. 20##, Accepted ## Mon. 20##, Published ## Mon. 20##

Abstract: *Drug development has traditionally been expensive and time consuming. Computational approaches such as machine learning have been widely applied to improve efficiency, yet interpreting prediction outcomes remains a challenge. This study aims to improve the efficiency of Alzheimer's drug discovery by conducting QSAR (Quantitative Structure Activity Relationship) modelling with Random Forest model to predict the inhibition potential (IC50 values) of each Alzheimer's drug candidate compound. A total of 5779 compounds were collected from ChEMBL and PubChem databases. The QSAR model in this study was built using features that were extracted by generating 1024 Morgan Fingerprints representing the substructure of compounds. In this study, SHapley Additive exPlanations (SHAP) are implemented to understand locally and globally important features from the prediction results of the developed model. The effectiveness of the QSAR model in this study was tested with 10-fold cross validation, where the developed regression model can achieve a MAPE score of 11.10% and the classification model achieves an AUC-ROC score of 84.77%. In this work, molecular docking is conducted to simulate how a drug binds to its target and verify the best molecules' effectiveness. Additionally, a web based application was developed in this study to facilitate predicting the bioactivity value of Acetylcholinesterase (AChE) inhibitors.*

Keywords: Random Forest, SHAP, QSAR modeling, Alzheimer, Drug Discovery, Molecular Docking

1. Introduction

Alzheimer's disease is a brain disease that causes a gradual decline in memory, thinking ability and behaviour. This disease is the most common cause of dementia, which is a condition that causes a decrease in mental function that interferes with daily activities. The early symptoms of Alzheimer's disease are usually not obvious, but some early symptoms may occur, such as difficulty remembering new things, difficulty completing daily tasks, difficulty finding words, difficulty understanding new information, and mood changes. or behaviour. Alzheimer's disease cannot be treated yet, but medication such as rivastigmine can help delay the illness's progression by blocking cholinesterase [1]. Annually, the number of people with Alzheimer's disease is still rising. However, designing one drug still takes more than ten years with expensive costs. In general, drug development consists of pre-discovery, preclinical development, clinical trials and reviewing stages. In the initial stage, researchers screen candidate drug compounds, and this stage up to preclinical development can take 5-6 years [2].

Due to the limitation of the wet lab approach, it is not efficient to test all possible chemicals as therapeutic candidates manually. Afterwards, in silico studies (computational approaches) became widely used to help increase efficiency. Where, the implementation of Artificial Intelligence in recent Drug Target Interaction (DTI) studies is enabling cost-effectiveness [3]. In the process of screening the Alzheimer's drug candidates, it is

E-mail:alyssa.imani@binus.ac.id, aagung@binus.edu, dsuhartono@binus.edu

crucial to analyze the drug target interaction with the target enzyme that is responsible for Alzheimer’s disease.

To be an effective drug, a compound must be able to reach the target enzyme in the body at a sufficient concentration level so that it can remain in bioactive form until the desired biological process occurs [4]. In this study, Acetylcholinesterase (AChE) was selected as the target enzyme that is responsible for Alzheimer’s Disease. This study aims to conduct a DTI study by designing a Quantitative structure-activity relationship (QSAR) model. QSAR is used in drug discovery to predict biological activities and toxicity in a way to screen out compounds that don't have drug-like properties [5]. Where, the drug compounds need to consider several basic aspects such as absorption, distribution, metabolism, excretion, and toxicity (ADMET).

In this research, to improve the efficiency of the screening process for candidate Alzheimer's drug compounds, a Random Forest model was developed to predict a bioactivity value. Aside from machine learning's capability to boost efficiency in processing vast amounts of data, interpretability issues in machine learning implementation have an impact on public confidence in its use in drug development and genomics. Ethical problems and discrimination have also contributed to the widespread discussion of machine learning interpretation approaches in recent years. So, in this study, a prediction interpretation study will be conducted on some data points using SHAP (SHapley Additive Explanations) to interpret prediction outcomes both locally and globally.

2. Related Works

A. SARS-CoV-2 3CLpro Inhibitor Classification

The study [6] developed a neural network to identify the bioactivity class of SARS-CoV-2 3CLpro protein inhibitor. The dataset used in this study collected from ChEMBL and PubChem databases contain over 300,000 experimental data from screening SARS-CoV-2 3CLpro inhibitors. In this study Lipinski and PaDEL descriptors were examined as feature extraction methods. A various ensamble models were trained in this study including Random Forest, Bagging, Extra Tree, LGBM, XGB, and AdaBoost. A neural network model was also designed in this study and outperformed the ML methods with 93% accuracy. The performance of models trained with PaDEL descriptors outperformed and suitable for high-throughput QSAR modeling.

The Explanatory factor identified in this study by implementing SHapley Additive exPlanations (SHAP) on the XGB classifier. The SHAP model could improve the interpretability of XGB model by finding the important fingerprints from PaDEL descriptors. The SHAP provides a more comprehensive and comprehensible depiction of the feature importances compared to the conventional approaches such as feature importance scores. Because SHAP values account for feature interaction, allowing for a deeper understanding of how each feature influences the model’s prediction.

B. Antimalarial Predictive Models

Antimalarial medication resistant happening for Chloroquine and Artemisinin-based Combination Treatment (ACT), consequently malaria became endemic in most locations. The study [7] implemented and compared five various ML including Support Vector Machine (SVM), Random Forest (RF), Extreme Gradient Boost (XGB), Logistic Regression (LR) and Artificial Neural Network (ANN) to build antimalarial predictive models. Those models were developed to predict the bioactivity class of drug against Plasmodium Falciparum Parasite. From the ChEMBL and PubChem databases, a total of 4794 compounds were retrieved and extracted into 1444 PaDEL descriptors.

