Assessment of Data-Driven and Deterministic Approaches for Predicting Bace-1 Inhibitor Binding Affinities

by
Amaithi Priya S

Under the supervision of
Dr.Arul Murugan

Submitted in partial fulfillment of the requirements for the degree of Master of Technology, Computational Biology

Center for Computational Biology Indraprastha Institute of Information Technology - Delhi
June, 2022

Certificate

This is to certify that the thesis titled *"Assessment of Data-Driven and Deterministic Approaches for Predicting Bace-1 Inhibitor Binding Affinities "* being submitted by **Amaithi Priya S** to the Indraprastha Institute of Information Technology Delhi, for the award of the Master of Technology, is an original research work carried out by him under my supervision. In my opinion, the thesis has reached the standards fulfilling the requirements of the regulations relating to the degree.

The results contained in this thesis have not been submitted in part or full to any other university or institute for the award of any degree/diploma.

June,2022

Dr Arul Murugan

Department of Computational Biology
Indraprastha Institute of Information Technology Delhi
New Delhi 110 020

Acknowledgements

I would like to thank Prof. Arul Murugan for his continuous guidance and mentoring, without him this work would not have been possible. From day one of starting my thesis, he constantly pushed my limits to aim for higher goals in life and to always do better at my work. Whenever I faced any issue or need some help in clarifying my doubts, he used to give his time even on days when he is busy. At times, when I was struck at my work and I approached him for some suggestions, he used to give some insights and would refine our approach to doing the work, and that has surely helped me to keep track and complete my work timely.

I thank my mother, brother and sister, who have helped me indirectly in innumerable ways during my work so that I can focus on my thesis completely. They have continuously supported me in my studies right from my under-graduation to graduation. I owe my deepest gratitude to them for this everlasting support.

During my thesis work, for every small or trivial doubt regarding my thesis work or career advice, My fellow batch mates were always there to help me and clarify any doubts related to biology and have always encouraged and supported me in my academic as well as personal life.I thank all of them.

I also thank all the faculty members and staff of the Department of Computational Biology and IIIT Delhi for always helping us throughout our college journey. My special thanks to Mr.Adarsh from the IT department, for his continuous support and help in providing access to the college IT infrastructure.

Abstract

BACE1 (beta-site amyloid precursor protein cleaving enzyme 1) is an aspartyl protease with a transmembrane domain that cleaves at the 671st position of amyloid precursor protein (APP) at the beta site. The production and release of amyloid-peptide, a pathological feature of Alzheimer's disease (AD), is caused by the cleavage of APP by beta-secretase and then by the gamma-secretase complex. BACE1 inhibitors have shown significant promise in reducing amyloid-beta load in the brain and preventing the progression of Alzheimer's disease. The BACE1 inhibitor binding affinities (IC50) reported in chEMBL for various chemical classes of human-beta secretase-1 inhibitors were computed using data-driven and deterministic approaches. Models for predicting binding affinities were developed using quantitative techniques, along with qualitative models for classifying inhibitors and non-inhibitors. The model was trained using 80 percent of the diverse data set, and the remaining 20 percent was used for external validation. In particular, machine learning models using various molecular descriptors and a deep learning model using molecular graph representation were developed. Based on the metrics of all approaches, the random forest model with 2d and 3d features outperforms other models with (R2 0.81) (RMSE 0.10). When compared to graph-based techniques such as GCN, GAN, and AttentiveFP, the success of the machine-learning approach in predicting the binding affinity of hBACE-1 inhibitors provides a strong thrust for systematically applying such methods for drug screening and developing successful BACE-1 inhibitors. The molecular docking-based virtual screening approach was unsuccessful in ranking the BACE-1 inhibitors, suggesting that such methods cannot be reliably used for identifying the lead compounds from different chemical libraries.

Keywords: Alzheimer's disease, BACE1, Graph Convolutional Network, Machine Learning,Deep Learning, Drug Discovery

Contents

List of Figures

Chapter 1

Introduction

1.1 Alzheimer's disease

Neurodegenerative diseases are incurable that lead to nerve cell degeneration or death. This disease causes problems with movement (ataxia) or mental functioning (dementia). A decline in reasoning, memory, and other cognitive abilities causes dementia. As the progression continues, it impacts daily activities. There are more than ten types of dementia. 1. vascular dementia; 2. mixed dementia; 3. Lewy body dementia; 4. Parkinson's disease; 5. frontotemporal dementia; 6. Creutzfeldt-Jakob disease; 7. hydrocephalus with normal pressure; 8. Huntington's disease; 9. Wernicke-Korsakoff syndrome; 10. mild cognitive impairment. Alzheimer's, which follows mild cognitive impairment, is the most common form of dementia. Alzheimer's disease (AD) is a progressive, irreversible brain disorder characterized by neurological and behavioral disabilities caused by plaque and tangle formation, synaptic loss, and neuroinflammation[1]. Neuronal cells use chemical transmitters (neurotransmitters) to transmit electrical messages to other body parts. However, in Alzheimer's, the brain tissue area is damaged, and some messages cannot transmit,resulting in disease symptoms. The first sign of Alzheimer's disease is a progressive loss of intellectual capacity accompanied by short-term memory loss, neuronal loss, cerebral cortex synapses, and certain subcortical regions. This loss causes gross atrophy of the affected region, which includes temporal, parietal, cingulate, and frontal cortex degeneration.

1.1.1 The pathological hallmark of Alzheimer's disease

Brain atrophy, extracellular amyloid plaques, and intracellular neurofibrillary tangles (NFTs) are the main characteristic features and pathological hallmarks of Alzheimer's disease.

1.1.2 Protein aggregates in the AD brain

Amyloid plaques:

Amyloid plaques are composed of Aβ fragments. The dense core and diffuse plaques are two types of fragments. The dense core is unique to beta-pleated sheets. These plaques' surroundings can reveal various abnormalities, such as neuronal and synaptic loss and dystrophic neurites. Plaque deposition begins early and can be seen in Down syndrome.

Neurofibrillary tangles:

NFT is a hyperphosphorylated form of the microtubule-associated tau protein that occurs within the neuron. It is found in other types of neurodegenerative diseases, such as frontotemporal dementia, Pick's disease, progressive supranuclear palsy, and corticob

1.1.3 Role of Aβ

There are several hypotheses for the pathogenesis of Alzheimer's disease, including the amyloid hypothesis, the cholinergic hypothesis, and the tau hypothesis. The progression of Alzheimer's disease is primarily associated with the amyloid and cholinergic hypotheses, which result in an elevated level of Aβ peptide. The aggregation and accumulation of these peptides in the brain increase neuronal toxicity, resulting in nerve degeneration and damage.

1.1.4 Aβ development:

β-secretase cleaves APP in its ectodomain to produce soluble APPβ (sAPPβ) and its β-C(C-99)-terminal fragment (β-CTF).In the amyloid pathway, γ-secretase cleaves the N-terminal side of β-CTF resulting in the generation of the APP intracellular domain (AICD). Aβ-species of varying lengths, namely A[1-38], A[1-40], soluble and non-neurotoxic, and A[1-42] is a less soluble and more hydrophobic, tendency to have undergone fast aggregation and tend to be toxic. α-secretase cleaves APP to produce

soluble APPα (sAPPα-neuroprotective) and its α-C(C-83)-terminal fragment (α-CTF). α-secretase cleaves γ-CTF resulting in generating the APP intracellular domain (AICD) and p3 (30kD peptide), peptides A [17β[17-40] and Aβ[17-42]. Both peptides are amyloid plaques, but this Aβ-42 is neurotoxic and can inhibit long-term potentiation.

1.1.5 Autosomal dominant and Sporadic AD:

Autosomal dominant inheritance of Alzheimer's disease begins before the age of 65. It is caused by an autosomal dominant mutation discovered in APP (PSEN1 or PSEN2), APP gene dosage (chr21). Aβ-42 is chronically overproduced, so either the A42/A40 ratio is increased, or Aβ aggregation kinetics are altered. The onset of sporadic Alzheimer's disease is later than autosomal dominant Alzheimer's disease. The primary causes are environmental and genetic risk factors. There are three types of APOE: 2 (decreased odd ratio), 4 (increased odd ratio), and mildly increased odd ratio. These loci affect gene products involved in endosomal vesicle trafficking, immune response, lipid metabolism, Aβ aggregation, and clearance.

1.1.6 Aβ Clearance:

In Sporadic AD, Aβ clearance from the central nervous system (CNS) slows down with increasing age. This clearance involves several mechanisms, including cellular uptake, enzymatic degradation, and cerebrovascular system-mediated clearance.

1.2 APP - physiological function

Amyloid precursor protein (APP) is a transmembrane protein of type I. Amyloid precursor-like protein 1 (APLP1) and amyloid precursor-like protein 2 (APLP2) are APP members (APLP2). If the sequence similarity between homologs is high in the cytosolic and ectodomain regions, then Aβ is unique to APP. APP has physiological functions in both neuronal and synaptic functions. APP processing is classified as amyloidogenic (Aβ-generating) or non-amyloidogenic. During ectodomain shedding (by α- or β-secretase), APP undergoes regulated intramembrane proteolysis (RIP), followed by -secretase-mediated proteolysis.

1.3 Secretase of APP

1.3.1 α-secretase:

alpha-protease belongs to metalloprotease enzyme which requires a cofactor (zinc) and belongs to the ADAM (A Disintegrin And Metalloproteinase) family. ADAM 9, ADAM 10, ADAM 17, and ADAM 19 are members of the ADAM family. ADAM19 cleaves APP, whereas ADAM17 induces α-secretase activity, which cleaves tumor necrosis factor (TNF), transforming growth factor (TGF), and intercellular adhesion molecule 1(IAM).

