

Prediction of pIC50 Values for the Acetylcholinesterase (AChE) using QSAR Model

Shobhana A Khedekar¹, Nidhi S Mhatre², Raju Mendhe³

¹Student, Computer Engineering Department, Mumbai University, Datta Meghe College of Engineering, Navi Mumbai, Maharashtra, India

²Student, Computer Engineering Department, Mumbai University, Datta Meghe College of Engineering, Navi Mumbai, Maharashtra, India

³Assistant Professor, Computer Engineering Department, Mumbai University, Datta Meghe College of Engineering, Navi Mumbai, Maharashtra, India

Abstract - To improve the efficiency, effectiveness, and quality of the results generated, targeted drug development and novel drug development approaches now typically include machine learning and deep learning algorithms. Machine learning approaches have been used to develop drug candidates in the field of drug discovery and development. As a result, the consumption and overall time spent on drug development was significantly reduced.

The development of powerful direct and indirect computational approaches such as Quantitative Structure-Activity Ratio (QSAR) has increased the potential of compounds to become drugs. The prediction of drug-target interactions in medicine has a significant impact. It helps in the development of new drugs for various diseases. Traditional drug interactions with targets has many drawbacks. The main disadvantage is that it is costly and time consuming. To solve this problem, a new approach to machine learning has been introduced. Machine learning approaches can be used to accurately and efficiently predict drug-target interactions.

In this paper, we will be using random forest regression model to train our machine to predict the pIC50 values of novel drugs against the Alzheimer's disease target protein, acetylcholinesterase (AChE), which is an enzyme that catalyzes the breakdown of the neurotransmitter acetylcholine, which is necessary for cognition and memory.

Key Words: Random Forest Regression Model, QSAR model, pIC50

1. INTRODUCTION

ChEMBL provided a large nonredundant data set of 5,103 compounds with published IC50 values against AChE, which were used in a quantitative structure activity relationship (QSAR) analysis to learn more about the origins of their bioactivity. AChE inhibitors were described using a set of 12 fingerprint descriptors, and prediction models were built using random forest.

Following that, the substructure fingerprint count was thoroughly examined in order to gain useful information on the inhibitory activity of AChE inhibitors. This information can be applied in a variety of ways. Our paper's goal is to predict pIC50 values using the SMILES notation and the ChEMBL ID as input.

1.1 Dataset

ChEMBL is a carefully selected database of biologically active compounds for medicinal use. It combines chemical data, biological activity data, and genetic data to help transform genomic data into effective new drugs. The human AChE inhibitor data set (subject ID ChEMBL220) was collected using the ChEMBL 20 database containing 7549 compounds. A total of 5,103 unique canonical smiles were obtained after removing duplicate canonical smiles and entries with null canonical smiles.

Model cleaning and building were done using Jupyter Notebook.

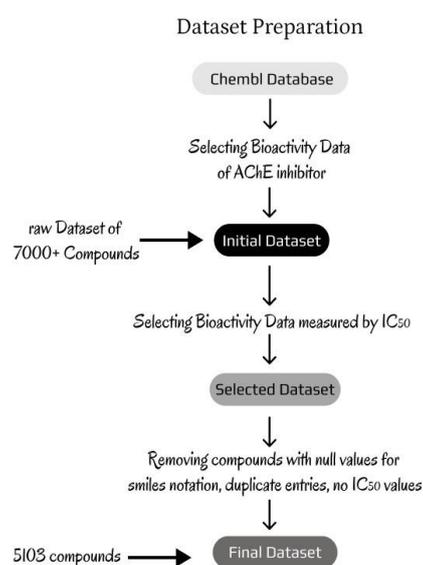


Fig -1: Work flow Dataset Preparation

1.2 Description of inhibitors

Several studies of the use of fingerprints to model biological activity have investigated performance differences between types of fingerprints. Riniker and Landrum (2013) compared and evaluated the performance of predictive models built using RDKit 2D fingerprint descriptors.

Salts were removed prior to descriptor computation using the built-in PaDELDescriptor software function. The

PaDELDescriptor software also calculated the four molecular descriptors used to define Lipinski's law: molecular weight (MW), logarithm of the octanol/water partition coefficient (LogP), number of hydrogen bond donors (NumHDonors), and number of hydrogen bonds. Bond Receptors (NumHAAcceptors).