The classification of anti-plasmodial activities in this study conducted with a threshold $IC_{50} \leq 1\mu M$ as active compounds and $IC_{50} > 1\mu M$ as inactive compounds. In this study various numbers of features were used and selected with Recursive Feature Elimination (RFE). The result shows XGB model with 361 features, reach the best recall of the ‘active’ label with 0.81 and F1 score of 0.83. The XGB model outperformed the designed ANN model which achieved the recall of the ‘active’ and F1 score of 0.79 and 0.80, respectively. This study implies that without compromising much precision, the XGB and ANN could identify the new anti-malaria drug formation around 81% and 79%, respectively.

C. ChemBERTa

The research [8] builds a model to forecast the molecular characteristics of SMILES strings using a Natural Language Processing (NLP) approach. Based on the RoBERTa transformer architecture, ChemBERTa is a model that was trained using the PubChem dataset, which has 77 million SMILES strings. ChemBERTa was created by combining six layers and twelve attention heads, which produced seventy-two distinct attention mechanisms. HuggingFace library's Byte-Pair Encoder (BPE) serves as the foundation for the tokenizer created on the ChemBERT model. Tokenization at both the character and word levels is combined in BPE, a hybrid tokenization technique. When it comes to several categorization tasks from MoleculeNet and attention-based visualization modalities, this model performs competitively. This model requires significant computational resources for training and inference compared to simple machine learning models. The size and interpretability of the model also needs to be considered, since it can be challenging to interpret the internal workings on complex models.

3. Material and Methods

A. Datasets

The first dataset collected for this study is obtained from the public open-source database ChEMBL. ChEMBL is a database that contains manually curated bioactive molecules with drug-like properties [9]. After cleaning 5664 data, 3575 compounds in total were obtained. Every compound in the dataset is represented using SMILES and has an IC50 bioactivity value. SMILES is a common representation of molecules that emphasizes the use of molecular graph theory to allow for rigorous structural specification using natural and basic grammar [10]. Table 7 provides a sample of the dataset. Every data point contains the ChEMBL ID, SMILES, and IC50 value. Where IC50 is a bioactivity value which can measure the concentration of compounds needed to inhibit the biological or biochemical function of a protein target by 50% [11].

TABLE I. Sample of CHEMBL Dataset

ChEMBL ID	SMILES	IC_{50}
CHEMBL133897	CCOc1nn(-c2cccc(OCc3ccccc3)c2)c(=O)o1	750.0
CHEMBL336398	O=C(N1CCCCC1)n1nc(-c2ccc(Cl)cc2)nc1SCC1CC1	100.0
CHEMBL131588	CN(C(=O)n1nc(-c2ccc(Cl)cc2)nc1SCC(F)(F)F)c1ccccc1	50000.0
CHEMBL130628	O=C(N1CCCCC1)n1nc(-c2ccc(Cl)cc2)nc1SCC(F)(F)F	300.0
CHEMBL130478	CSc1nc(-c2ccc(OC(F)(F)F)cc2)nn1C(=O)N(C)C	800.0

The second dataset collected from PubChem Database. In total, there are 115 AChE inhibitors collected in SMILES string. This dataset does not contain any bioactivity value. In this study, this collected dataset used for validating the predicted inhibition potency by conducting molecular docking.

B. QSAR Modeling

QSAR Modeling was developed in this study with Random Forest to determine the association between chemical compounds' structural features and the biological activity of Alzheimer's medicines. Using a variety of mathematical techniques, QSAR aims to associate structural, chemical, statistical, and physical attributes with biological potency. The physicochemical properties are taken into account, including partition coefficient and the existence of certain chemical features.

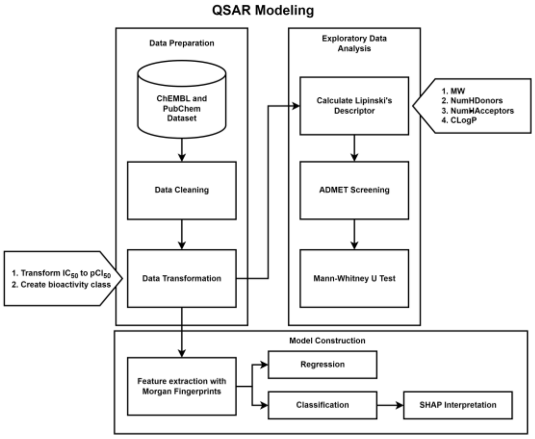


Figure 1. Workflow of QSAR Modeling for examining the AChE inhibitor candidates.

The QSAR Modeling begins with data preparation, which includes data collection from the ChEMBL and PubChem databases, as well as data cleansing and transformation. Then, exploratory data analysis was carried out by computing Lipinski's Descriptor, ADMET Screening, and statistical analysis using the Mann-Whitney U Test. In the final stage of QSAR Modeling, a model is created using Random Forest and trained using compound data that has been extracted using Morgan Fingerprints. Then, the Random Forest model's predictions were analyzed using the SHAP approach.