1.3.2 γ-secretase:

PSEN1/PSEN2, APH1a/APH1b, and PEN2 form a multi-complex protein. Apart from the APP CTF's more than 90 membrane substrates, such as the p75 neurotrophin receptor, E-cadherin has been identified.

1.3.3 β-secretase:

β-secretase (the aspartic acid protease BACE1) is also known as aspartyl protease 2 (Asp2) and membrane-associated aspartic protease 2 (memapsin 2). BACE-1 and BACE-2 are aspartyl proteases with two catalytically active aspartyl residues (D93 and D289). These proteases are closely related to the pepsin family and distantly related to the retroviral aspartyl protease family. BACE-1 is a 501 amino acid protease with a gene on chromosome 1 that cleaves an APP site at M596-D597 and, to a lesser extent, Y606-E607 in an acidic environment. The two active site motifs are asp-thr-gly-ser (residues 93–96) and asp-ser-gly-thr (residues 289–292). The BACE-1 orthosteric active site utilizes two aspartate residues (D32 and D228) to anchor endogenous substrates or synthetic ligands [2]. It is located between the N- and C-terminal domains. The incredibly flexible hairpin loop shields the shallow and highly exposed binding site. The X-ray co-crystal structure of BACE1 with peptidomimetic inhibitor interactions was critical in developing BACE1 inhibitors[3]. R289 and the hydrophobic pockets formed by the active site were crucial in substrate binding. The active site size of BACE1 (28 amino acids) is relatively large, and finding small molecules to occupy such a large active site is difficult. The blood-brain barrier's permeability is another major challenge. The developed BACE1 inhibitors were susceptible to efflux by P-group proteins. Therefore, drug entry into the brain is complex, even if the drug has crossed the barrier.

Mutation	Site	Effect
Double point (Swedish mutation)	K670N/M671L	yields a 6- to 8-fold increase in Aβ production and leads to Autosomal dominant AD
point	A673T	β's deliverance from β-CTF by γ-secretase,Aβ aggregation kinetics
point (Flemish)	A692G	Increase in Aβ aggregation
point (Dutch)	E693Q	Increase in Aβ aggregation
point (Austrian)	T714I	Increase in Aβ42 levels
point (French)	V715M	Increase in Aβ42 levels
point (Florida)	I716V	Increase in Aβ42 levels
point (Indiana/London)	V717L/V717F	Increase in Aβ42 levels
point (Austrian)	L723P	Increase in Aβ42 levels
point	E682K	β' cleavage is blocked
Substitution		
point	A673V	progression of BACE1-mediated APP cleavage

Figure 1.1: BACE1 in AD

1.4 APP cleavage BACE1's functions:

γ-secretase-mediated release of Aβ initiates the BACE1 cleavage within the Aβ sequence(L703 and M704). BACE1 cleavage results in Aβ34 fragments generated from substrates Aβ40 and 42. Thus, BACE1 has not only played an important role in Aβ

production (β-cleavage) but also anti-amyloidogenic cleavage (β'-cleavage within the Aβ sequence) and A34 generation (enzyme degradation). The physiological function of BACE-1 is impaired in synaptic plasticity and memory, muscle spindle formation and maintenance, retinal pathology, silent epileptic seizures, hypomyelination, and axonal guidance defects. Non-enzymatic functions are involved in modulating voltage-gated sodium and potassium channel functions. The effects of these functions are mediated by BACE1-mediated proteolysis of the accessory subunits[4].

1.5 BACE1 as a therapeutic target in AD:

The majority of BACE1 Inhibitors also inhibit BACE2. NB-360 long-term treatment for BACE1 inhibition shows potential benefits. The drug prevents the inflation level of CSF-tau of β-amyloidosis.LY2886721 (liver toxicity), LY2811376(auto-fluorescent deposits in the retina), and verubecestat (MK-8931) help in lowering CSF levels of sAPPβ, Aβ40, and Aβ42. Suppressing BACE1's expression to attenuate Aβ production aids in preventing AD's development.three cleavage sites within APP for BACE1: canonical β-site, β'-site and β34-site. β' and β34 cleavages could serve anti-amyloidogenic activities. In order to ensure safe prevention treatments in the correct clinical state, it is important to understand the biochemical changes leading to AD onset.BACE1 targeted drug able to penetrate through the blood-brain barrier (BBB); large and conformationally flexible active site; homology to other aspartic proteases; the presence of an efflux transporter, phosphoglycoprotein (P-gp); cell membrane permeation improvement

1.6 Scope of the current study:

Even with the availability of numerous publications, patents, and proceedings related to the pathogenesis of Alzheimer's disease, technological advancement, structural information, and molecular modeling for the development of BACE1 inhibitors. There are still unmet clinical needs in the treatment of Alzheimer's disease.

To understand the binding properties of BACE1 inhibitors, we used machine learning, deep learning, and virtual screening approaches in this study. Our goal is to compare the results and determine the best method for predicting the binding affinities of BACE1 active compounds.

1.7 Thesis outline

In chapter 2, we will see how some of the data driven approach (machine and deep learning-based methods) are used to understand the binding properties of BACE1 inhibitors. We will also see, how binding properties is impacted when using the concept of data-driven and deterministic approach and will explore different consensus algorithms which are used for the same.

chapter 3, will be a concluding part of the thesis, where the results and overall analysis of the methods used will be summarized followed by the future scope of work.

Chapter 2

Materials and Methods

2.1 Data Driven Approaches

A problem-solving technique that entails analysing and interpreting data to inform decision-making and solving problems. It entails gathering and analysing data from various sources, employing statistical techniques and machine learning algorithms to identify patterns and trends, and applying these insights to make informed decisions or develop solutions to problems. The advantage is that decisions are objective and unbiased.

2.2 Why Data Driven Approaches?

2.2.1 To avoid dealing with human subjects : Clinical trial phases

Data-driven approaches can help design and analyze clinical trials, as they can help objectively evaluate the safety and effectiveness of new medical treatments and interventions. However, it is essential to note that clinical trials are just part of the data-driven medical research and development approach.

There are several ways in which data-driven approaches can help reduce the number of clinical trials that are needed:

Predictive modeling: By using machine learning algorithms and historical data, it may be possible to predict which treatments are likely safe and effective and which clinical trials are most worth pursuing. This can reduce the number of trials conducted and save time and resources.

Real-world data: By collecting and analyzing data from real-world clinical settings, it may be possible to gain insights into the safety and effectiveness of new treatments without conducting formal clinical trials. This can be especially useful for rare diseases or conditions where it may be difficult to recruit sufficient subjects for a clinical trial.

Clinical trial simulation: By using computational models to simulate clinical trials, it may be possible to predict the outcomes of a trial without actually conducting it. This can reduce the number of needed trials and save time and resources.

Overall, data-driven approaches can help optimize the design and analysis of clinical trials.

2.2.2 Availability of numerous compound library for lead compounds

Using a data-driven approach, there are several ways to search for lead compounds in a compound library for drug discovery.

A computer-aided drug design (CADD) tool to help you identify compounds likely to have the desired pharmacological properties based on their chemical structure and other characteristics. These tools can use machine learning algorithms to analyze large datasets of compounds and their known biological activities and predict which compounds are most likely to be effective as lead compounds in drug discovery.

Text-mining tools to search the scientific literature for compounds that have the potential to be lead compounds in drug discovery. These help to identify papers that discuss compounds with particular biological activity or that have been studied in particular disease areas and can provide valuable insights into the potential of these compounds as lead compounds in drug discovery.

ChEMBL with known biological activity that contain information on millions of compounds and their activity against various targets—searches using various criteria, including chemical structure, biological activity, and other characteristics.

2.2.3 To reduce Time duration and cost

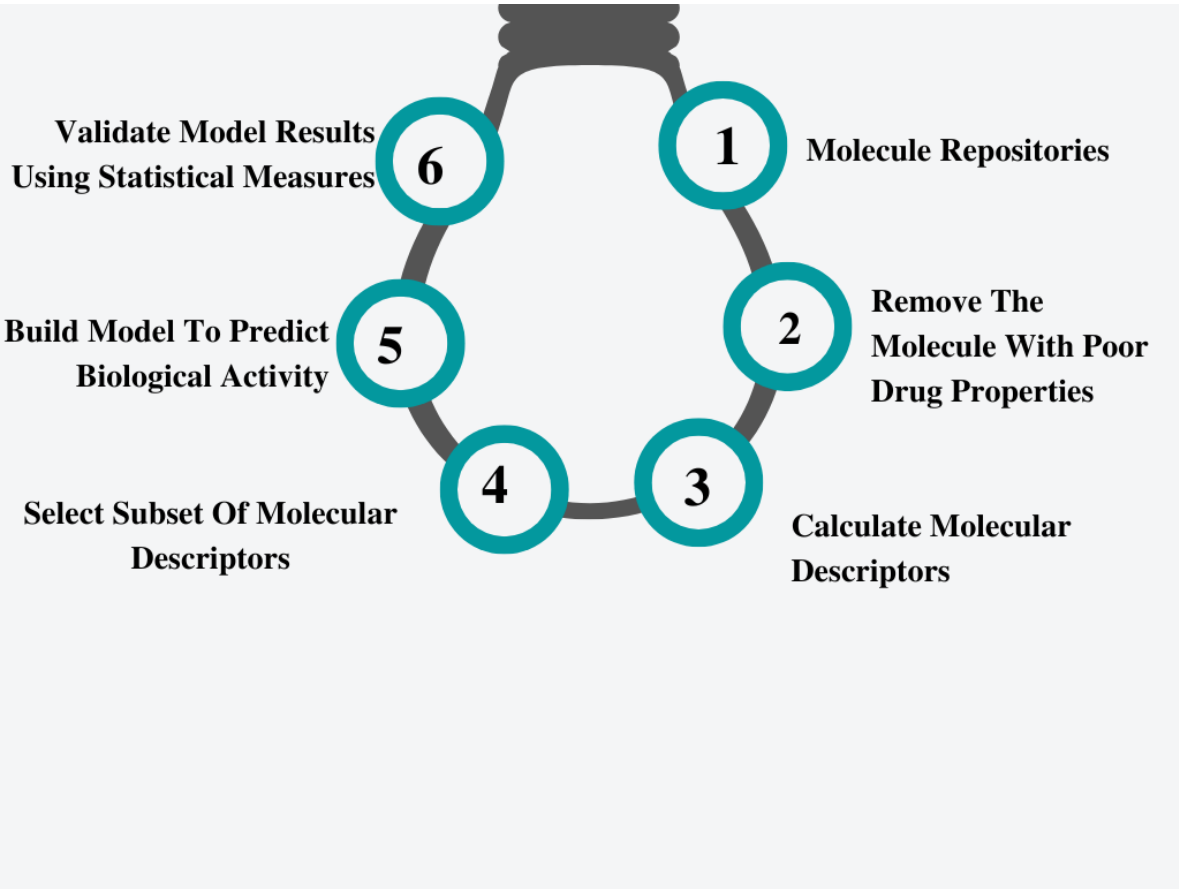
The time and duration of a data-driven approach can depend on several factors, including the targeted disease's complexity, the availability of relevant data and resources, and the efficiency of the research and development process.