	MW	LogP	NumHDonors	NumHAcceptors
0	312.325	2.8032	0.0	6.0
1	376.913	4.5546	0.0	5.0
2	426.851	5.3574	0.0	5.0
3	404.845	4.7069	0.0	5.0
4	346.334	3.0953	0.0	6.0
...
5098	306.406	2.7027	2.0	4.0
5099	436.489	4.5050	1.0	7.0
5100	331.441	3.2431	1.0	5.0
5101	447.506	5.1143	1.0	5.0
5102	496.376	5.8682	1.0	4.0

5103 rows × 4 columns

Fig -2 Lipinski Descriptors

3. QSAR MODELING

The development of machine learning algorithms has become an important tool in the process of drug discovery. Quantitative Structure Activity Relationships (QSAR) currently uses a variety of machine learning tools to develop QSAR models. Using machine learning algorithms, 2DQSAR analysis explores quantitative relationships between molecular descriptors and biological activities. Established QSAR models based on machine learning techniques can help to understand the structural requirements needed to develop novel compounds with improved biological activity.

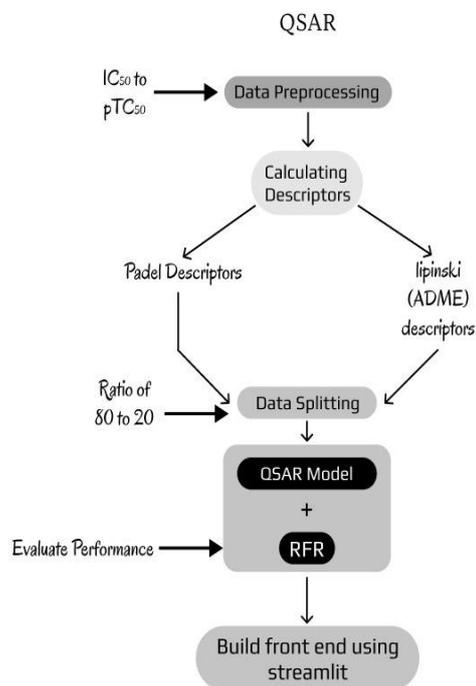


Fig -3 Flowchart of QSAR Model

Above is a flow chart of the research workflow. Briefly, this includes a large-scale QSAR model for the prediction and assessment of AChE inhibition performed in accordance with OECD guidelines.

- (i) A clear method of learning;
- (ii) specific application domains of the QSAR model;
- (iii) Use acceptable measures of quality, sustainability and predictability.
- (iv) Mechanical analysis of the QSAR model.

3.1 Multivariate Analysis

Supervised training is the process of training a model on labeled training data, which can be used to predict unknown or future data. This study builds a regression model that predicts a continuous response variable (i.e., pIC50) as a function of a predictor variable (i.e., fingerprint descriptors).

This study builds a regression model that predicts a continuous response variable (e.g., pIC50) as a function of a predictor variable (ego fingerprint descriptor).

A random forest (RF) classifier is an ensemble classifier made up of multiple decision trees. Simply put, the main idea of RF is that instead of building a deep decision tree with an ever-increasing number of nodes that may be susceptible to data overfitting and overtraining, it creates multiple trees to minimize variance rather than maximize accuracy. As a

result, the result is noisier than the result of a well-trained decision tree, but is generally reliable and robust.

	canonical_smiles	molecule_chembl_id	pIC50
0	<chem>CCOC1nn(-c2cccc(OCc3cccc3)c2)c(=O)o1</chem>	CHEMBL133897	6.124939
1	<chem>O=C(N1CCCC1)n1nc(-c2ccc(Cl)cc2)nc1SCC1CC1</chem>	CHEMBL336398	7.000000
2	<chem>CN(C(=O)n1nc(-c2ccc(Cl)cc2)nc1SCC(F)(F)F)c1cccc1</chem>	CHEMBL131588	4.301030
3	<chem>O=C(N1CCCC1)n1nc(-c2ccc(Cl)cc2)nc1SCC(F)(F)F</chem>	CHEMBL130628	6.522879
4	<chem>CSc1nc(-c2ccc(OC(F)(F)F)cc2)nn1C(=O)N(C)C</chem>	CHEMBL130478	6.096910
...
5098	<chem>CN(C)C(=O)Oc1ccc(C(O)CNC2CCCC2)cc1.Cl</chem>	CHEMBL4645476	3.575118
5099	<chem>COc1ccc(CCC(=O)Nc2nc(-c3cc4cccc4oc3=O)cs2)cc1OC</chem>	CHEMBL4645659	6.130768
5100	<chem>COc1ccc(-c2csc(NC(=O)CCN3CCCC3)n2)cc1</chem>	CHEMBL513063	6.292430
5101	<chem>COc1cc(C2C3=C(CCCC3=O)NC3=C2C(=O)CCC3)ccc1OCc1...</chem>	CHEMBL4640608	3.903090
5102	<chem>O=C1CCCC2=C1C(c1ccc(OCc3cccc(F)c3)c(Br)c1)C1=C...</chem>	CHEMBL4173961	4.000000

Fig -4 Canonical Smiles with pIC50 values and Chembl ID

3.2 QSAR Model Validation

Model validation is a critical step that should be taken to ensure that a fitted model can accurately predict responses for future or unknown subjects.