1) Data Preparation

Before the ChEMBL dataset was used for model construction, first prepared by removing redundant data and missing values data. Each data is SMILES of a compound that represents a candidate for AChE inhibitor. The bioactivity value in IC50 is used as a label for constructing the regression model. However, the collected dataset has a wide range of IC_{50} values, so it converted into negative logarithmic in molar concentration units (M). The conversion calculation shown below:

$$pIC_{50} = -\log_{10}(IC_{50}) \tag{1}$$

Two bioactivity classes 'active' and 'inactive' compounds are created for performing classification prediction. According to prior research, 'active' compounds have an IC_{50} value < 1µM and 'inactive' compounds have an IC_{50} value > 10µM [12]. The calculations for converting IC_{50} to pIC_{50} for each bioactivity class are shown below:

- Active

$$\begin{gathered} IC_{50} < 1\mu M = IC_{50} < 10^{-6}M \\ pIC_{50} = -\log_{10}(10^{-6}M) \\ pIC_{50} > 6 \end{gathered} \tag{2}$$

- Inactive

$$\begin{gathered} IC_{50} > 10\mu M = IC_{50} > 10^{-5}M \\ pIC_{50} = -\log_{10}(10^{-5}M) \\ pIC_{50} < 5 \end{gathered} \tag{3}$$

2) *Exploratory Data Analysis*

The exploratory data analysis in this study was carried out to investigate the bioactivity class 'active' and 'inactive' from two different populations. It is conducted by doing statistical analysis and screening of drug candidate molecules based on Lipinski's Rule of Five, where medications that can be ingested orally need to match the following requirements:

- Molecular weight < 500 Daltons
- Hydrogen bond donors < 5
- Hydrogen bond acceptors < 10
- The logarithm of octanol-water partition coefficient (ClogP) < 5 or (MlogP < 4.15)

Four new features including molecular weight (MW), hydrogen bond donors (NumHDonors), hydrogen bond acceptors (NumHAcceptors), and the logarithm of octanol-water partition coefficient (ClogP) calculated to conduct a statistical analysis. The statistical analysis performed with Mann-Whitney U test to evaluate the hypothesis H0: bioactivity classes ‘active’ and ‘inactive’ come from the same population.

TABLE II. MANN-WHITNEY U TEST RESULT

Feature	**Statistics**	**P-value**	**α**	**Interpretation**
pIC50	882716.5	0.048372	0.05	Reject H0
MW	823729.0	0.000001	0.05	Reject H0
NumHDonors	850778.5	0.000198	0.05	Reject H0
NumHAcceptors	879139.0	0.028911	0.05	Reject H0
ClogP	859996.5	0.002051	0.05	Reject H0

Based on the Mann-Whitney U Test result, the statistical test results on all features successfully rejected hypothesis H0. In summary, the groupings of compounds with the bioactivity classes "active" and "inactive" do not originate from the same data population.

3) *Model Construction*

All SMILES in ChEMBL dataset were extracted into Morgan Fingerprints before used in the model construction process. Morgan Fingerprints, also known as circular fingerprints, are vectors that depict the substructure of molecules with different atomic radii [13]. In total there are 1024 Morgan Fingerprints that were generated from all SMILES strings that represent each compound in ChEMBL dataset.

This study constructed two models for different tasks: regression and classification. The regression model was created to predict the bioactivity value pIC_{50}, and a classification model developed to distinguish the bioactivity class ‘active’ and ‘inactive’. Both models developed using Random Forest, which is an ensemble model consisting of multiple decision trees. A simple algorithm was chosen in order to help the prediction easily to interpret. Where each model construction is evaluated with 10-fold cross validation.

It is easy to interpret the prediction from a single decision tree, but it would be challenging to interpret multiple decision trees. Therefore, in this study an explainer model known as SHAP was implemented to help explain the prediction results.

C. *Evaluation Metrics*

The main metric that will be used in evaluating the regression model is Mean Absolute Percentage Error (MAPE) score.

$$M = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_t}\right| \tag{4}$$

Where A_t is the actual value and F_t is the forecast value. MAPE was chosen to evaluate the regression model because it shows the error in percentage and makes it easy to compare with different datasets or model performances.

Apart from MAPE, R^2 is used to measure the dependency between features and the prediction result. R^2 calculations can be done as follows:

- The sum of squares of residuals

$$SS_{res} = \sum_i (y_i - f_i)^2 \tag{5}$$

- The total sum of squares

$$SS_{tot} = \sum_i (y_i - \bar{y})^2 \tag{6}$$

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} \tag{7}$$

Evaluation of the classification model was done using the AUC-ROC (Area Under the Receiver Operating Characteristic Curve) metric. AUC-ROC measures the two-dimensional area under the ROC curve, with values ranging from 0 to 1. AUC-ROC equals to one indicates a model with perfect performance. The ROC curve has two parameters:

- TPR (True Positive Rate)

$$TPR = \frac{TP}{TP + FN} \tag{8}$$

- FPR (False Positive Rate)

$$FPR = \frac{FP}{FP + TN} \tag{9}$$

The F1 metric is also used to evaluate the performance of the classification model. F1 calculates the average of precision and recall which is mathematically defined as follows:

$$precision = \frac{TP}{TP + FP} \tag{10}$$

$$recall = \frac{TP}{TP + FN} \tag{11}$$

$$F1 = \frac{2 \times precision \times recall}{precision + recall} \tag{12}$$

D. SHAP

Shapley Additive exPlanations (SHAP) is a method developed to explain prediction of an instance by calculating the contribution of each feature to the prediction result [14]. SHAP was designed based on Shapley values which is one of the game theory concepts. SHAP was developed with unification concept and shows improved computational performance and/or better consistency of human intuition than previous approaches [15].