In some cases, a data-driven approach may require more time and resources due to the gathering and analysis of large amounts of data to inform decision-making. For example, suppose a drug target is poorly understood or lacks relevant data. Using a

data-driven approach, it may take more time and resources to identify and validate potential drug candidates.

On the other hand, a deterministic approach may reduce time and duration costs by following a predetermined set of steps and minimizing the need for data gathering and analysis.

2.3 How does Data Driven work?



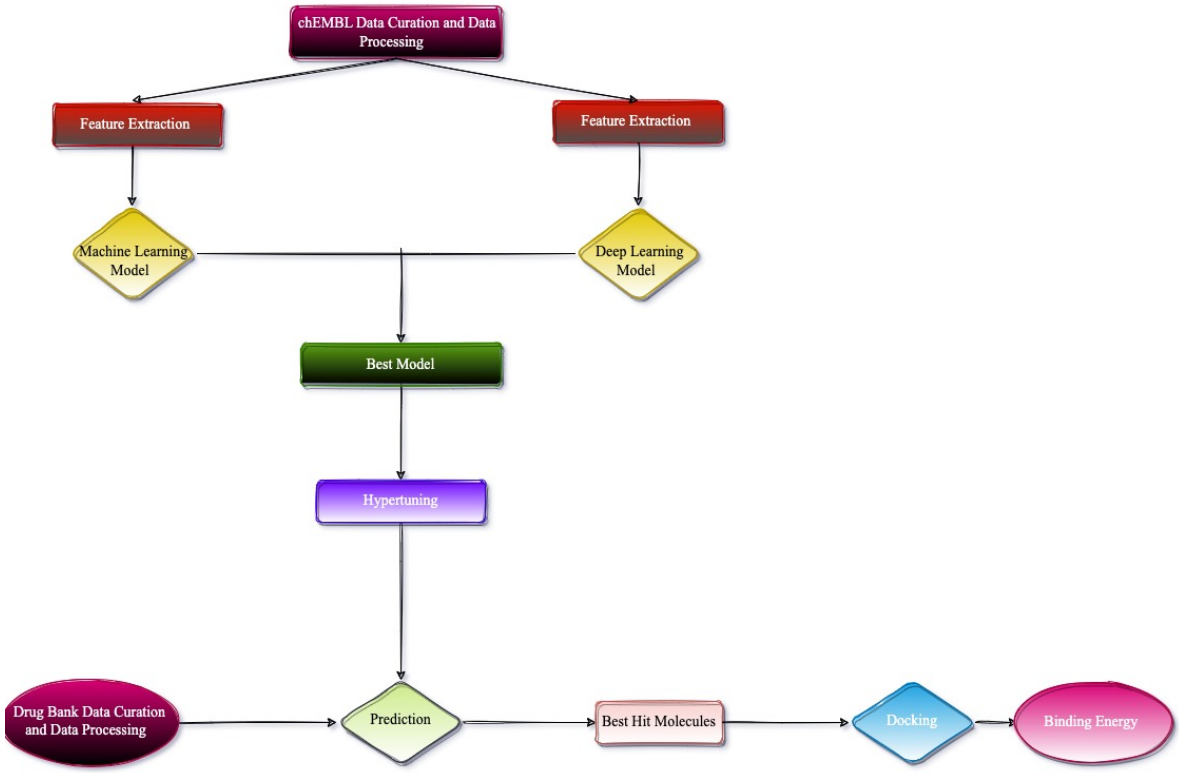


Figure 2.1: Work Flow

2.4 Data Collection

2.4.1 Assembling Druggable Dataset

chEMBL is an open-access database and web interface with open-source code available on GitHub. chEMBL contains the curated biological activities of more than two billion compounds. First, the data with confidence scores provided by chEMBL of greater than or equal to 6 were selected. The confidence score indicates the confidence level of the target protein assigned to the compounds. The assay type belongs to binding assay (B) by an in-vitro experiment. Standard units of nm were filtered. p-activity values (pIC_{50}) defined by -log() where () = IC_{50}. Further IC_{50} values were chosen, and they were directly proportional to activity, "activity-comment" was neither "None" nor "Not Determined". Duplicates were removed. Chemical compound structures were extracted in SMILES format (simplified molecular-input line-entry system) from chEMBL. Salts and solvents were removed, neutralized, and converted into canonical SMILES.

DATA RETRIVAL FROM CHEMBL

```
target = new_client.target
target_query = target.search("BACE1")
targets = pd.DataFrame.from_dict(target_query)
targets
```

	cross_references	organism	pref_name	score	species_group_flag	target_chembl_id	target_components	target_type	tax_id
0	[]	Cricetulus griseus	Beta-secretase 1	13.0	False	CHEMBL3297642	[{'accession': 'G3IAK4', 'component_descriptio...	SINGLE PROTEIN	10029
1	[{'xref_id': 'Beta-secretase_1', 'xref_name': ...	Homo sapiens	Beta-secretase 1	11.0	False	CHEMBL4822	[{'accession': 'P56817', 'component_descriptio...	SINGLE PROTEIN	9606
2	[{'xref_id': 'P56818', 'xref_name': None, 'xre...	Mus musculus	Beta-secretase 1	11.0	False	CHEMBL4593	[{'accession': 'P56818', 'component_descriptio...	SINGLE PROTEIN	10090
3	[]	Rattus norvegicus	Beta-secretase 1	11.0	False	CHEMBL3259473	[{'accession': 'P56819', 'component_descriptio...	SINGLE PROTEIN	10116
4	[]	Homo sapiens	Beta-secretase (BACE)	6.0	False	CHEMBL2111390	[{'accession': 'Q9Y5Z0', 'component_descriptio...	PROTEIN FAMILY	9606

Figure 2.2: Data Retrival from chEMBL

- **chEMBL (https://www.ebi.ac.uk/chembl/)**

2.4.2 Preparing Data for Feature generation: SMILES to mol2 conversion

A simplified molecular-input line-entry system (SMILES) is a way to represent molecules as strings of characters. It is a very compact and human-readable representation that can describe a molecule's structure. The representation of the smiles for the Heavy atoms has explicit atomic symbols, while hydrogen atoms do not (implicit), For example, CH4-C and NH3-N. Elements that do not belong to the organic subset Eg: [Au]. Bonds that are single, double, triple, or aromatic -,=,#, Branches are represented by () and stereochemistry @.

SMILES can be imported into various molecular editors or cheminformatics software to generate 2D or 3D structures that can be used to develop various machine learning models (data-driven approach), virtual screening, or molecular docking. A tool like Open Babel can convert a SMILES string to a mol2 file. Open Babel is a free and open-source chemical file format converter that can convert between various chemical file formats, including SMILES and mol2. The RDKit is a collection of cheminformatics and machine-learning software in C++ and Python. It can perform various chemical informatics tasks, such as generating 2D and 3D structures from SMILES strings, performing structural searches, and calculating various chemical properties. Many other tools and online services can be used to convert SMILES to mol2, such as the RDKit, ChemAxon, and the PubChem website. We have used RDKit and Open Babel for our study.

source for data conversion:

```
#!/bin/bash

mvalue=7234
nvalue=0
for smile in $(cat bace1.smi)
do
nvalue=$((nvalue + 1))
echo $nvalue
if [ "$nvalue" -gt "$mvalue" ]; then
/usr/bin/obabel -ismi -:"$smile" -osdf -Obace$nvalue.sdf -h --gen2d
/usr/bin/obabel -isdf bace$nvalue.sdf -omol2 -Obace$nvalue.mol2 --gen3d
fi
done
```

Figure 2.3: mol2 conversion using OpenBabel

- **OpenBabel** (https://openbabel.org/)

- **RDKit** (https://www.rdkit.org/)