The performance of the QSAR models was evaluated using four statistical parameters: Pearson's correlation coefficient (r), root mean squared error (RMSE), mean squared error (MSE) and coefficient of determination (r²).

Mean squared error (MSE): 0.35
 Coefficient of determination (R²): 0.87
 Pearsons correlation: 0.93
 RMSE value: 1

Fig -5 Model Evaluation

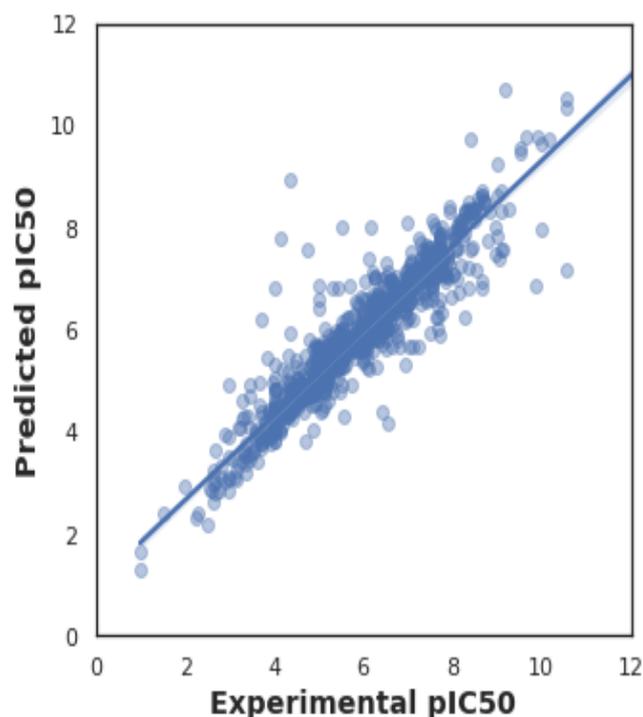


Fig -6 Plot of experimental pIC50 values versus predicted pIC50 values

```
model = RandomForestRegressor(n_estimators=500, random_state=42)
model.fit(X, Y)
r2 = model.score(X, Y)
r2
0.8652930583474272
```

Fig -7 Model Score

4. DEPLOYMENT OF MODEL

The model was then converted to a pickle file for further distribution. An application using various features of Streamlit is created. Users are prompted to upload a text file containing standard Smiles and ChEMBL ids.

Once the prediction is made, the user can download it in .csv file format and use it for further investigation.

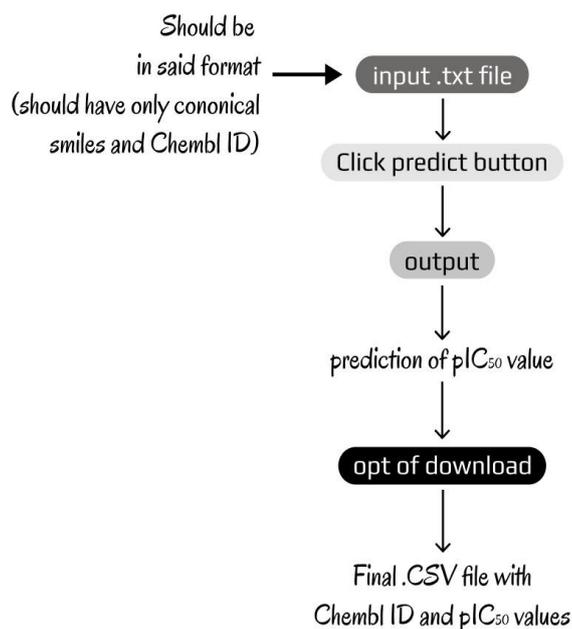
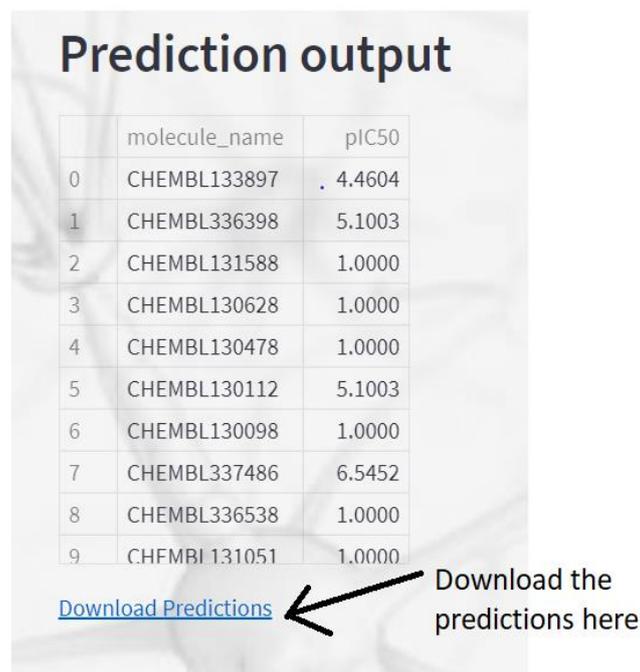