Suppose f is the original prediction model which will be explained by the explanation model g. Explanation model g is a simpler model that estimates a more complex model f. The explanation model g is additive, which means that the explanation carried out based on the sum of the contributions of each feature. Mathematically, the concept of additive feature attribution in the SHAP method is defined as a linear function of binary variables as follows:

$$g(z') = \phi_0 + \sum_{i=1}^{M} \phi_i z'_i \tag{13}$$

Where $z' \in \{0,1\}^M$, M is the number of input features simplified to binary values (0 or 1), and $\phi_i \in R$ is the attribution for each feature.

4. Results and Discussion

A. Model Performances

This research develops a Random Forest model to predict the bioactivity of Alzheimer's drug candidate compounds in two schemes, regression and classification. 10-fold cross-validation was used to evaluate the Random Forest performance to predict the pIC$_{50}$ value. The result is shown in the following table:

TABLE III. Regression Performance

Fold	MAPE	R^2
Fold-1	0.1155	0.7080
Fold-2	0.1119	0.7168
Fold-3	0.1017	0.7638
Fold-4	0.1257	0.6902
Fold-5	0.1003	0.7688
Fold-6	0.1124	0.7493
Fold-7	0.1092	0.7624
Fold-8	0.1020	0.8030
Fold-9	0.1121	0.7023
Fold-10	0.1191	0.6772
Average	**0.1110**	**0.7342**
Std	**0.0077**	**0.0388**

The standard deviation value for the 10-fold cross-validation indicates that there is not much variation in the MAPE and regression model values between folds. In other words, the model has sufficient stability for ten trials using random data. The prediction performance is deemed acceptable, with an average MAPE score of 11.10% indicating a reasonably low error. According to the average R^2 value, 73.42% of the variability in the target data can be explained by the regression model.

The table below shows the performance of the classification model using the Random Forest Classifier created for this study.

TABLE IV. Classification Performance

Fold	AUC-ROC	F1 Score
Fold-1	0.8381	0.8703
Fold-2	0.8350	0.8692
Fold-3	0.8368	0.8730
Fold-4	0.8590	0.8847
Fold-5	0.8663	0.8872

Fold-6	0.8277	0.8575
Fold-7	0.8595	0.8835
Fold-8	0.8292	0.8656
Fold-9	0.8781	0.9008
Fold-10	0.8468	0.8764
Average	0.8477	0.8768
Std	0.0162	0.0118

Based on the standard deviation values, Table IV illustrates the F1 and AUC-ROC values in the classification model are stable in 10-fold cross-validation. Based on the average AUC-ROC score, which is 84.77%, this classification model performs well in differentiating across classes. The F1 value of 87.68% indicates that the model's classification performance also demonstrates a good balance between precision and recall values.

B. SHAP Interpretation

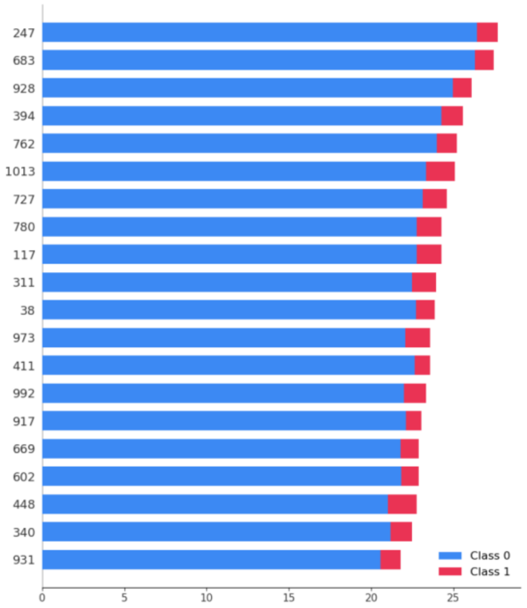


Figure 2. SHAP Summary of PubChem Dataset Prediction.

Among 1024 Morgan fingerprints, the bar chart shows the 20 most important substructures.

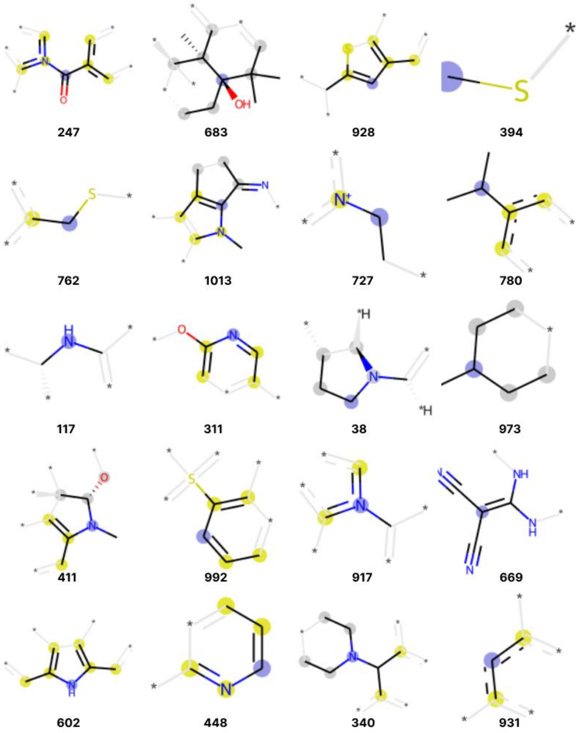


Figure 3. 20 most important Morgan Fingerprints.

Fig.3 shows the visualization of 20 most important Morgan Fingerprints based on SHAP summary results. Those important features include Morgan Fingerprints 247, 683, 928, 394, 762, 1013, 727, 780, 117, 311, 38, 973, 411, 992, 917, 669, 602, 448, 340, and 931. Overall, the features displayed in the SHAP summary have a class 0 value higher than class 1. This indicates that the absence of these features further increases their dominance in the predicted results.