2.4.3 Adding Features to Data

For feature extraction, signaturizer and PaDEL were used. The signaturizer provides information about small molecules' physiological and structural properties. However, there are no available small-molecule bioactivity descriptors for many applications. Signaturizer is capable of detecting 25 distinct bioactivities of molecules (including cellular responses, clinical outcomes, and target profiles) and working on the signature-activity relationship (SAR). PaDEL can calculate molecular descriptors and fingerprints, and paDEL currently calculates 1875 descriptors (1444 - 1D,2D,431- 3D, and 12 fingerprints). The features are calculated based on The Chemistry Development Kit with additional fingerprints and descriptors like atom type electron topological state descriptors, Crippen's logP, and MR, extended topochemical atom (ETA) descriptors, McGowan volume, molecular linear free energy relation descriptors, ring counts, and counts of chemical substructures.

source for Molecular Descriptor Calculation:

- **PaDEL-Descriptor** (http://www.yapcwsoft.com/dd/padeldescriptor/)

```
from rdkit import Chem
from rdkit.Chem import AllChem


smile = 'smilesinput'

uncharged_mol_1D = Chem.MolFromSmiles(smile)

uncharged_mol_3D = Chem.AddHs(uncharged_mol_1D)
AllChem.EmbedMolecule(uncharged_mol_3D)
AllChem.UFFOptimizeMolecule(uncharged_mol_3D)

charged_mol_3D = uncharged_mol_3D
AllChem.ComputeGasteigerCharges(charged_mol_3D)

fout = Chem.SDWriter('./charged_test.mol')
fout.write(charged_mol_3D)
fout.close()
```

Figure 2.4: mol2 conversion using RDKit

```
#!/bin/bash
for i in {1..7234}
do
java -jar /storage/priya/bace1_obabel/PaDEL-Descriptor.jar -removesalt -maxruntime -1 -2d -3d -fingerprints -dir ./bace$i.mol2 -file ./bace$i.csv -descriptortypes ./descriptors.xml -usefilenameasmolname
done
```

Figure 2.5: Feature generation using paDEL

- **signaturizer** (https://gitlabsbnb.irbbarcelona.org/packages/signaturizer/)

2.5 System setup (tools and library installation)

This section will describe the system requirements and configuration for this study. We used the Ubuntu 18.04 operating system for data processing, and we used the HPC server at IIIT Delhi for modeling. Although we used Ubuntu 18.04 for data

preprocessing, the codes are also compatible with Windows systems using a Python compiler. We followed the steps below to set up the Python environment. Python is a user-friendly and versatile programming language. The developers wanted to create a programming language that was fun to use, so they named it after the British comedy group Monty Python. Python 3 is the most recent language version and is widely regarded as Python's secure future. We needed a computer system with Ubuntu 18.04 installed, administrative access to that machine, and an internet connection to install Python. After preparing the machines, we began with the Python installation and configuration via the command line. A command-line interface, also known as a shell, is a non-graphical way to interact with your computer.

On Ubuntu, you can open the terminal by heading to the start menu and typing "terminal" into the search bar; alternatively, you can use the CTRL+ALT+T keyboard shortcut to open the terminal application automatically.

Python 2 and Python 3 are already installed in Ubuntu 18.04. We use apt-get to upgrade the system to ensure the Python version is up-to-date.

pip install pandas

The "-y" flag indicates that we agree to have all items installed in the system. When the run is finished, we can check the version of Python 3 installed on the system.

Python3 -V

The output in the terminal window shows the version number.

Python 3.8.0

We now have pip installed, which will aid in the management of Python software packages.

sudo apt-get install -y python3-pip

Pip is a Python tool for installing and managing programming packages that we may use in the future for the project. We can install all of the Python packages after installing the pip tool by typing:

pip3 install 'package name'

In this case, 'package name' can refer to any Python package or library, such as Pandas, Numpy, Biopython, etc. Ubuntu also includes the option to create a virtual environment. Virtual environments allow you to have a separate space dedicated to a Python project on your machine. This ensures that each of your projects can have its own set of dependencies and versions of Python and Python packages that will not interfere with the other projects. The virtual environment for separate projects gives us more control over our projects and how we can use different versions of packages on the same machine. We can create as many environments as we want.Each environment is a directory/folder in the system that contains a few scripts that make it act like an environment. To create a virtual environment, we must first install the venv module,

which is part of the standard Python 3 library.

sudo apt-get install -y python3-venv

To create an environment, we must first select a directory in which we want to install Python. We made the directory with mkdir, as shown in To create an environment, run

python3 -m venv Th env

This will create a new directory with a few items needed to create the environment. To use this environment, we must first activate it by issuing the following command:

source allows Th env

This command invokes the activate script, and we can now install all the packages and libraries required for our project.

2.5.1 Installing Jupyter Notebook

Jupyter notebook is handy for creating and sharing interactive codes, data visualization, and other tasks. Jupyter notebook can provide a variety of programming languages, such as Python, R, and Ruby. It is frequently used for statistical analysis, machine learning, modeling, and other tasks. Pip was used to integrate Jupyter notebook into our already-running virtual environment.

pip install jupyter

This will successfully install all of the software required to run Jupyter. We can now start the Jupyter notebook server by running the following command:

Jupyter notebook Or jupyter notebook –ip 'IP Address'

2.5.2 Settingup python environment

We used Pandas and NumPy libraries for data visualization and preprocessing, as well as Biopython to obtain GenBank data and query NCBI. The following commands were used to install all of these libraries in the same environment.

Using terminal

pip install pandas
pip install numPy
pip install biopython
pip install perl

Using jupyter notebook

!pip install pandas
!pip install numpy
!pip install biopython
!pip install perl1

2.5.3 Autodock installation

Molecular docking predicts the preferred orientation of one molecule with another to form a stable complex in the field of Molecular modeling. Using methods such as scoring functions, preferred orientation can be used to predict the strength of association or binding affinity between two molecules. Because molecular docking can predict the binding configuration between a suitable target binding site and a small ligand, it is one of the most widely used methods in structure-based drug design. Binding behavior characterization is essential for rational drug design and elucidating fundamental biochemical processes. In this study, we used AutoDock Vina to perform molecular docking.

2.6 selection of significant descriptors

Feature selection is an essential step in any machine-learning pipeline because it aims to identify relevant features while eliminating a group of irrelevant features that would introduce unnecessary noise. Traditional feature selection methods are minimally optimal, whereas Boruta is an appropriate FS(Feature Selection) method that aims to find all relevant features for prediction rather than a possibly compact feature subset. Feature ranking, automatic n-estimator selection, and faster run times are some of the advantages of the boruta algorithm. source for Molecular Descriptor Calculation:

- **Boruta https://github.com/mi2-warsaw/Boruta/**

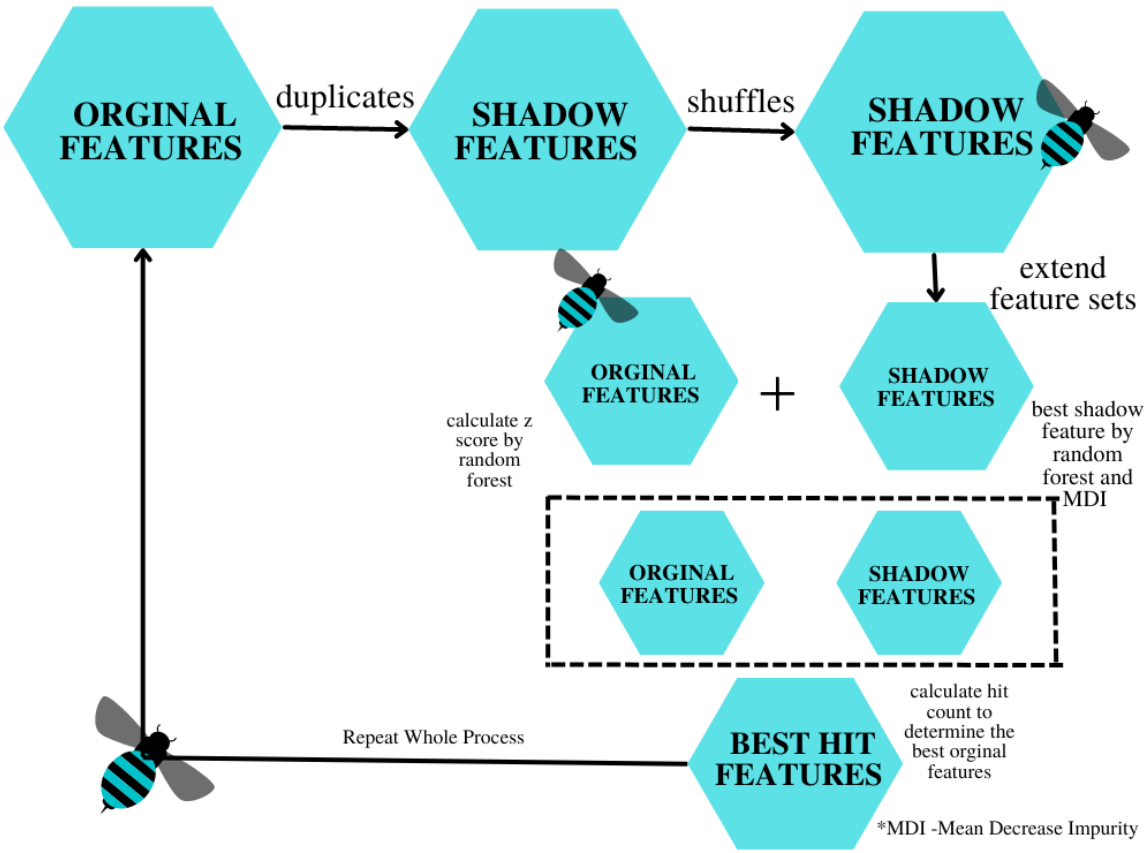


Figure 2.6: boruta feature selection

2.7 Machine Learning and Deep Learning Models

We performed machine learning (classification and regression) and deep learning (classification and regression). We used a graphical-based method to build DL models for binary classification and regression to obtain the best performance model. Both machine learning (ML) and deep learning (DL) were performed using Python 3.9. For the ML model, we used the scikit library. To avoid bias in the model, we oversampled the data before building it. We used random over-sampling and the synthetic minority oversampling technique (SMOTE) for up-sampling. SMOTE generates synthetic minority class samples that outperform random samplers. The data was divided into 80 and 20 percent for training and testing, respectively. The DeepChem library (version 2.4.0) was used for the DL models. TensorFlow (an open-source library for DL ML), SciPy,

