IN-SILICO APPROACH TO AID DRUG DISCOVERY PROCESS FOR ALZHEIMER'S DISEASE

Project submitted to the

SRM University – AP, Andhra Pradesh

for the partial fulfillment of the requirements to award the degree of

Bachelor of Technology

In

Computer Science and Engineering

School of Engineering and Sciences

Submitted by

Akash Ghosh

(AP21110010057)

Under the Guidance of

Dr. Anuj Deshpande

SRM University–AP

Neerukonda, Mangalagiri, Guntur

Andhra Pradesh – 522240

Nov, 2023

Certificate

Date: 1-Dec-23

This is to certify that the work present in this Project entitled "**In-silico Approach To Aid Drug Discovery Process For Alzheimer's Disease**" has been carried out by **Akash Ghosh** under my/our supervision. The work is genuine, original, and suitable for submission to the SRM University – AP for the award of Bachelor of Technology/Master of Technology in **School of Engineering and Sciences**.

Supervisor

(Signature)

Dr. Anuj Deshpande

Designation,

Affiliation.

Acknowledgements

I express my sincere gratitude to my supervisor, Dr. Anuj Deshpande, who made this work possible. He has walked us through the process of doing the research and effectively communicating the results of the research. I feel incredibly blessed and delighted to do this project with his help. I would also like to thank all the members of my team – Jenil Padshala, Yatharth Tomar, and Karthik Reddy for all the assistance and efforts to finish the project.

Table of Contents

Abstract

This research study aims to leverage computational techniques for data analysis and machine learning models to identify potential drug candidates for Alzheimer's disease (AD) efficiently by predicting the IC_{50} values of the target protein inhibitors. The study primarily focuses on Acetylcholinesterase (AChE) and Butyrylcholinesterase (BChE) as the target proteins for AD. A thorough QSAR (Quantitative structure-activity relationship) study procedure was developed to predict the IC_{50} values for inhibitors of each target protein. This procedure involved collecting bioactivity data of the inhibitors from the ChEMBL database, eliminating redundancy from the dataset, calculating descriptors, and building an efficient machine-learning model to predict the IC_{50} values.

Statement of Contributions

We, Jenil Padshala, Karthik Reddy, Yatharth Tomar and Akash Ghosh, hereby declare that the following is an accurate statement of our contributions to the research study:

Jenil Padshala assumed responsibility for the conceptualization of ideas, conducted the data collection and subsequent analysis, prepared figures and tables, created the ML model, and performed error analysis.

Yatharth Tomar undertook the critical tasks of conceptualization, conducting an extensive literature review, engaged in descriptor calculations, and contributed significantly to error analysis.

Karthik Reddy carried out data collection and analysis, performed descriptor calculations, and made substantive contributions to the drafting and refinement of the research report.

Akash Ghosh contributed to the research study by conducting a comprehensive literature review, crafting and refining the written material, and actively participating in the creation of the ML model.

Each of us has played distinct yet complementary roles, thereby collectively shaping and enriching the research study through our concerted efforts and expertise.

Abbreviations

Ach	Acetylcholine
AChE	Acetylcholinesterase
AD	Alzheimer′s Disease
BChE	Butyrylcholinesterase
ChEMBL	Chemical Biology Database
IC_{50}	Half-maximal Inhibitor Concentration
pIC_{50}	Negative log base 10 of IC_{50}
QSAR	Quantitative Structure Activity Relationship
RMSE	Root Mean Square Error
RSS	Sum of Squares of Residuals
SMILES	Simplified Molecular Input Line Entry System
TSS	Total Sum of Squares

List of Tables

List of Figures

List of Equations

1. Introduction

Alzheimer's disease (AD) is a progressive neurological disorder characterized by the deterioration of memory, thinking skills, and behaviour. As the leading cause of dementia, AD significantly impacts cognitive abilities, hindering daily life. This research delves into the examination of potential therapeutic targets for AD, focusing on the cholinergic hypothesis and utilizing Quantitative Structure-Activity Relationship (QSAR) methodology.

Various hypotheses, such as the cholinergic, amyloid, and tau hypotheses, have been proposed to elucidate the mechanisms underlying Alzheimer's disease. The cholinergic hypothesis posits a significant deficiency in the neurotransmitter Acetylcholine (ACh) in AD. Consequently, inhibiting Acetylcholinesterase (AChE) emerges as a promising treatment strategy. AChE is an enzyme responsible for the hydrolysis of acetylcholine into acetic acid and choline [1]. Additionally, Butyrylcholinesterase (BChE), another enzyme catalysing the hydrolysis of esters of choline, including acetylcholine, is considered [2].