Fig -8 Flowchart of Front-end

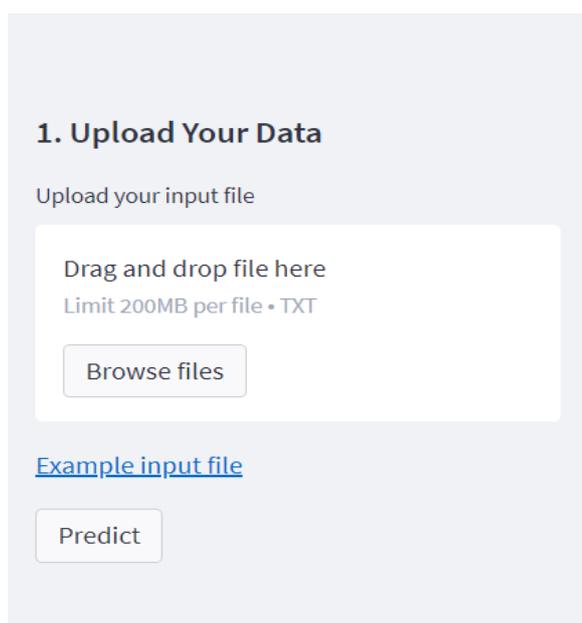


Prediction output

	molecule_name	pIC50
0	CHEMBL133897	4.4604
1	CHEMBL336398	5.1003
2	CHEMBL131588	1.0000
3	CHEMBL130628	1.0000
4	CHEMBL130478	1.0000
5	CHEMBL130112	5.1003
6	CHEMBL130098	1.0000
7	CHEMBL337486	6.5452
8	CHEMBL336538	1.0000
9	CHEMBL131051	1.0000

[Download Predictions](#) ← Download the predictions here

Fig -10 Predictions of pIC50 values



1. Upload Your Data

Upload your input file

Drag and drop file here
Limit 200MB per file • TXT

[Browse files](#)

[Example input file](#)

[Predict](#)

Fig -9 Front-end Sidebar to upload file

5. CONCLUSIONS

The goal of this project is to create applications that help medical researchers in the early stages of drug discovery, where different methods have been studied and carefully integrated.

Finally, we built a QSAR model that can predict pIC50 values for new drugs using fingerprint descriptors. We hope that the results of this study will be a general guideline for the development of novel AChE inhibitors..

REFERENCES

- [1] "How do drugs for Alzheimer's disease work?" *Alzheimer's Society*, 22 December 2021, <https://www.alzheimers.org.uk/about-dementia/treatments/drugs/how-do-drugs-alzheimers-disease-work>. Accessed 1 March 2022.
- [2] Aykul S, Martinez-Hackert E. Determination of half-maximal inhibitory concentration using biosensor-based protein interaction analysis. *Anal Biochem*. 2016 Sep 1;508:97-103. doi: 10.1016/j.ab.2016.06.025. Epub 2016 Jun 27. PMID: 27365221; PMCID: PMC4955526.
- [3] Bosc, N., Atkinson, F., Felix, E. *et al*. Large scale comparison of QSAR and conformal prediction methods and their applications in drug discovery. *J Cheminform* **11**, 4 (2019). <https://doi.org/10.1186/s13321-018-0325-4>

- [4] Yap, Chun Wei. (2011). PaDEL-Descriptor: An Open Source Software to Calculate Molecular Descriptors and Fingerprints. *Journal of computational chemistry*. 32. 1466-74. 10.1002/jcc.21707.
- [5] Kuz'min, Victor & Polishchuk, Pavel & Artemenko, Anatoly & Andronati, Sergey. (2011). Interpretation of QSAR Models Based on Random Forest Methods. *Molecular Informatics*. 30. 10.1002/minf.201000173.
- [6] M. Stitou, H. Toufik, M. Bouachrine, H. Bih and F. Lamchouri, "Machine learning algorithms used in Quantitative structure-activity relationships studies as new approaches in drug discovery," 2019 International Conference on Intelligent Systems and Advanced Computing Sciences (ISACS), 2019, pp. 1-8, doi: 10.1109/ISACS48493.2019.9068917.
- [7] Bhure, Soham, et al. "Drug Generation Using Generative Models." vol. 08, no. 10 | Oct 2021, p. 8, <https://www.irjet.net/archives/V8/i10/IRJET-V8I1097.pdf>.