Scikit-learn, Pandas, and NumPy were installed to support DeepChem. DeepChem has several features for feature extraction. These feature generators use data as input to generate features for molecules. Our models used the molecular graph convolution featurizer (MolGraphConvFeaturizer). MolGraphConvFeaturizer is a molecule-specific graph convolution network. Many splitters are available in DeepChem to split data into training and testing; we used ScaffoldSplitter, and for DL, we split data into 80 (Training) and 20 (Testing) (Testing). Four classifiers were used to build models: Attentive, CNN, GCNetwork, and GAT. Aside from training and testing, various parameters such as precision score, MCC, F1 score, recall, and AUC value were used to determine model accuracy. For regression, R2(coefficient of determinants), mse(mean square error), and make (mean absolute error)metrics were used. After performing ML and DL, we compared all the parameter scores across all models. The best model performed from the hyperparameter tuning (HPT) was chosen[5]. HPT provides the best model architecture. We used the Gaussian process and then performed k-fold cross-validation. source for library used for data driven approaches:

- **Scikit-learn** **https://scikit-learn.org/stable/index.html**

- **DeepChem** **https://deepchem.io/**

2.8 Ensemble learning:

Ensemble learning is a machine learning technique in which multiple models are built and combined. When compared to individual models, these methods improve prediction accuracy. Bagging and random forest belong to ensemble methods that minimize variance. Functional gradient descent, boosting, and ensemble selections belong to ensemble methods that minimize bias. Random Forest uses K randomly chosen attributes at each node to build a tree. It also allows the estimation of class probabilities (or target mean in the regression case) based on a hold-out set (back fitting) and no pruning. In this study, we used this method by simply averaging the individual outputs without weighting.

2.9 Graph Convolutional Network:

DeepChem's RDKit transformed canonical SMILES into a binary vector with 75 dimensions per atom. This vector includes physiochemical properties, atomic type, formal

charges, hybridization, and several valences. Neighboring atoms were added in GCN using the initial vector as input, and atom information was updated with the maximum value of a neighboring atom in the graph pooling layer and then converted into one dense layer. This layer generates a numerical vector, which is then combined on the graph to produce a neural fingerprint. The output neuron layer links to the graph-gathering layer. The parameters used include the Adam optimizer, relu (convolutional layer), and tanh (graph gathering layer) as activation functions and batch normalization to prevent overfitting and improve learning efficiency[6].

2.9.1 Hyperparameter tuning and model training:

DeepChem 2.1.0 and the pyGPGO package were used to perform hyperparameter optimization using Bayesian optimization with Gaussian processes. The hyperparameters were investigated independently for architectures with a wide range of convolutional layers. A Bayesian optimization search was performed 100 times with the Matérn kernel as a covariance function and "expected improvement" as an acquisition function. This calculation was repeated the required number of times with different weights initialized by a random seed value. The mean absolute error (MAE), root-mean-square error (RMSE), and coefficient of determination (R2) are widely used as statistical metrics of model performance in quantitative activity prediction and were calculated using the given equation. The final model training was performed with a fixed initial seed on the best hyperparameter set (excluding epochs). One hundred epochs were first calculated for each model. The learning was stopped if the minimum MAE on the held-out validation set did not decrease further in the subsequent 100 epochs. When the MAE value fell, another 100 epochs of learning were performed. The procedure was repeated without setting an upper limit for the total number of epochs until the previous minimum MAE no longer changed during the additional 100 epochs.

2.10 Molecular Docking

2.10.1 Ligand Preparation:

OpenBabel was used to convert the ligands' smi format (SMILES) to SDF format and then from SDF to pdbqt. The standardization procedure included the removal of salts, mixtures, and metal ions and correcting the chemical structure's geometry.

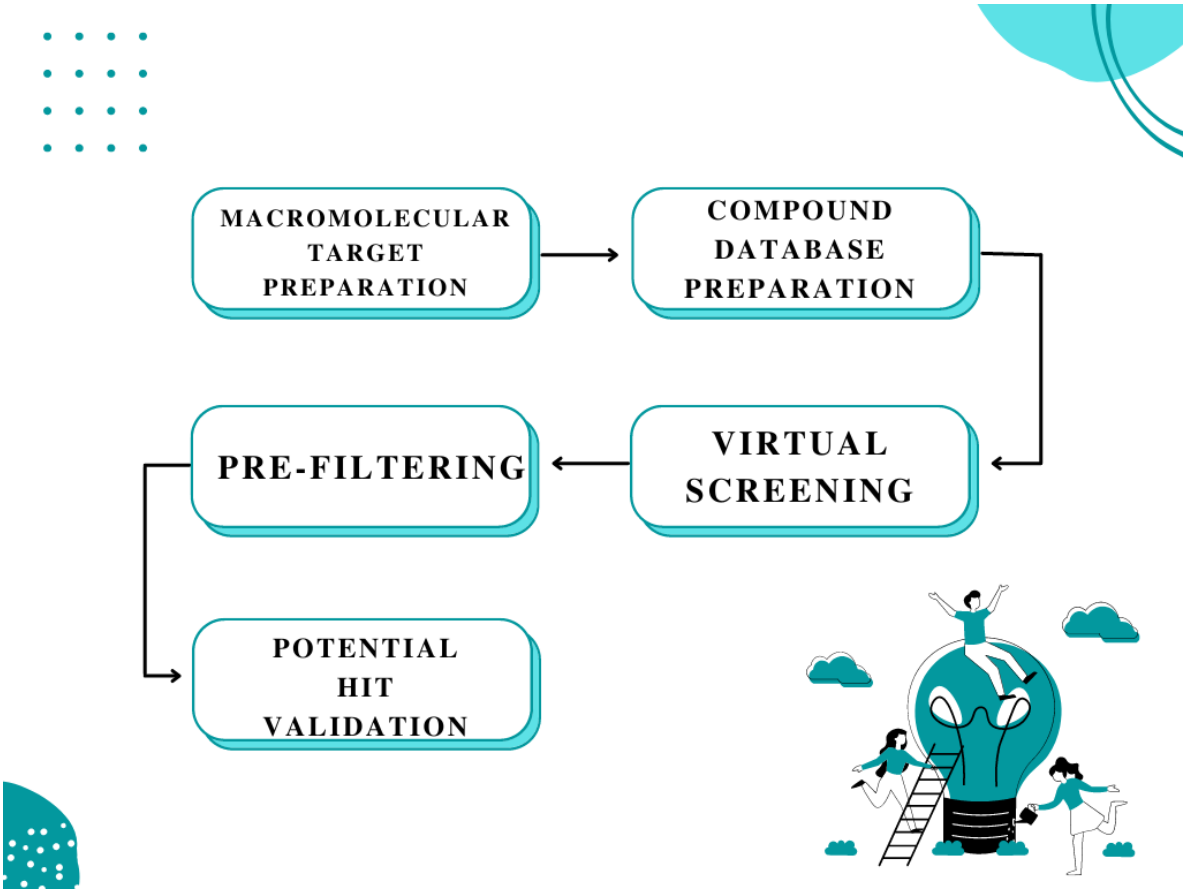


Figure 2.7: deterministic approach workflow

```
#!/bin/bash
ls -1 *mol2  |sed 's/\.mol2//g' >files_lists

for i in $(cat files_lists)
do
file=$i.mol2
/storage/priya/mgltools/bin/pythonsh /storage/priya/mgltools/MGLToolsPckgs/AutoDockTools/Utilities24/prepare_ligand4.py -l $file
#echo "$file"
done
```

Figure 2.8: MOL2 to PDBQTconversion

2.10.2 Protein Preparation:

The Protein Data Bank is a three-dimensional structural database of large biological molecules such as proteins and nucleic acids. The available structure is obtained by X-ray crystallography, NMR spectroscopy, or cryo-electron microscopy. The three-

dimensional structure of the chosen target was retrieved from RCSB-PDB better to understand the functional and physical aspects of a protein and to study the interaction of the amino acid in the protein with the ligands. BACE1 (PDB id: 6EJ2) was downloaded from RCSB-PDB in PDB format. The structure of BACE1 (Aa position 1–501 Aa) was studied in X-RAY DIFFRACTION with a resolution of 1.46 Angstroms. To prepare the protein (BACE1), we downloaded AutoDock4 and the MGL tool and prepared it. Multiple steps were involved in the preparation of the protein (BACE1). First, we loaded protein molecules on AutoDock4, deleted water molecule heteroatoms, and added polar hydrogen. We removed water molecules and heteroatoms because they interfere with the binding of ligands to the protein molecule. Polar hydrogen helps to find the hydrogen bond's interaction, making it more favorable for the ligand to bind with the protein. The Kollman charge (0.0) were also added to the protein, and then the pdbqt format of the prepared protein was downloaded.

