PROJECT TITLE: Developing a machine learning model for any other drug-like properties such as toxicity, permeability, BBB permeability, Oral bioavailability, solubility

AIM: To predict pIC50 values for CDK5 inhibitors in Alzheimer's and Parkinson's Disease using Machine Learning

1. INTRODUCTION

Neurodegenerative disorders include a variety of ailments that cause nerve cells to die or degenerate, which can cause problems with mental functioning (dementia) or movement (ataxia). Cognitive faculties like memory and reasoning deteriorate with dementia, which eventually affects day-to-day functioning.

Some neurodegenerative diseases are as follows:

1. Wernicke-Korsakoff syndrome
2. Mixed dementia
3. Alzheimer's disease
4. Parkinson's illness
5. Normal-pressure hydrocephalus
6. The disease Creutzfeldt-Jakob
7. Huntington's disease
8. Lewy body dementia
9. Vascular dementia
10. Mild cognitive impairment, and
11. Frontotemporal dementia

Alzheimer's disease (AD) is the most common neurological condition. It is a relentless robber of brain function that causes a gradual but noticeable decline. The hallmarks of this decline include the gradual buildup of plaques and tangles in the brain, the loss of essential synapses—connections between nerve cells—and ongoing inflammation. Alzheimer's disease symptoms are caused by damage to brain tissue that prevents messages from being sent through chemical transmitters. The temporal, parietal, cingulate, and frontal cortex are among the affected areas that usually exhibit gross atrophy as a result of early symptoms, which typically include a progressive deterioration in cognitive function accompanied by short-term memory loss, neuronal loss, and degeneration in cerebral cortex synapses and specific subcortical regions.

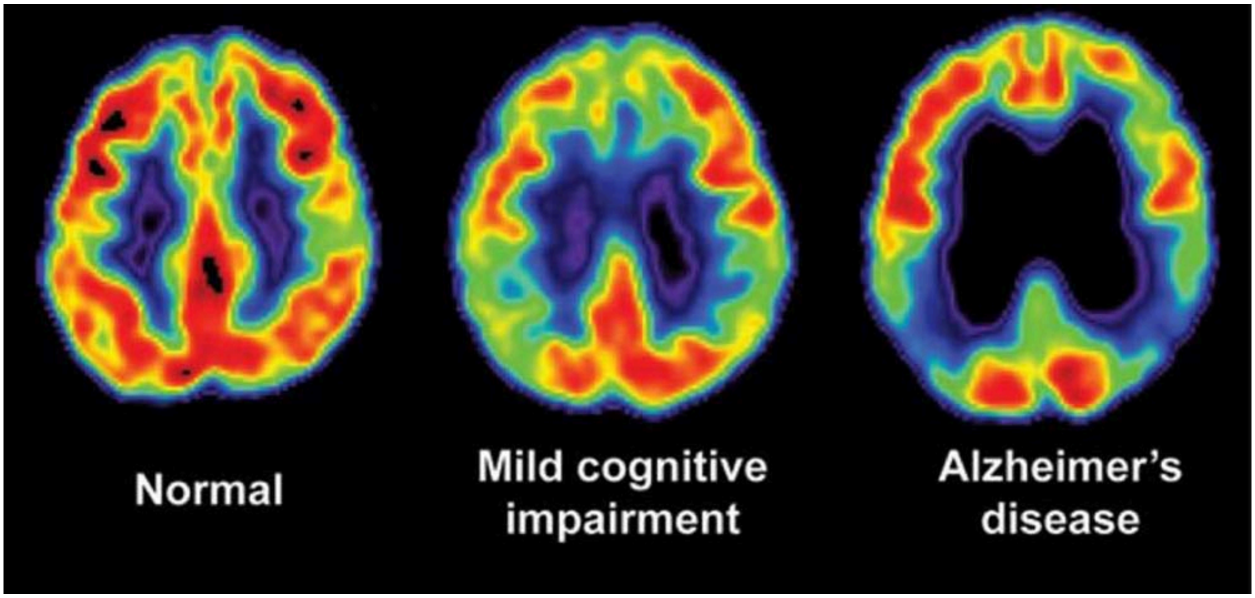


Figure 1. Brain Imaging for Alzheimer's and Dementia

Selected Target for our Study

After studying different research papers, we selected **Cyclin-dependent Kinase-5 (CDK-5)** inhibitors or targets for our project.