To find out feature importance locally, you can see the results of the SHAP force plot on a data point. A SHAP force plot on a data point with the bioactivity class prediction "active" is shown below.

TABLE V. DETAIL OF SELECTED 'ACTIVE' COMPOUND

SMILES	*COC1=C(C=C2C(=C1)CC(C2=O)CC3CCN(CC3)C C4=CC=CC=C4)OC*
pIC$_{50}$	*8.1036*
Bioactivity class	*Active*

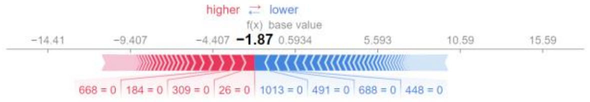


Figure 4. SHAP Force Plot of Selected ‘Active’ Compound.

The chosen 'active' compound's forecast result, f(x), in the force plot is -1.87 below average. As seen by the 'red' arrow displays 668=0, 184=0, 309=0, and 26=0, meaning the absence of these substructures has a greater impact on the bioactivity prediction into the 'active' class. In the other hand, the 'blue' arrow displays 1013=0, 491=0, 688=0, and 448=0, indicating that the absence of these substructures decreases the selected compound predicted to be an 'active' class.

The following is an example of a force plot for compounds with predicted bioactivity class 'inactive' classification results.

TABLE VI. DETAIL OF SELECTED 'INACTIVE' COMPOUND

SMILES	CCCCN(CCCC)SN(C)C(=O)OC1=CC=CC2=C1OC(C2)(C)C
pIC$_{50}$	*4.9145*
Bioactivity class	*Inactive*

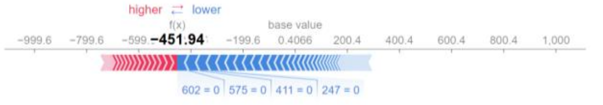


Figure 5. SHAP Force Plot of Selected ‘Inactive’ Compound.

The force plot findings indicate that there is a significant difference of -451.94 between the basis value and the expected outcomes of f(x). This demonstrates that the expected value for this compound is significantly less than the average expected value derived from the data that was utilized to train the model. Subsequently, it demonstrates that no feature increases the prediction outcomes of these data points in the bioactivity class 'inactive' categorization significantly. On the other hand, the 'blue' arrow indicates that features 602, 575, 422, and 247 have a value of 0. The number 0 denotes the lack of substructures 602, 575, 422, and 247, which lessens the impact on the compound data's specific "inactive" bioactivity class prediction.

C. *Molecular Docking*

In general, the molecular properties of compounds such as pIC$_{50}$ are obtained manually through research results from the wet lab. Molecular Docking is a method that investigates interactions of ligand which is a small molecule with a target protein's binding site [16]. Protein-ligand docking in this research predicts the position and orientation of an Alzheimer's drug candidate compound as ligand on the Acetylcholinesterase protein as receptor.

In this study, molecular docking is used to validate the prediction results of the Random Forest model, for a new dataset that does not yet have labels. The dataset used in this analysis is a dataset taken through the PubChem database. In this study, only compounds with the highest and lowest pIC$_{50}$ bioactivity values are selected to be conducted on molecular docking.

TABLE VII. MOLECULAR DOCKING RESULTS ON THE 'ACTIVE' LIGAND

Mode	Affinity (kcal/mol)	Dist from best mode	
		rmsd l.b.	rmsd u.b.
1	-13.18	0	0
2	-11.42	3.32	10.33
3	-11.34	2.023	3.081
4	-11.02	1.609	2.223
5	-10.92	3.584	10.3
6	-10.7	2.843	4.706
7	-10.67	3.593	9.67
8	-10.6	1.594	2.529
9	-10.51	3.551	10.97
Std	**0.7778**	**0.8326**	**3.6691**

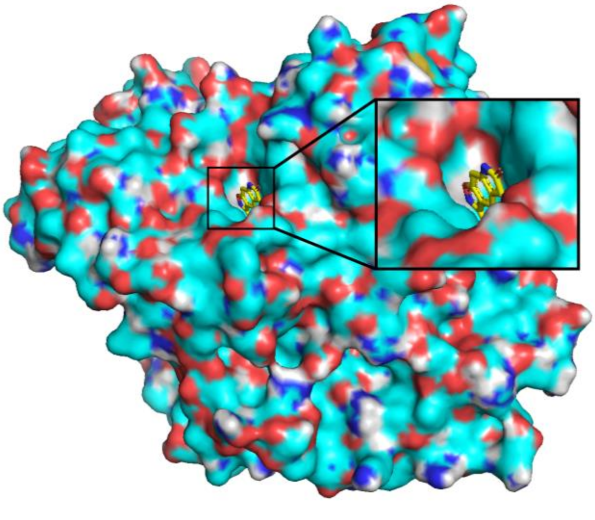


Figure 6. The docking pose of the ‘active’ ligand.

Table VII shows molecular docking result on an ‘active’ ligand with nine docking modes on the target protein Acetylcholinesterase. The docking mode represents one possible orientation and conformation of the ligand at the protein target binding site. Based on these results, docking mode 1 has the lowest affinity value at -13.18 kcal/mol, indicating that the ligand can bind very strongly to the protein target.

Fig. 6 shows the docking pose of the 'active' ligand located in the pocket binding site of target protein. This shows that the compound chosen as the 'active' ligand has the potential to properly inhibit the biological function of the target protein Acetylcholinesterase. In other words, the compound could be a good candidate for an Alzheimer's drug.