For this QSAR study, Acetylcholinesterase (AChE) and Butyrylcholinesterase (BChE) were selected as the target proteins associated with Alzheimer's disease. The research employed a systematic procedure, as illustrated in Figure 1.

2. Methodology

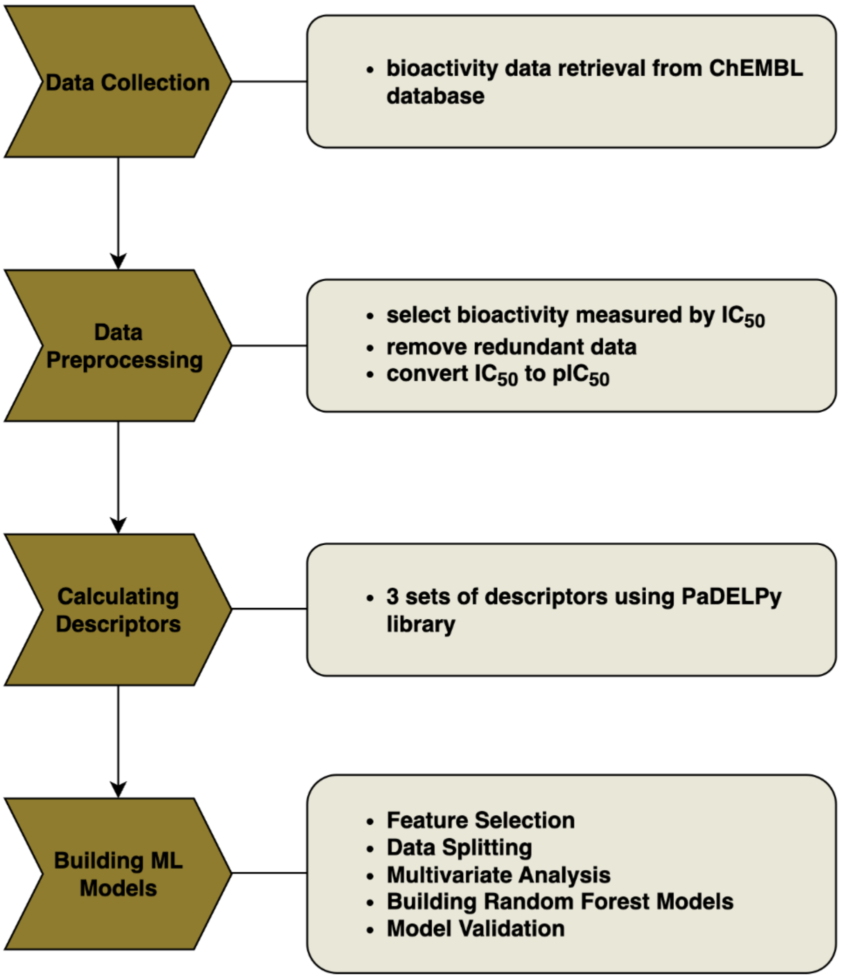

Figure 1. Workflow of the study

2.1 Data set collection and pre-processing

The study focused on inhibitors of human AChE, a specific enzyme, and human BChE. To gather information, a data set was collected from the ChEMBL online database [3], specifically for human AChE (id: CHEMBL220) and human BChE (id: CHEMBL1914).

For human AChE, the initial data set included 16,879 records of AChE inhibitors. These records had various bioactivity measurements, such as IC_{50}, EC_{30}, % inhibition, % reactivation, etc. However, our focus was on the IC_{50} values, so the data set was filtered to include only 8,832 compounds with available IC_{50} values.

During this process, it was noticed that 1,285 compounds either lacked recorded IC_{50} values or had missing Simplified Molecular Input Line Entry System (SMILES) notations [4], which are molecular structure representations. Consequently, these compounds were excluded from the data set. In addition, 1,390 redundant compounds were deleted due to duplicate SMILES notations. The final data set for human AChE consisted of 6,157 compounds, each with a canonical SMILES notation and IC_{50} values measured in nanomolar (nM) units.

A similar curation process was applied to the data set for human BChE. Initially, there were 3,772 compound records. After eliminating 947 redundant records, the final data set for human BChE comprised 3,772 compounds, each with a canonical SMILES notation and IC_{50} values in nanomolar (nM) units.