1.1 Cyclin dependent Kinase-5 (CDK5)

The multifunctional kinase Cdk5 is necessary for neuron survival as well as apoptosis. In psychological situations, it requires its activator, p35, to carry out a number of functions, such as learning, memory consolidation, and cerebellar development, in addition to neuronal processes, including migration, differentiation, and neurotransmission. However, elevated calcium levels in neurotoxic settings can result in calpain-mediated p35 cleavage into p25 and p10. The resulting p25 is more stable but neurotoxic and may activate Cdk5 on its own. This results in a stable p25/Cdk5 complex that sets off pathological processes, such as tau phosphorylation, Aβ accumulation, dysfunctional synapses, apoptosis, cell cycle reactivation, and oxidative stress, which ultimately cause neurons to die.

1.2 Role of Cdk5 in Alzheimer's and Parkinson's Disease

A serine/threonine kinase called Cdk5, one of its activators in the nervous system, is essential for synaptic functioning, synapse formation, and neuronal migration. Additionally, Cdk5 is linked to the regulation of neuronal survival in both development and disease; either excessive or insufficient Cdk5 activity reduces neuronal survival. Regarding the role of Cdk5 in the pathophysiology of Alzheimer's disease, significant advancements have been made. Notwithstanding early debate regarding the finding of p25 and increased Cdk5 activity in postmortem samples of Alzheimer's patients, further research employing animal models of the illness mainly demonstrates that Cdk5 dysregulation plays a role in the disease's neuronal death.

Overactivation of Cdk5 has been shown to cause degeneration of dopaminergic neurons in the SNpc by translocating SIRT2 from the cytoplasm to the nucleus.30 Furthermore, Cdk5 has the ability to phosphorylate Raf-kinase inhibitor protein (RKIP), a negative regulator of the mitogen-activated protein kinase (MAPK) pathway that causes autophagy to recognize RKIP.31 Transgenic PD mice and brain tissues from PD patients demonstrated that Cdk5 activation results in the degeneration of dopaminergic neurons in the SNpc by activating the MAPK pathway.

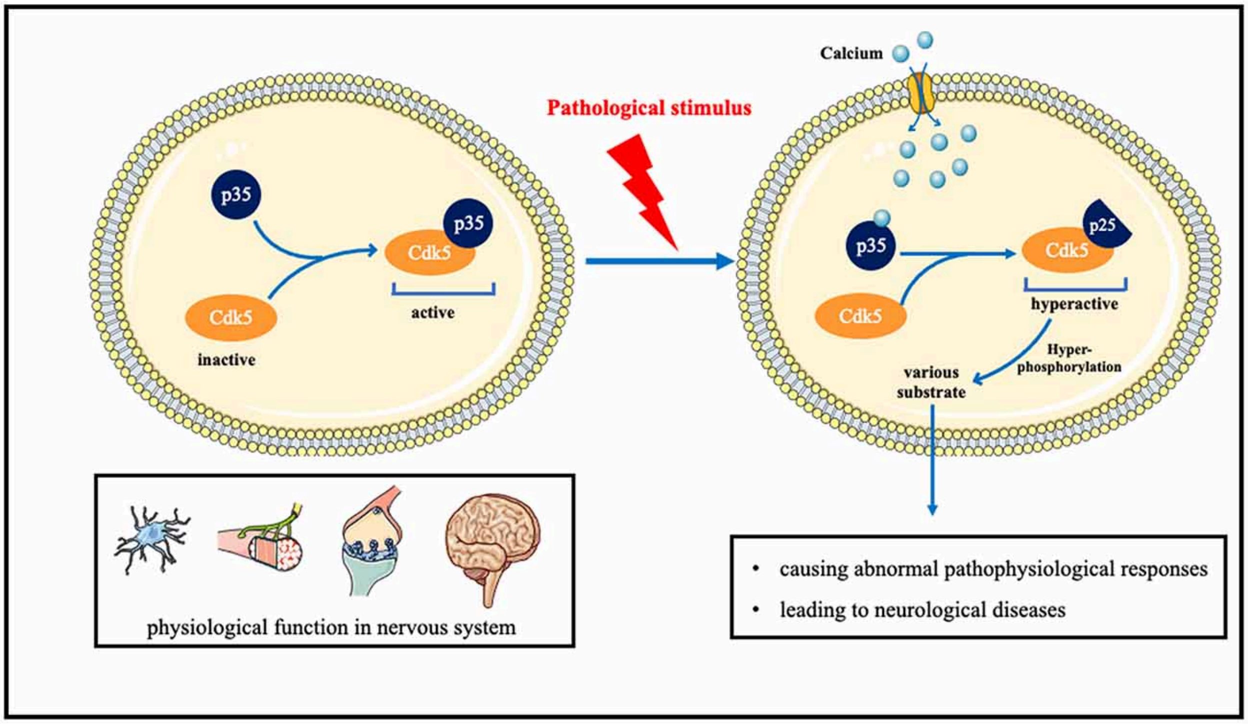
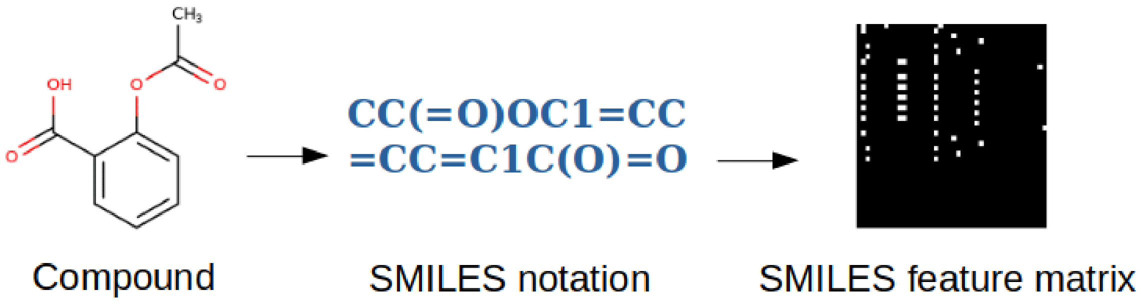