TABLE VIII. MOLECULAR DOCKING RESULTS ON THE 'INACTIVE' LIGAND

Mode	Affinity (kcal/mol)	Dist from best mode	
		rmsd l.b.	rmsd u.b.
1	-5.845	0	0
2	-5.488	3.303	6.034
3	-4.694	3.34	5.013
4	-4.586	9.233	10.2
5	-4.44	8.409	10.04
6	-4.272	31.37	32.14
7	-4.202	8.214	10.15
8	-4.173	9.814	12.34
9	-4.123	28.61	29.96
Std	**0.5802**	**10.2252**	**9.8350**

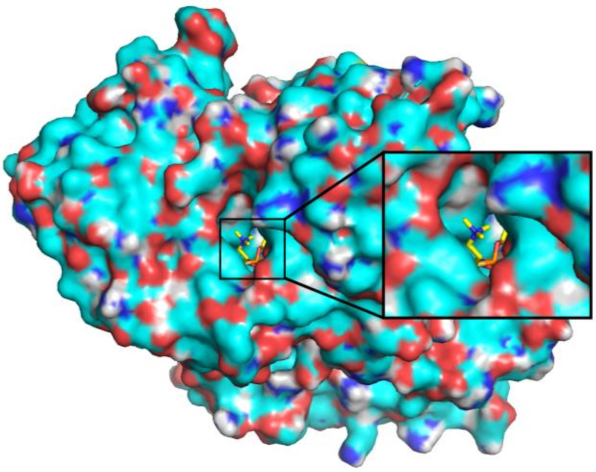


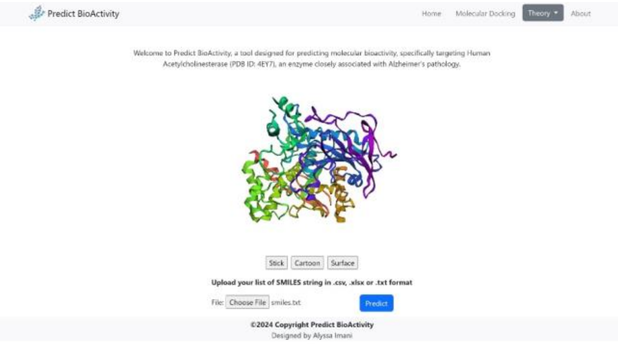
Figure 7. The docking pose of the 'inactive' ligand.

Table VIII and Fig. 7 show the results of molecular docking on 'inactive' ligands. The lowest binding affinity of the 'inactive' ligand is -5,845 kcal/mol, which is much higher than that of the 'active' ligand. This shows that the 'inactive' ligand has a weaker binding interaction than the 'active' ligand.

Based on the visualization of the docking pose for the 'inactive' ligand, the position of the ligand is still in the pocket binding site of the target protein. This would make sense considering that all data collected through the PubChem database is Acetylcholinesterase inhibitors. So, all PubChem ligands that are predicted to be 'inactive' also still have the potential to inhibit the function of the protein target. However, the standard deviation values for the RMSD lower bound and upper bound for 'inactive' ligands are much higher than those for 'active' ligands. This shows that the 'inactive' ligand has much lower conformational flexibility or specificity when binding to the protein target. Conformational flexibility in 'inactive' ligands can make the ligand structure less able to lock properly at the binding site of the protein target. This allows the inhibitory potential of 'inactive' ligands to be lower compared to 'active' ligands.

D. Web Application Development

The development of a web application has also been done in this research. The web application developed with Django framework and Python as the programming language. The system was designed as a web application since mostly bioactive prediction tools such as SwissTargetPrediction [17] were developed as web applications. Here are some interfaces and features of designed web application in this research:



Welcome to Predict BioActivity, a tool designed for predicting molecular bioactivity, specifically targeting Human Acetylcholinesterase (PDB ID: 4EY7), an enzyme closely associated with Alzheimer's pathology.

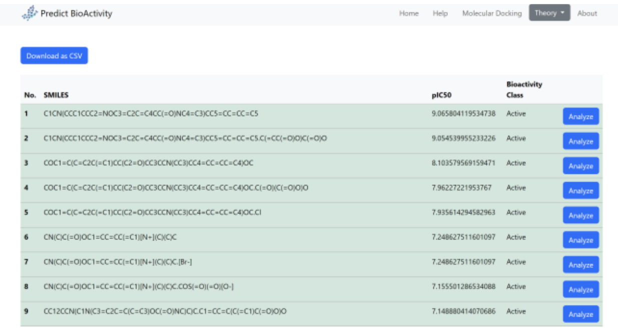
Upload your list of SMILES string in .csv, .xlsx or .txt format

File: Choose File smiles.txt

©2024 Copyright Predict BioActivity
Designed by Alyssa Imani

Figure 8. The Home Page.

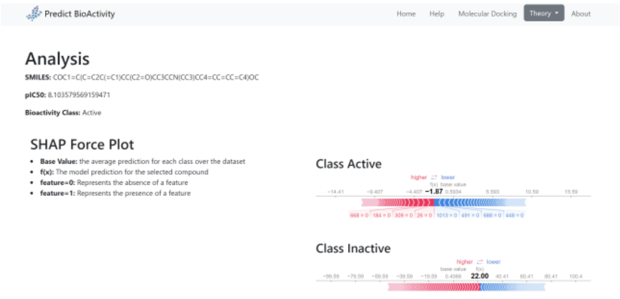
In the Home page, user can see the visualization of 3D structure of target enzyme AChE. To obtain the bioactivity prediction result, users can input the chemicals in the SMILES string using a format file (.csv, .xlsx, or .txt) and then click the predict button.