These experimental IC_{50} measurements of compounds vary across a wide range of values. To enhance interpretability and facilitate statistical analysis, the data is simplified by converting IC_{50} values to pIC_{50} values [10].

$$pIC_{50} = -1 \times \log_{10}(IC_{50} \times 10^{-9}) \tag{1}$$

2.2 Calculating Descriptors

Molecular descriptors serve as mathematical representations depicts the properties of a given molecule. They emerge through a systematic application of mathematical and logical procedures, effectively converting chemical information into numerical values or reflecting the outcomes of experimental processes. Despite their widespread use, the efficacy of these descriptors in facilitating bioactivity calculations and modeling can exhibit considerable variation. Numerous research studies have been conducted to scrutinize and evaluate the performance discrepancies and accuracy levels associated with these descriptors [5,6].

This study specifically explores into the assessment of three molecular descriptors concerning their suitability in predicting the bioactivity of inhibitors targeting acetylcholinesterase (AChE) and butyrylcholinesterase (BChE). The details of these descriptors, along with pertinent references, are comprehensively presented in Table 1 [7,8].

Table 1: Molecular Descriptors with their size, description and references

Sr No.	Descriptor	Size	Description	Reference
1	CDK Extended	1024	Fingerprints of length 1024 and search depth of 8 with additional bits describing ring features	[7, 8]
2	PubChem	881	Binary representation of substructures defined by PubChem	[9]
3	Substructure	307	Presence of SMARTS patterns for functional groups	[6]

2.3 Building Machine Learning Models

2.3.1 Feature Selection

Collinearity refers to the situation where pairs of descriptors exhibit intercorrelation, introducing complexity to the model and potentially leading to biased outcomes. To prevent this, the *VarianceThreshold* function in the sklearn python library was utilised to obtain features with a variance greater than the variance threshold of 0.1.

Variance of feature can be calculated by the following formula:

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - x_{mean})^2 \qquad (2)$$

2.3.2 Data Splitting

To avoid the possibility of bias that may arise from building regression models from a single dataset, the data set was split into Train and Test sets in which the former contained 80% of the initial data and the latter constituted 20% of the initial data.

2.3.3 Multivariate Analysis

In supervised learning, the algorithm learns from labelled training data. It involves training a model on input-output pairs, where the algorithm tries to learn the mapping between the input data and the corresponding output labels [11].

This study required the construction of regression models, which allow to prediction of a continuous value from infinitely many possible values (here, pIC_{50}).

First, we utilized the *LazyPredict* library in Python to quickly evaluate multiple machine-learning models without extensive manual setup. The results are shown in figure 2 in the Discussion section.

2.3.4 Building ML Models

Based on the initial evaluation, the decision was made to build Random Forest Models for predicting pIC_{50} values as it showed consistent performance for all the descriptors.

Random forest (RF) [12] is a robust and versatile ensemble classifier that leverages the collective wisdom of multiple decision trees to enhance predictive accuracy and minimize overfitting. Unlike a single decision tree that may become overly complex and susceptible to overtraining, RF employs a collection of decision trees, each constructed using a random subset of the training data. This strategy, known as bagging (bootstrap aggregating), introduces diversity among the trees, effectively averaging out individual biases and reducing the overall variance of the ensemble. While the individual trees may be noisier than a well-trained single tree, the collective output of the ensemble is typically more reliable and robust. This approach not only improves generalization performance but also enhances the ensemble's ability to handle high-dimensional and noisy data.

The *RandomForestRegressor* class provided in the sklearn python library was used for constructing the models.

2.3.5 Model Validation

Two statistical parameters were used to evaluate the accuracy of the models: Coefficient of determination (r^2) and Root Mean Square Error (RMSE).

Formula to calculate the r^2 value:

$$R^2 = 1 - \frac{RSS}{TSS} \tag{3}$$

$$RSS = \sum_{i=1}^{n} \left(y_i - f(x_i)\right)^2 \tag{4}$$

$$TSS = \sum_{i=1}^{n} (y_i - y_{mean})^2 \tag{5}$$

Formula to calculate the RMSE value:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (y_i - y_{i\,predicted})^2}{n}} \tag{6}$$

Table 2 and Table 3 show the parameter values for each descriptor set for AChE and BChE respectively.

Table 2: Error and accuracy analysis of Random Forest Models for AChE inhibitor descriptors.