Calcium
Pathological stimulus
p35
Cdk5
p35
active
Cdk5
inactive
p35
p25
Cdk5
hyperactive
Cdk5
Hyper-
phosphorylation
various
substrate
physiological function in nervous system
• causing abnormal pathophysiological responses
• leading to neurological diseases

Figure 2. The role of Cdk5 in neurological disorders and the underlying molecular mechanisms.

2. METHODOLOGY

2.1 Data Collection and SMILES generation

SMILES (Simplified Molecular Input Line Entry System) serve as a method to convert the three-dimensional configuration of a chemical compound into a simple string of symbols, facilitating easy interpretation by computer software.
SMILES are generated by using the ChEMBL web resource.



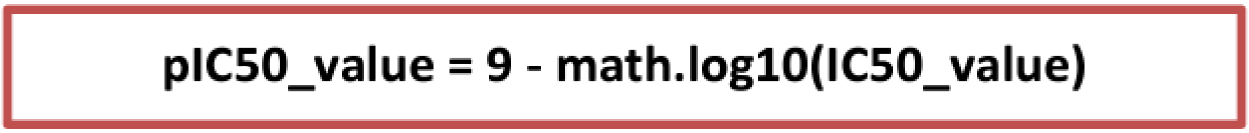
2.2 IC50 and pIC50

IC50: The IC50 is a drug concentration required to inhibit 50% of a specific biological activity, and the test conditions affect its value. IC90 or IC99 might be employed when total inhibition is required. Assuming equilibrium binding at a single site with a Hill coefficient of 1, fractional occupancy calculations show that the IC90 concentration is approximately ten times greater than the IC50 concentration. Likewise, the concentration of IC99 is roughly 100 times higher than that of IC50.

pIC50: The pIC50 factor, or negative logarithm of the half maximum inhibitory concentration (IC50), is critical in medical assessment

Converting IC50 to pIC50

The predicting model is made based on the pIC50 value, so we need to convert the IC50 value to pIC50 using the formula,



```
pIC50_value = 9 - math.log10(IC50_value)
```

2.3 Generating Molecular Descriptors

Molecular descriptors are essential in studying chemical compounds, enabling scientists to gain insights into their properties, predict their behaviour, and accelerate drug discovery and development.

The descriptors are generated by converting the SMILES into Isomeric SMILES, and then all descriptors are generated using Rdkit Mordred.

2.4 Feature Selection

The data will be processed before model training.