No.	SMILES	pIC50	Bioactivity Class	
1	C1CN(CCC1CCC2=NOC3=C2C=C4CC(=O)NC4=C3)CC5=CC=CC=C5	9.065804119534738	Active	Analyze
2	C1CN(CCC1CCC2=NOC3=C2C=C4CC(=O)NC4=C3)CC5=CC=CC=C5.C(=CC(=O)O)C(=O)O	9.054539955233226	Active	Analyze
3	COC1=C(C=C2C(=C1)CC(C2=O)CC3CCN(CC3)CC4=CC=CC=C4)OC	8.103579569159471	Active	Analyze
4	COC1=C(C=C2C(=C1)CC(C2=O)CC3CCN(CC3)CC4=CC=CC=C4)OC.C(=O)(C(=O)O)O	7.9622722195376‌7	Active	Analyze
5	COC1=C(C=C2C(=C1)CC(C2=O)CC3CCN(CC3)CC4=CC=CC=C4)OC.Cl	7.935614294582963	Active	Analyze
6	CN(C)C(=O)OC1=CC=CC(=C1)[N+](C)(C)C	7.248627511601097	Active	Analyze
7	CN(C)C(=O)OC1=CC=CC(=C1)[N+](C)(C)C.[Br-]	7.248627511601097	Active	Analyze
8	CN(C)C(=O)OC1=CC=CC(=C1)[N+](C)(C)C.COS(=O)(=O)[O-]	7.155501286534088	Active	Analyze
9	CC12CCN(C1N(C3=C2C=C(C=C3)OC(=O)NC)C)C.C1=CC=C(C(=C1)C(=O)O)O	7.148880414070686	Active	Analyze

Figure 9. The Prediction Result Page.

The prediction result will be shown in a table and sorted descending based on the pIC_{50}. On this page users can download the prediction result as csv file and proceed local analysis for a compound.



Predict BioActivity

Analysis

SMILES: COC1=C(C=C2C(=C1)CC(C2=O)CC3CCN(CC3)CC4=CC=CC=C4)OC

pIC50: 8.103579569159471

Bioactivity Class: Active

SHAP Force Plot

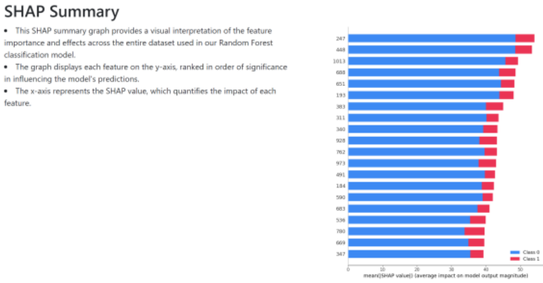
- **Base Value:** the average prediction for each class over the dataset
- **f(x):** The model prediction for the selected compound
- **feature=0:** Represents the absence of a feature
- **feature=1:** Represents the presence of a feature

Class Active

Class Inactive

Figure 10. SHAP local interpretation.

The local analysis will show SHAP force plot for each class ‘active’ and ‘inactive’.

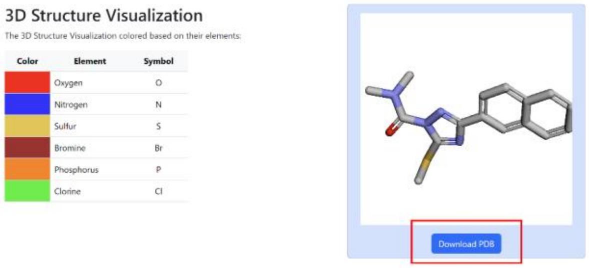


SHAP Summary

- This SHAP summary graph provides a visual interpretation of the feature importance and effects across the entire dataset used in our Random Forest classification model.
- The graph displays each feature on the y-axis, ranked in order of significance in influencing the model's predictions.
- The x-axis represents the SHAP value, which quantifies the impact of each feature.

Figure 11. SHAP Summary

On the same page, users will get the SHAP Summary, which shows a bar chart highlighting the most important features.



3D Structure Visualization

The 3D Structure Visualization colored based on their elements:

Color	Element	Symbol
🟥	Oxygen	O
🟦	Nitrogen	N
🟨	Sulfur	S
🟫	Bromine	Br
🟧	Phosphorus	P
🟩	Clorine	Cl

Figure 12. The docking pose of the ‘inactive’ ligand.

In this web application, the user can also download the 3D structure of compound in PDB format. Where, a molecular docking analysis requires the compound's three-dimensional structure.

5. Conclusion

In Conclusion, the Random Forest Method can be utilized as a suitable model for QSAR modeling in Alzheimer's drug discovery, considering ease of interpretation and maintaining a respectable degree of prediction accuracy (MAPE regression model 11.10% and AUC-ROC classification model 84.77%). Aside from that, adopting SHAP as an explanation model can help in both local and global interpretation by comprehending the essential elements of the Random Forest classification model prediction outputs in QSAR modeling. According to the molecular docking validation result, the binding affinity and the pIC50 have a negative correlation as expected. Moreover, a web-based tool has been created in this study to help with the screening process of Alzheimer's medication candidate. Code for web application in this study could be access from, https://github.com/alyssaimani/Predict-BioActivity.

Acknowledgment

The author would like to extend sincere gratitude to Mr. Rudi Nirwantono, S.Si., M.Mol.Bio. from Bioinformatics and Data Science Research Center (BDSRC) at Binus University for his invaluable assistance and insightful discussions throughout the course of this research. His expertise and guidance have greatly contributed to the development and refinement of this work.