Sr No.	**Descriptor**	**No. of Features**	**R^2**	**RMSE**
1	CDK Extended	931	0.75	0.62
2	PubChem	216	0.72	0.71
3	Substructure	24	0.51	1.22

Table 3: Error and accuracy analysis of Random Forest Models for BChE inhibitor descriptors

Sr No.	**Descriptor**	**No. of Features**	**R^2**	**RMSE**
1	CDK Extended	940	0.75	0.61
2	PubChem	215	0.70	0.80
3	Substructure	23	0.48	1.35

3. Discussion

The bar graphs in Figure 2 depict the accuracy of various regression models for predicting the pIC_{50} value of AChE inhibitors based on different molecular descriptors using the *LazyPredict* Python library. It was observed that the *Random Forest Regressor* performed consistently well for all the molecular descriptors.

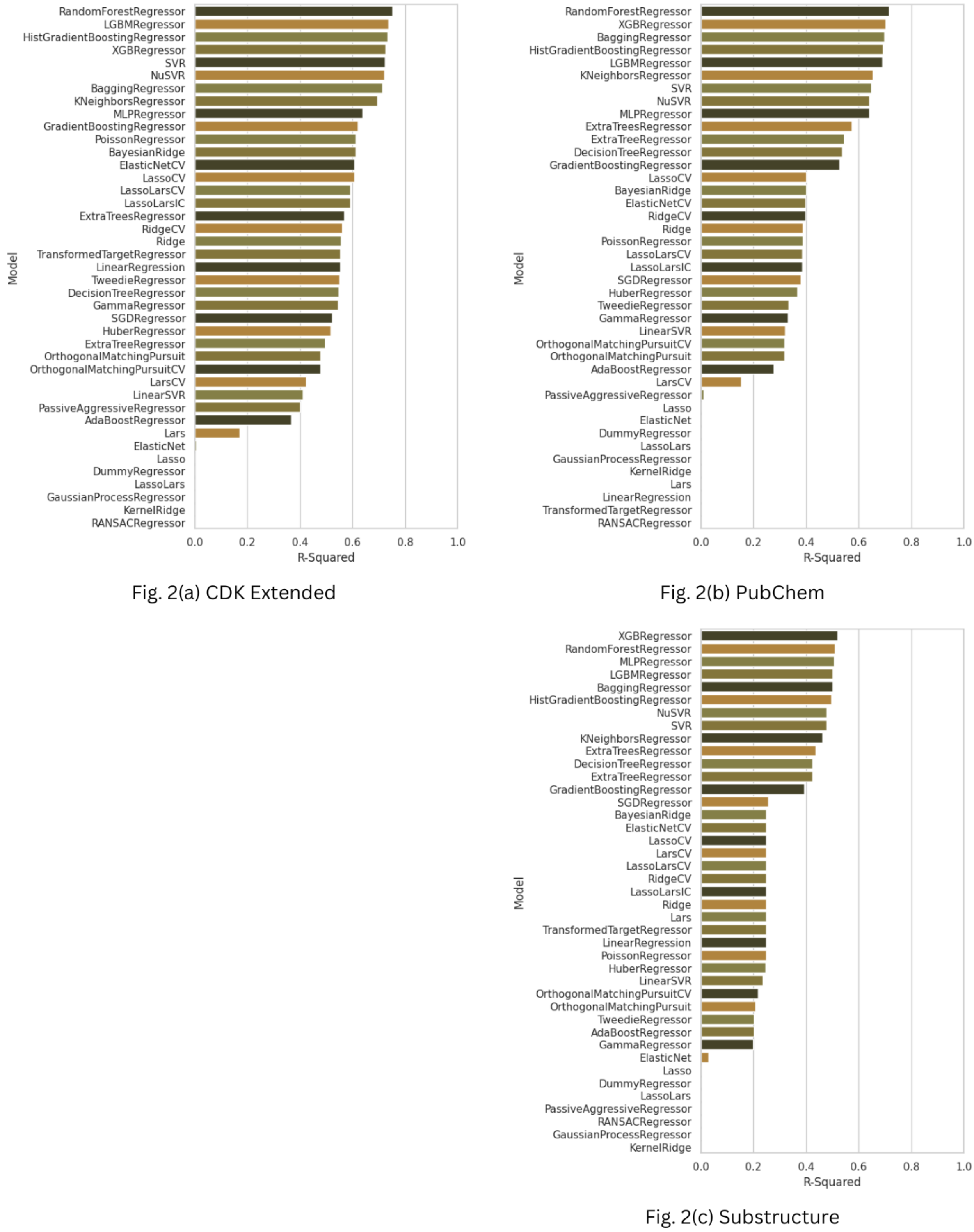

Fig. 2(a) CDK Extended

Fig. 2(b) PubChem

Fig. 2(c) Substructure

Figure 2. Quick evaluation of regression models for AChE inhibitors based on different molecular descriptors.