The steps are as follows:

Importing the libraries: In this step, we import all the necessary Python libraries like NumPy, Pandas, Matplotlib

Importing the Dataset: In this step, the descriptor dataset of different targets for pre-processing is imported.
Dropping unnecessary column: Dropping the column that has no use, like duplicates, character values, etc.   NAN value handling
Splitting the dataset into train and test data, i.e., 80:20

2.5 Algorithms Used

1. LazyPredict

LazyPredict is a Python library designed to help users quickly compare multiple machine-learning models on a given dataset with minimal code. It simplifies the process of model selection by automatically training and evaluating various algorithms, enabling users to identify the best-performing models for their data without manually coding for each one.

2. Random Forest Regressor (RF)

RF or Random forest machine learning technique that combines the strengths of many decision trees. Imagine a forest where each tree is a simple decision-making process. Random forests work by creating a large number of these decision trees, each trained on a slightly different subset of the data. When a new piece of data comes along, it's passed through all the trees in the forest. Each tree votes for the most likely category, and the final prediction is based on the most popular vote. Random forests exhibit notable strengths such as robustness in handling outliers, resilience against overfitting and noise, and high classification accuracy. Consequently, they have emerged as a favoured research approach within the realms of data mining and have found extensive application in diverse fields, including biological research.

3. Support Vector Regressor (SVR)

One type of support vector machine is support vector regression (SVR) variation explicitly designed for regression applications. Its main goal is to find a function that can reliably anticipate continuous output values from corresponding input values. Both linear and non-linear kernels may be used with SVR. A linear kernel only does a dot product operation between two input vectors, but a non-linear kernel employs a more intricate function that can spot minute patterns in the data. The specifics of the dataset and the complexity of the task will determine whether to use a linear or non-linear kernel.

2.6 Assessing Performance: Scoring Metrics for Regression Models

Mean Squared Error (MSE): MSE is anticipated, and actual values of average squared difference are often assessed using regression analysis. MSE is a commonly used metric for this purpose. It is a crucial indicator of the overall accuracy of the regression model, with lower MSE values indicating superior performance. Notably, the MSE continuously yields values that are either positive or zero, highlighting the squared nature of the error component. Because the Mean

Squared Error (MSE) computes the average of errors, the existence of outliers can dramatically distort the metric's assessment, making it less reliable. Root Mean Squared Error: A commonly used measure for evaluating the differences between values obtained from a sample of the population and the projected values generated by a model or estimator is the Root Mean Squared Error (RMSE). RMSE differs from MSE because it provides mistakes in squared units. It is computed as the square root of the Mean Squared Error (MSE). The range of RMSE values is 0 to positive infinity, which indicates how far the actual and anticipated values deviate from each other.

3. RESULTS AND DISCUSSION

3.1. Machine Learning Model Generation, Optimization and Evaluation

In order to develop ML models for identifying novel ligands potentially active against Cdk5, we searched for compounds with bioactivity data related to Cdk5 inhibition available on ChEMBL 25. Compounds with biological activity measured as IC50 were retrieved and subjected to a data curation process, eventually obtaining the training set used to generate our ML models. Then, the IC50 values were converted to PIC50 values.

To identify the best-performing models for predicting pIC50 values, we initially employed the LazyPredict library, which provided a comparative analysis of multiple regression algorithms. Among the evaluated models, Random Forest and Support Vector Regression (SVR) demonstrated the highest potential based on their performance metrics. To further refine the model selection, we conducted a Grid Search with Cross-Validation (Grid Search CV) on the hyperparameters of the shortlisted models. The results revealed that SVR achieved the highest accuracy, with an R^2 score of 0.65 on the test set. This indicates that SVR is the most suitable model for our dataset, providing the best balance between bias and variance.