References

[1] P. H. Patel and V. Gupta, *Rivastigmine*. 2024.

[2] N. Singh, P. Vayer, S. Tanwar, J.-L. Poyet, K. Tsaioun, and B. O. Villoutreix, “Drug discovery and development: introduction to the general public and patient groups,” *Frontiers in Drug Discovery*, vol. 3, May 2023, doi: 10.3389/fddsv.2023.1201419.

[3] D. Suhartono, M. R. N. Majiid, A. T. Handoyo, P. Wicaksono, and H. Lucky, “Towards a more general drug target interaction prediction model using transfer learning,” *Procedia Comput Sci*, vol. 216, pp. 370–376, 2023, doi: 10.1016/j.procs.2022.12.148.

[4] A. Daina, O. Michielin, and V. Zoete, “SwissADME: A free web tool to evaluate pharmacokinetics, drug-likeness and medicinal chemistry friendliness of small molecules,” *Sci Rep*, vol. 7, Mar. 2017, doi: 10.1038/srep42717.

[5] C. S. Kue and S. Kumar, “Nonmammalian models in toxicology screening,” in *Encyclopedia of Toxicology*, Elsevier, 2024, pp. 971–985. doi: 10.1016/B978-0-12-824315-2.00598-4.

[6] F. Bin Ashraf, S. Akter, S. H. Mumu, M. U. Islam, and J. Uddin, “Bio-activity prediction of drug candidate compounds targeting SARS-Cov-2 using machine learning approaches,” *PLoS One*, vol. 18, no. 9, p. e0288053, Sep. 2023, doi: 10.1371/journal.pone.0288053.

[7] M. E. Mswahili, G. L. Martin, J. Woo, G. J. Choi, and Y.-S. Jeong, “Antimalarial Drug Predictions Using Molecular Descriptors and Machine Learning against Plasmodium Falciparum,” *Biomolecules*, vol. 11, no. 12, p. 1750, Nov. 2021, doi: 10.3390/biom11121750.

[8] S. Chithrananda, G. Grand, and B. Ramsundar, “ChemBERTa: Large-Scale Self-Supervised Pretraining for Molecular Property Prediction,” Oct. 2020.

[9] B. Zdrazil *et al.*, “The ChEMBL Database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods,” *Nucleic Acids Res*, vol. 52, no. D1, pp. D1180–D1192, Jan. 2024, doi: 10.1093/nar/gkad1004.

[10] M. Krenn *et al.*, “SELFIES and the future of molecular string representations,” *Patterns*, vol. 3, no. 10, p. 100588, Oct. 2022, doi: 10.1016/j.patter.2022.100588.

[11] Navre M, “Why using pIC50 instead of IC50 will change your life.” Accessed: Dec. 11, 2023. [Online]. Available: https://www.collaborativedrug.com/cdd-blog/why-using-pic50-instead-of-ic50-will-change-your-life

[12] S. Simeon *et al.*, “Probing the origins of human acetylcholinesterase inhibition via QSAR modeling and molecular

docking," *PeerJ*, vol. 4, p. e2322, Aug. 2016, doi: 10.7717/peerj.2322.

[13] B. Sharma *et al.*, "Accurate clinical toxicity prediction using multi-task deep neural nets and contrastive molecular explanations," *Sci Rep*, vol. 13, no. 1, p. 4908, Mar. 2023, doi: 10.1038/s41598-023-31169-8.

[14] C. Molnar, "Interpretable Machine Learning A Guide for Making Black Box Models Explainable," Apr. 2021. [Online]. Available: https://christophm.github.io/

[15] S. M. Lundberg, P. G. Allen, and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," 2017. [Online]. Available: https://github.com/slundberg/shap

[16] N. S. Pagadala, K. Syed, and J. Tuszynski, "Software for molecular docking: a review," *Biophys Rev*, vol. 9, no. 2, pp. 91–102, Apr. 2017, doi: 10.1007/s12551-016-0247-1.

[17] A. Daina, O. Michielin, and V. Zoete, "SwissTargetPrediction: updated data and new features for efficient prediction of protein targets of small molecules," *Nucleic Acids Res*, vol. 47, no. W1, pp. W357–W364, Jul. 2019, doi: 10.1093/nar/gkz382.

Alyssa Imani is an undergraduate student pursuing Computer Science and Mathematics at Bina Nusantara University, Indonesia. She was a research analyst intern at Bioinformatics and Data Science Research Center (BDSRC) in Bina Nusantara University. Her research interests include Artificial Intelligence and Bioinformatics.

Alexander Agung Santoso Gunawan is an Associate Professor at Bina Nusantara University, Jakarta in Indonesia, with 15 years of research experience. He holds a Bachelor of Science in Mathematics from Bandung Institute of Technology, a Master of Automation Engineering from Darmstadt University of Applied Sciences, Germany and a Master of Electrical Engineering from Bandung Institute of Technology. He completed his PhD in Computer Science at University of Indonesia. Alexander has published over 100 research papers. His research interests include Data Science, Geo-AI, Computer Vision, IoT, and Robotics.

Derwin Suhartono received the Ph.D. degree in computer science from Universitas Indonesia, in 2018. He is currently a Faculty Member of Bina Nusantara University, Indonesia. His research interest includes natural language processing. Recently, he is continually doing research in argumentation mining and personality recognition. He actively involves in Indonesia Association of Computational Linguistics (INACL), a National Scientific Association in Indonesia, IndoCEISS, and Aptikom. He has professional memberships in ACM, INSTICC, and IACT. He also takes role as a reviewer in several international conferences and journals.