Similarly, the below bar graphs in figure 3 depict the accuracy of various regression models for predicting pIC$_{50}$ value of BChE inhibitors based on different molecular descriptors.

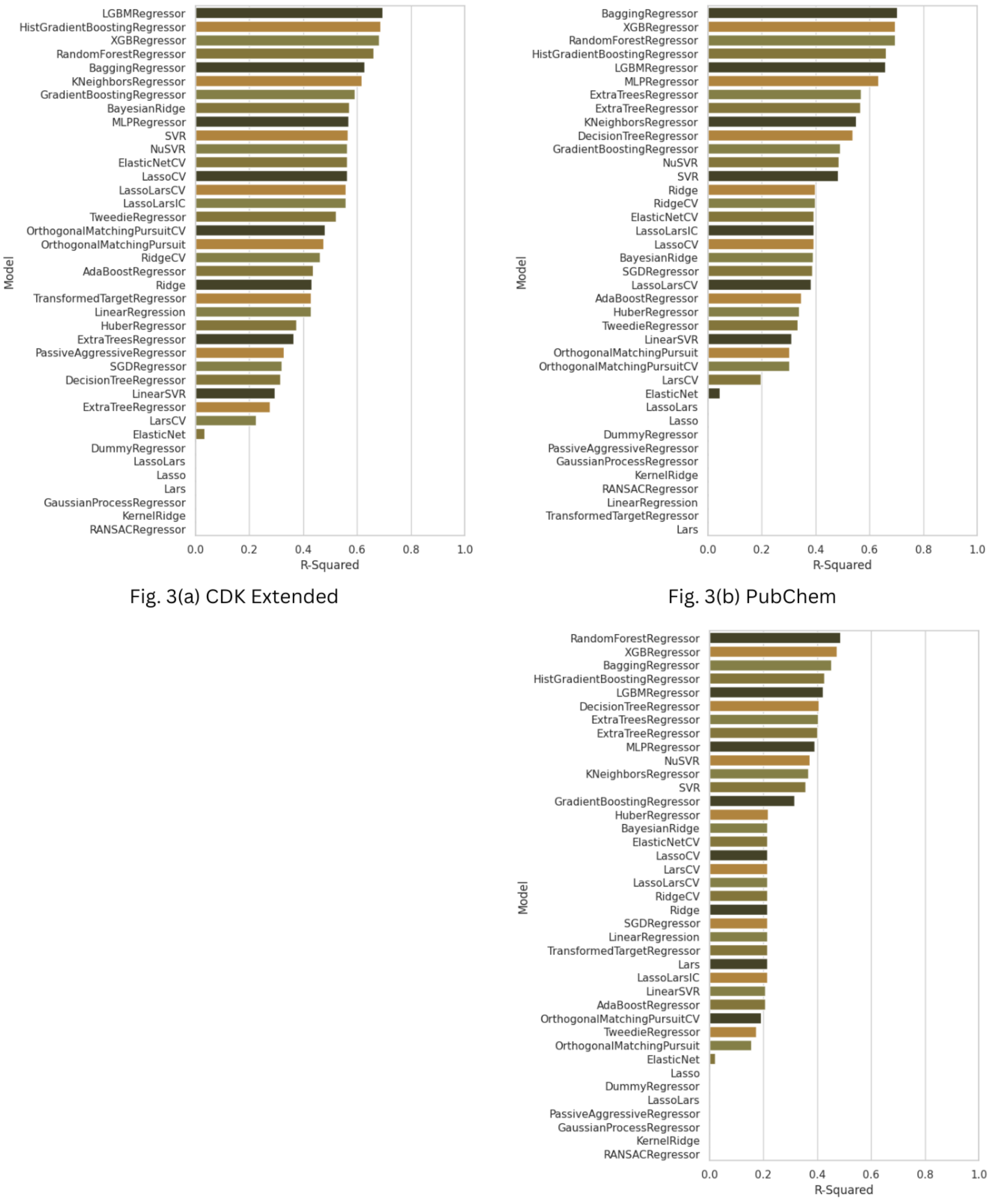

Fig. 3(a) CDK Extended

Fig. 3(b) PubChem

Fig. 3(c) Substructure

Figure 3. Quick evaluation of regression models for BChE inhibitors based on different molecular descriptors.