3.2 Scatter Plot for CDK5: Predicted Values vs True Values

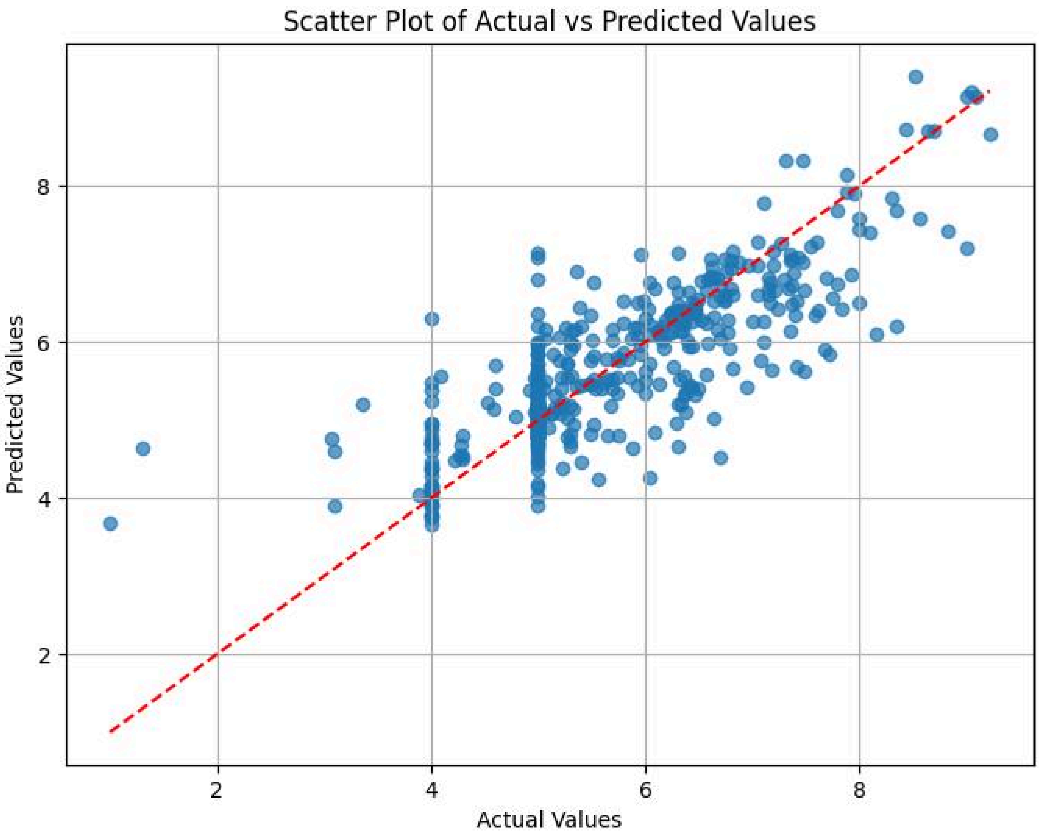


Figure 3. Scatter Plot for CDK5: Predicted Values vs True Values

The clustering of points close to the diagonal line indicates the model's performance, and deviations from this line show prediction errors.
From this, we can conclude:

- Tight clustering near the line: Indicates good prediction accuracy for many data points.
- Spread of points: Suggests some variance in predictions, which could be attributed to the dataset's complexity or model limitations.