Figure 2.9: BACE1 protien structure preparation

- **RCSB-PDB** https://www.rcsb.org/

2.10.3 AutoDock Vina Docking:

Molecular docking is a vital tool in structural molecular biology and computer-assisted drug design. The objective of ligand-protein docking is to predict the predominant binding interaction between the ligand with the protein (chosen target) of the available three-dimensional structure. The open-source software for molecular docking, Vina (v.1.2.0), was used. Virtual screening software for computational drug discovery is used to screen the compounds and to study the interactions against the potential drug targets[7]. We created two files: one is a script file (vina. sh) that generates a log file and a pdbqt file for each ligand, and the other is a configuration (config.txt) file that contains information about the size and center of the grid box that covers the protein inside the box. We performed site-specific docking, in which we gave information about the active center position of the binding site. It covers both the site and the ligand itself. X, Y, and Z centers were -16.519, -37.499, and -10.874, respectively, and the size (dimension) of x, y, and z were 80, 80, and 80, respectively. The CPU, seed, exhaustiveness, and number of models were 10, 1, 5, and 5, respectively.

```
#!/usr/bin/perl
$ligfile="Ligand.txt";
chomp $ligfile;
open (FH,$ligfile)||die "Cannot open file\n";
@arr_file=<FH>;

for($i=0;$i<@arr_file;$i++)
{
print"@arr_file[$i]\n";
@name=split(/\./,@arr_file[$i]);
}
for($i=0;$i<@arr_file;$i++)
{
    chomp @arr_file[$i];
    print"@arr_file[$i]\n";
    system("/disk2/priya/autodock_vina_1_1_2_linux_x86/bin --config config.txt --ligand @arr_file[$i] --log @arr_file[$i]_log.log");
}
```

Figure 2.10: Docking Scripts

- **AutoDock Vina** https://vina.scripps.edu/download.html

2.10.4 Molecular Visualization of Interacting Molecules:

Visualization of the docked structure was performed using PyMOL, which is a molecular visualization software. Molecular visualization of the target protein which binds with the ligand molecule can be identified. Also, the amino acid residue in which the ligand

gets bonded can be visualized. It was also used to compare the binding locations of the target protein with the ligand and its interactions.

- **PyMOL** https://pymol.org/2/

Chapter 3

Results and Discussion

3.1 chEMBL Dataset

The previous section's procedure was used to select a unique canonical smiles dataset of 7234 from chEMBL. The proper representation of inactive compounds has improved classification accuracy and prediction. A wide range of activity values is desirable for regression analysis. We used qualitative measurements above and below the detection limit of the assay.

This R2 is proportional to the p"IC_{50}" distribution range. For BACE1 inhibitors, the maximum and minimum p"IC_{50}" distribution ranges were "12.6987" and "1.06550" respectively. Outliers can have an adverse effect on predictability. As a result, the dataset was divided randomly into validation and test sets for hyperparameter optimization and model prediction.

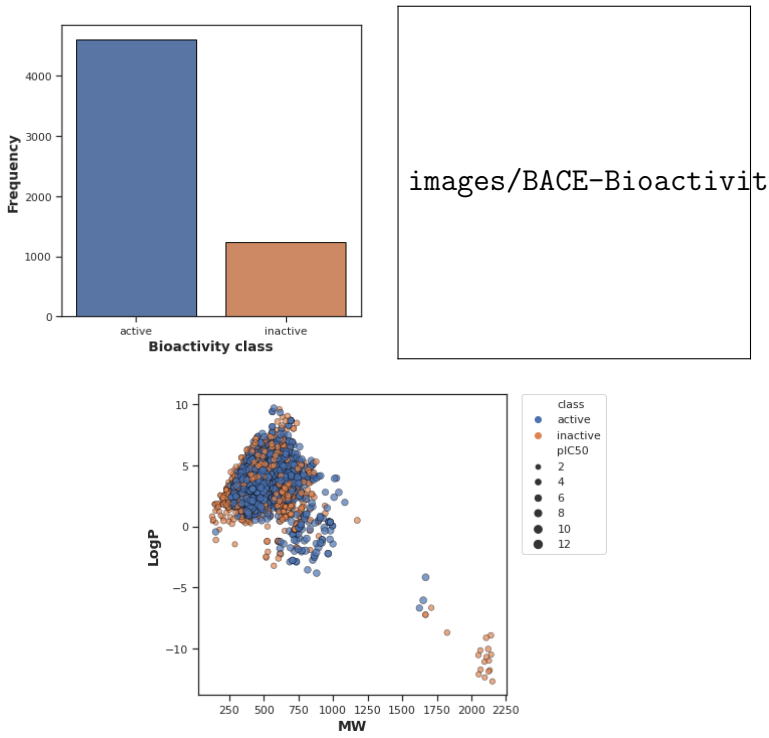


Figure 3.1: Data Curation from chEMBL dataset

3.2 Feature Selection

For molecular descriptor calculation using Signaturizer, canonical SMILES (can) as input(1D=3200) were calculated, and using PaDEL mol2 as input (2D/3D=1972,2D/3D/FINGERPRINT= were calculated[8].

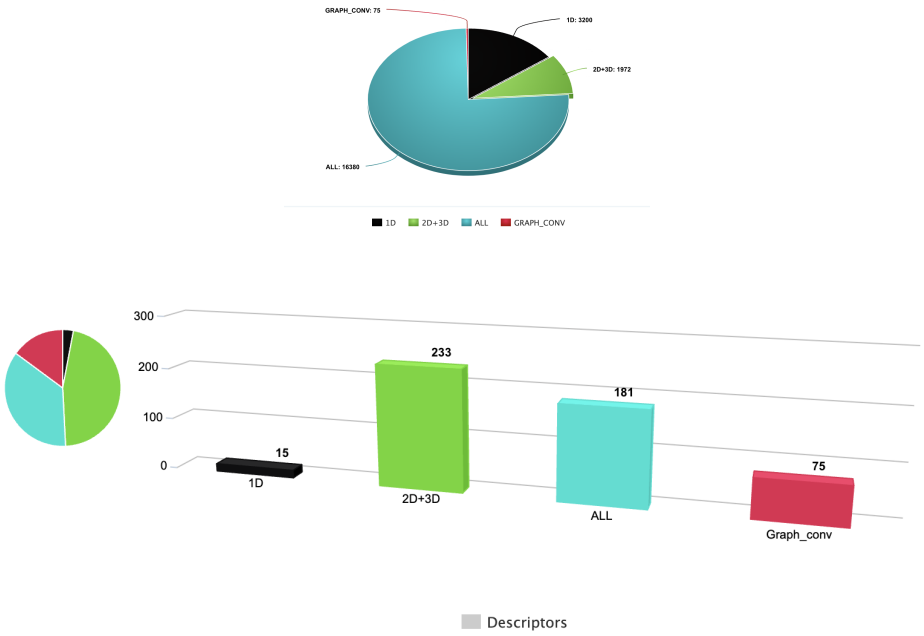


Figure 3.2: Data selection from Boruta

3.3 Machine learning with hyperparameter tuning and model training

For qualitative modelling the data set was divided into two categories: active ("IC_{50}" = 100 nM) and inactive ("IC_{50}" = 100 nM) (using a self-imposed cutoff for the experimental "IC_{50}"). On the training set, these models achieve 98 percent classification accuracy but only 92 percent prediction accuracy on the validation set. The validation set's model performance was used to select RF hyperparameters. The RF models had 10 or 100 trees and used square root, log, or complete cutoffs for the number of features used in tree split computation. Grid hyperparameter search on the validation set was used to determine the number of trees and feature cutoff. For quantitative modelling Rf performed best in the quantitative approach. The results also reveal that using 3D-based approaches provides only minor prediction enrichment compared to 1D and 2D descriptor-based statistical techniques. Given that the experimental observations are incorporated from various labs in various countries, the prediction error of the models is reasonable. The 1D, 2D, and 3D-based RF methods outperform the 3D-based techniques, proving that reliable models can still be developed using traditional descriptors and machine learning methods[9].

Overall, quantitative modeling of the data set provides vital prediction accuracy and can be used successfully in a lead optimization setting.

The observations from the qualitative (Figure 3.2) and quantitative (Figure 3.3) approaches were not game changers, implying that the data size considered here may affect the modeling outcome. The findings support that reasonable models with good prediction capabilities can still be developed with a relatively diverse and small data set.