4. Concluding Remarks

In conclusion, this research establishes a robust computational framework utilizing machine learning techniques and data analysis to predict inhibitor activity against Acetylcholinesterase (AChE) and Butyrylcholinesterase (BChE), pivotal enzymes in Alzheimer's disease. The study successfully curated and processed data from the ChEMBL database, optimizing it for predictive modelling. While employing diverse molecular descriptors and implementing Random Forest regression models, the general trend observed in its performance indicated that the accuracy followed the order of CDK Extended > PubChem > Substructure descriptors. This observation emphasizes the significance of descriptor selection in influencing the predictive accuracy of the models, highlighting the potential impact of feature representation on the performance of machine learning algorithms in drug discovery processes.

5. Future Work

Future research endeavours could involve the exploration of a wider range of target proteins, the diversification of descriptor analyses, and the expansion of this approach to encompass studies across various diseases. Such extensions could support drug repurposing initiatives and provide valuable insights for researchers aiming to explore novel therapeutic strategies for different medical conditions.

References

[1] Rees, Tina M., and Stephen Brimijoin. "The role of acetylcholinesterase in the pathogenesis of Alzheimer's disease." Drugs of today (Barcelona, Spain: 1998) 39, no. 1 (2003): 75-83. doi: 10.1358/dot.2003.39.1.740206. PMID: 12669110.

[2] Darvesh, Sultan, David A. Hopkins, and Changiz Geula. "Neurobiology of butyrylcholinesterase." Nature Reviews Neuroscience 4, no. 2 (2003): 131-138. https://doi.org/10.1038/nrn1035

[3] Gaulton, Anna, Louisa J. Bellis, A. Patricia Bento, Jon Chambers, Mark Davies, Anne Hersey, Yvonne Light et al. "ChEMBL: a large-scale bioactivity database for drug discovery." Nucleic acids research 40, no. D1 (2012): D1100-D1107. doi: 10.1093/nar/gkr777. Epub 2011 Sep 23. PMID: 21948594; PMCID: PMC3245175.

[4] Weininger, David. "SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules." Journal of chemical information and computer sciences 28, no. 1 (1988): 31-36. https://doi.org/10.1021/ci00057a005

[5] Kim, Sunghwan, Evan E. Bolton, and Stephen H. Bryant. "PubChem3D: conformer ensemble accuracy." Journal of Cheminformatics 5 (2013): 1-17. https://doi.org/10.1186/1758-2946-5-1

[6] Simeon, Saw, Nuttapat Anuwongcharoen, Watshara Shoombuatong, Aijaz Ahmad Malik, Virapong Prachayasittikul, Jarl ES Wikberg, and Chanin Nantasenamat. "Probing the origins of human acetylcholinesterase inhibition via QSAR modeling and molecular docking." PeerJ 4 (2016): e2322.doi: 10.7717/peerj.2322. PMID: 27602288; PMCID: PMC4991866.

[7] Willighagen, Egon L., John W. Mayfield, Jonathan Alvarsson, Arvid Berg, Lars Carlsson, Nina Jeliazkova, Stefan Kuhn et al. "The Chemistry Development Kit (CDK) v2. 0: atom typing, depiction, molecular formulas, and substructure searching." Journal of cheminformatics 9 (2017): 1-19. https://doi.org/10.1186/s13321-017-0220-4

[8] Steinbeck, Christoph, Yongquan Han, Stefan Kuhn, Oliver Horlacher, Edgar Luttmann, and Egon Willighagen. "The Chemistry Development Kit (CDK): An open-source Java library for chemo-and bioinformatics." Journal of chemical information and computer sciences 43, no. 2 (2003): 493-500. DOI: 10.1021/ci025584y

[9] Kim, Sunghwan. "Exploring chemical information in PubChem." Current protocols 1, no. 8 (2021): e217.

[10] Selvaraj, Chandrabose, Sunil Kumar Tripathi, Karnati Konda Reddy, and Sanjeev Kumar Singh. "Tool development for Prediction of pIC50 values from the IC50 values-A pIC50 value calculator." Current Trends in Biotechnology and Pharmacy 5, no. 2 (2011): 1104-1109.

[11] Nasteski, Vladimir. "An overview of the supervised machine learning methods." Horizons. b 4 (2017): 51-62.

[12] Breiman, Leo. "Random forests." Machine learning 45 (2001): 5-32. https://doi.org/10.1023/A:1010933404324