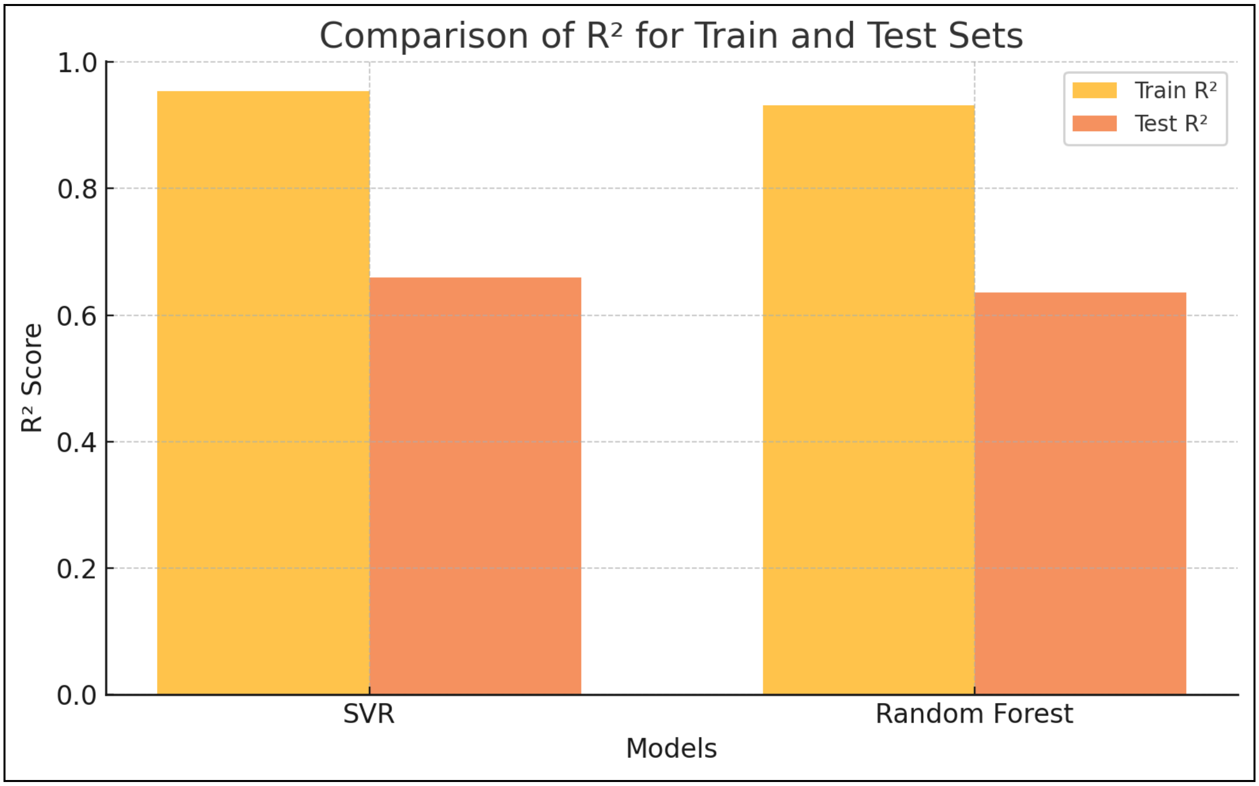


Figure 4. R² Comparison for train and test sets

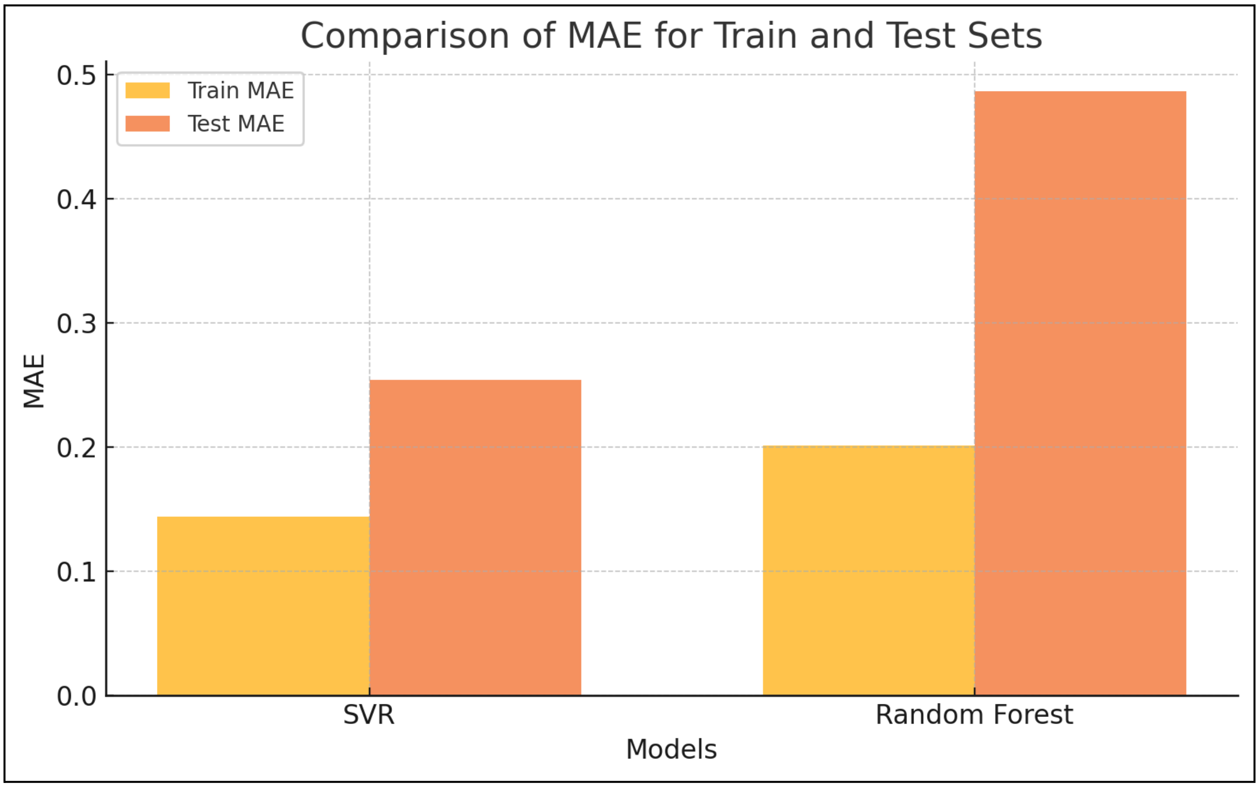


Figure 5. MAE Comparison for train and test sets

Here are the comparison plots:
R^2 Comparison: This bar chart shows the R^2 scores for the train and test sets of both SVR and Random Forest models. SVR achieved better test R^2 than Random Forest, demonstrating slightly better generalisation performance.
MAE Comparison: This bar chart compares the Mean Absolute Error (MAE) for the train and test sets. SVR shows lower MAE values, indicating better predictions compared to Random Forest.

4. CONCLUSION AND FUTURE SCOPE

Cyclin-dependent kinase 5 (Cdk5) is considered a promising target in the drug design field for its role in the progression of neurodegenerative diseases such as AD and PD, as well as in the development and progression of a variety of tumours. In this work, we employed a machine learning-based protocol with the aim of identifying novel potential Cdk5 inhibitors. Among 24 different machine learning models herein developed for this purpose, the best one in terms of MCC and Precision was used to filter two focused libraries of commercial compounds. The prediction is made on the DrugBank and Zinc Natural Product, and we get the five drugs from DrugBank, which have the pIC50 value above seven and show good binding affinity against all our targets after performing the molecular docking.

4.1 Future Scope

- The model can be used to predict pIC50 values.
- Since the dataset was small, we couldn’t achieve higher accuracy. Thus, more data can be collected, modelled and fine-tuned.
- Other inhibitors can be used for ML model building.

5. LIMITATIONS