		TRAINING SET				TESTING SET					
APPROACH	DESCRIPTORS	RECALL	PRECISION	F1	ACCURACY	ROC-AOC	RECALL	PRECISION	F1	ACCURACY	ROC-AOC
	1D	0.975	0.9884	0.9817	0.9818	0.9991	0.8064	0.82	0.8131	0.8148	0.8838
	1D+2D+3D	0.9998	1	0.9994	0.9999	1	0.7288	0.8697	0.7926	0.9194	0.9437
RF	ALL	0.9997	1	0.9998	0.9999	1	0.7426	0.876	0.8035	0.9233	0.9493
	1D	0.9661	0.9974	0.9815	0.9818	0.9993	0.7284	0.7353	0.7317	0.7333	0.7346
	1D+2D+3D	1	1	1	1	1	0.6923	0.7064	0.6988	0.8735	0.8074
DT	ALL	1	1	1	1	1	0.716	0.7247	0.7201	0.8821	0.8214
	1D	0.7996	0.8965	0.8453	0.8541	0.9293	0.774	0.8652	0.8169	0.8266	0.9011
	1D+2D+3D	0.025	0.9031	0.0501	0.7938	0.7261	0.0178	0.7966	0.0348	0.7906	0.7249
SVC	ALL	O.0256	0.9032	0.0499	0.7938	0.7375	0.0178	0.7966	0.0348	0.7906	0.7345
	1D	0.8467	0.9075	0.8749	0.8795	0.9462	0.793	0.845	0.8165	0.8223	0.9003
	1D+2D+3D	0.4127	0.3876	0.3492	0.6959	0.6679	0.4185	0.394	0.3533	0.6968	0.6629
SGDC	ALL	0.4326	0.4346	0.3772	0.7166	0.687	0.4273	0.435	0.3729	0.7128	0.6821
	1D	0.7864	0.8412	0.8129	0.8195	0.9054	0.7511	0.7911	0.7704	0.7764	0.858
	1D+2D+3D	0.712	0.848	0.7709	0.9107	0.9584	0.6371	0.7835	0.7021	0.8855	0.9096
ADA	ALL	0.7342	0.8422	0.7844	0.9149	0.9612	0.6885	0.79	0.723	0.8919	0.9201
	1D	0.8215	0.9175	0.867	0.8744	0.9566	0.7708	0.8688	0.8168	0.8273	0.8937
	1D+2D+3D	0.6461	0.8977	0.7513	0.9098	0.9677	0.5307	0.8031	0.6386	0.873	0.8811
KNN	ALL	0.6283	0.8901	0.7365	0.9052	0.9651	0.5149	0.8049	0.6275	0.8707	0.8633

(a) statistical parameter for qualitative models

		TRAINING SET				TESTING SET			
APPROACH	DESCRIPTORS	MAE	MSE	RMSE	R2	MAE	MSE	RMSE	R2
	1D	0.793	1.0301	1.0149	0.4511	0.213	0.0871	0.2952	0.9554
	1D+2D+3D	0.213	0.0871	0.2952	0.9554	0.5713	0.6104	0.7812	0.6727
RF	ALL	0.2099	0.085	0.2916	0.9565	0.5672	0.6142	0.7837	0.6707
	1D	0.9763	1.8348	1.3545	0.0224	1.234	5.481	2.3411	0.9432
	1D+2D+3D	1.234	5.481	2.3411	0.9432	0.7992	1.3198	1.1488	0.2924
DT	ALL	1.6199	9.2492	3.041	0.6342	0.7557	1.1101	1.0536	0.4048
	1D	1.1214	1.8389	1.356	0.02	0.7007	0.8156	0.9031	0.5835
	1D+2D+3D	0.7007	0.8156	0.9031	0.5835	0.7202	0.8566	0.9255	0.5407
LR	ALL	0.7023	0.8143	0.9023	0.5841	0.734	0.8974	0.9473	0.5188
	1D	0.9736	1.4263	1.1942	0.24	0.5943	0.5848	0.7647	0.7013
	1D+2D+3D	0.5943	0.5848	0.7647	0.7013	0.6559	0.7139	0.8449	0.6172
GB	ALL	0.5829	0.5695	0.7546	0.5792	0.6418	0.6972	0.835	0.6262
	1D	1.1204	1.8358	1.3549	0.0219	0.7098	0.8358	0.7647	0.7013
	1D+2D+3D	0.7098	0.8358	0.7647	0.7013	0.73	0.8728	0.9342	0.532
RIDGE	ALL	0.7072	0.8239	0.9076	0.5792	0.7369	0.9006	0.949	0.5171
	1D	1.1135	1.8208	1.3493	0.0298	0.8926	1.2483	1.1172	0.3625
	1D+2D+3D	0.8926	1.2483	1.1172	0.3625	0.8608	1.1519	1.0732	0.3824
ELASTIC	ALL	0.9016	1.2621	1.1234	0.3555	0.88	1.1949	1.0931	0.3593

(b) statistical parameter for quantitative models

Figure 3.3: Machine Learning

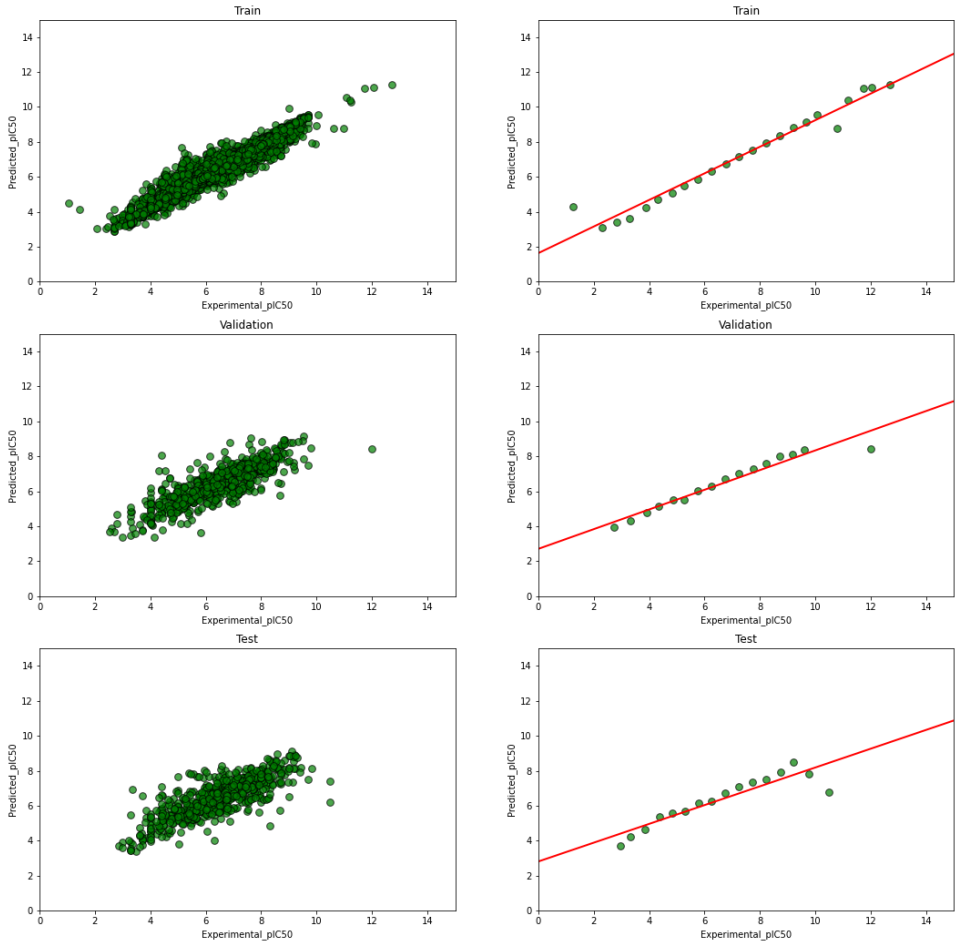


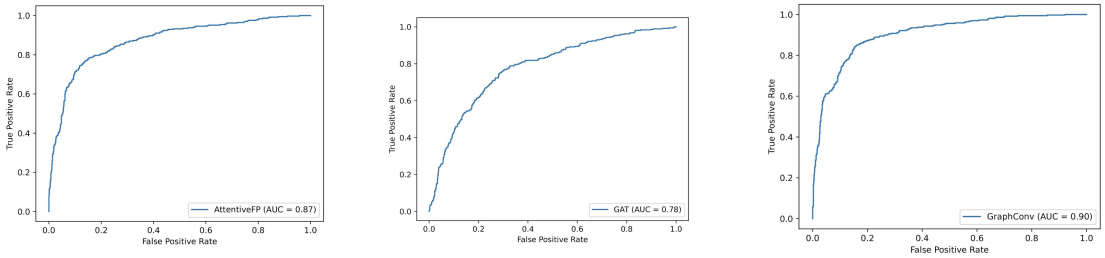
Figure 3.4: Best Machine learning model for quantitative

3.4 Deep Learning with hyperparameter tuning and Model Training

Like other machine learning models, GCN is prone to hyperparameters. The graph's maximum number of convolutional layers is times the value. We set default values for Deep Chem for the hyperparameters not listed in the table[10]. R2, RMSE, and MAE values are often used to evaluate the regression. The lower the R2 value, the less accurate the prediction. An R2 value of 1 indicates a perfect prediction. The lower the RMSE and MSE, on the other hand, the better the prediction accuracy. The activity range of the dataset influences the performance of R2. RMSE tends to increase as the size of the dataset grows. The size of convolutional layers and dense layers varied with the dataset[11].