The dataset used for this study contained only a little over 2,000 entries, which limits the amount of data available for training and testing the machine learning models. This relatively small dataset size may have impacted the model's ability to generalize effectively, resulting in an R^2 score of 0.65. With a larger dataset, it is likely that the model's performance could improve, yielding higher accuracy and better predictive capabilities. Future work should focus on collecting a more extensive dataset to address this limitation and enhance the robustness of the model.

6. REFERENCES

1. Rami, L., Sala-Llonch, R., Solé-Padullés, C., Fortea, J., Olives, J., Lladó, A., Peña-Gómez, C., Balasa, M., Bosch, B., Antonell, A., Sanchez-Valle, R., Bartrés-Faz, D., & Molinuevo, J. L. (2012). Distinct functional activity of the precuneus and posterior cingulate cortex during encoding in the preclinical stage of Alzheimer's disease. *Journal of Alzheimer S Disease*, *31*(3), 517–526. https://doi.org/10.3233/jad-2012-120223

2. Arosio, B., Mastronardi, L., Vergani, C., & Annoni, G. (2010). Intereleukin-10 promoter polymorphism in mild cognitive impairment and in its clinical evolution. *International Journal of Alzheimer S Disease*, *2010*, 1–5. https://doi.org/10.4061/2010/854527

3. Zhan, X., Cox, C., Ander, B. P., Liu, D., Stamova, B., Jin, L., Jickling, G. C., & Sharp, F. R. (2015). Inflammation Combined with Ischemia Produces Myelin Injury and Plaque-Like Aggregates of Myelin, Amyloid-β and AβPP in Adult Rat Brain. *Journal of Alzheimer S Disease*, *46*(2), 507–523. https://doi.org/10.3233/jad-143072

4. Lin, J., Yu, J., Zhao, J., Zhang, K., Zheng, J., Wang, J., Huang, C., Zhang, J., Yan, X., Gerwick, W. H., Wang, Q., Cui, W., & He, S. (2017). Fucoxanthin, a marine carotenoid, attenuates Β‑Amyloid Oligomer‑Induced neurotoxicity possibly via regulating the PI3K/AKT and the ERK pathways in SH‑SY5Y cells. *Oxidative Medicine and Cellular Longevity*, *2017*(1). https://doi.org/10.1155/2017/6792543

5. Lenjisa J.L., Tadesse S., Khair N.Z., Kumarasiri M., Yu M., Albrecht H., Milne R., Wang S. CDK5 in oncology: Recent advances and future prospects. Future Med. Chem. 2017;9:1939–1962. doi: 10.4155/fmc-2017-0097.