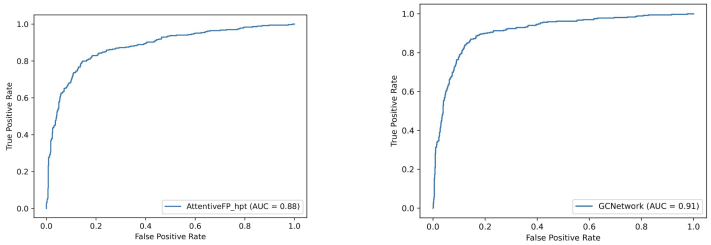
	0	1
graph_conv_layer_size	int	[32, 2048]
batch_size	int	[10, 100]
dropout	cont	[0.0, 0.5]
nb_epoch	int	[20, 200]
learning_rate	cont	[0.0001, 0.002]
dense_layer_size	int	[16, 2048]

Figure 3.5: GCN HYPERTUNING PARAMETERS



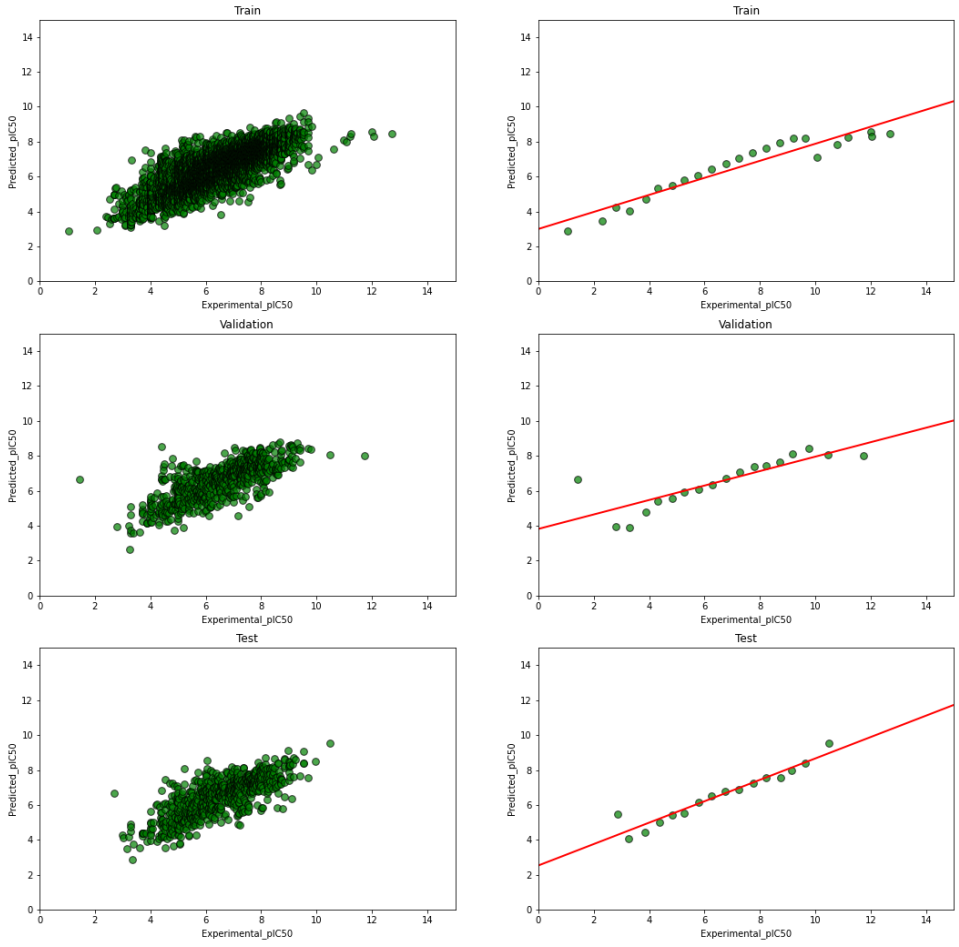
(a) AttentiveFP (b) GAT (c) GCN

Figure 3.6: Deep Learning - Classification



(a) HPT - AttentiveFP (b) HPT - GCN

Figure 3.7: Deep Learning - Classification with hyperparameter tuning


Train
Predicted_pIC50
Experimental_pIC50
Train
Predicted_pIC50
Experimental_pIC50
Validation
Predicted_pIC50
Experimental_pIC50
Validation
Predicted_pIC50
Experimental_pIC50
Test
Predicted_pIC50
Experimental_pIC50
Test
Predicted_pIC50
Experimental_pIC50

3.5 Virtual Screening

Screening is the process of identifying bioactive compounds from an extensive collection of libraries. Screening scores are interpreted as binding free energy. The lower binding energy indicates that the targets can easily bind with the specific ligand and attain a stable conformation with significantly less energy expenditure. The most potent ligand for the targets was mentioned in the table. We also found that the high-quality targets are homologs or functionally correlated, while unrelated targets are poor or mediumly correlated[12].

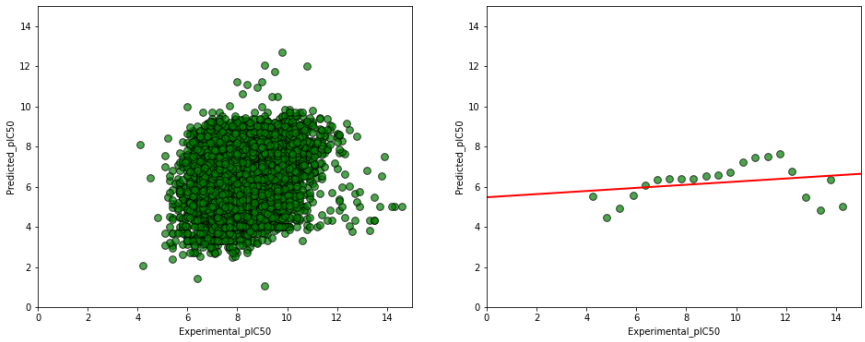


Figure 3.8: Virtual Screening

3.6 Molecular Visualization of the Docked Molecules

Molecular visualization aids in understanding the interactions between the amino acid, the chain in which it gets binded, the bonds created for the interaction, the types of bonds, and the angle. The interaction between the target and the ligand was visualized using PyMOL.

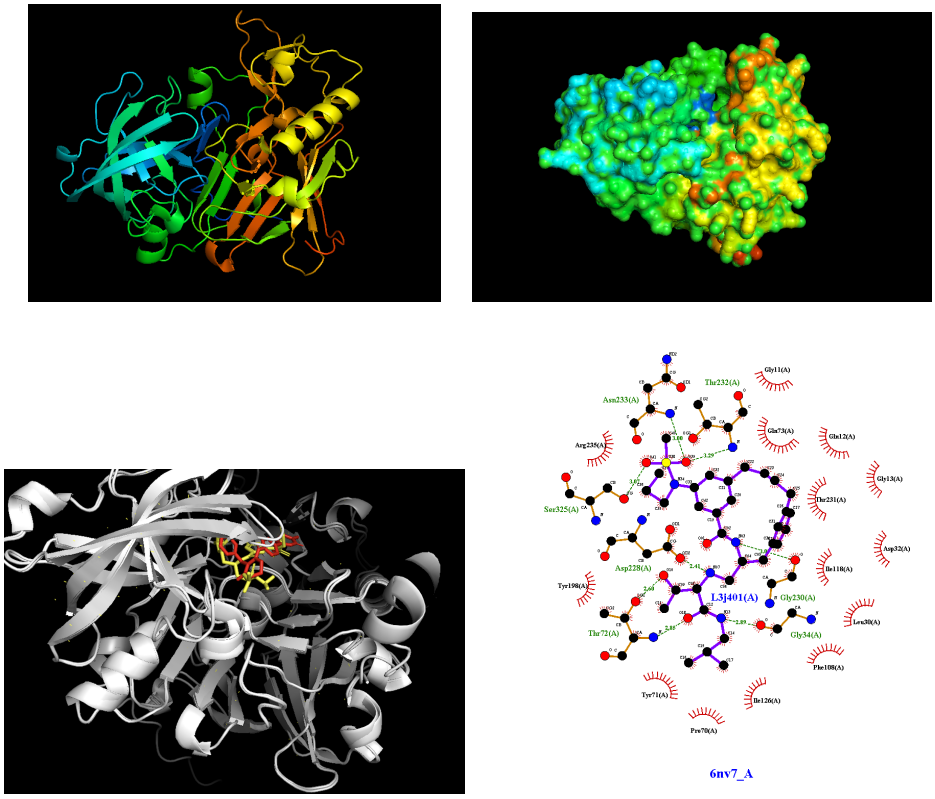


Figure 3.9: Molecular Visualization of the Docked Molecules

Chapter 4

Conclusion & Future Scope

4.1 Conclusion

In this thesis, titled "Assessment of Data-Driven and Deterministic Approaches for Predicting BACE-1 Binding Affinities," we investigated qualitative and quantitative BACE1 prediction models in chEMBL using features extracted from one-, two-, and three-dimensional structural information of compounds using a signaturizer, paDEL, and GCN architecture. We broadened the range of hyperparameters reported in the classification and regression tasks. Top best models rely on a non-linear relationship. Because of the non-linear relationship, more feature experimentation is required to produce a more model-ready data set.RF is an ideal application for this type of data.

In both qualitatively and quantitatively, based on the metrics of all approaches, the random forest model with 2d and 3d features outperforms other models with (R2 0.81) (RMSE 0.10).

As a result, when compared to graph-based techniques such as GCN, GAN, and AttentiveFP, the success of the machine-learning approach in predicting the binding affinity of hBACE-1 (human beta-site amyloid precursor protein cleaving enzyme 1) inhibitors provides a strong thrust for systematically applying such methods for drug screening and developing successful BACE-1 inhibitors. It should be noted, however, that the dataset's data preparation scheme and treatment of qualitative and quantitative measurements differ from those in our dataset.

The dataset, compiled from various sources, includes measurements and noise under various experimental conditions. The accuracy of activity prediction models has improved. For example, when multiple experimental values were available for the same compound-target pair, the maximum value was used. Other studies used median and mean values, implying that a different data pre-processing method could improve the

prediction performance of the models even more.

Even if the activity prediction target models are "unsatisfactory," they can provide valuable hints for drug discovery and development by directing researchers to likely off-targets, identifying polypharmacological drugs, and aiding in the search tool for target compounds.Qualitative classification and quantitative regression models also indicate that RF with all (1D, 2D, 3D, and fingerprint) descriptors achieves statistical accuracy, frequently requiring molecular alignment of diverse chemical scaffolds in one universal chemical space.The top 20 metabolites' binding energy is -9.9 kcal/mol. This demonstrates that these ligands have high binding affinities for BACE1.A model that ranks compounds quantitatively rather than qualitatively is an effective tool in drug discovery research. We believe our GCN architecture can help predict binding affinity.

4.2 Assumptions and Limitations

The model was trained using around 7432 compounds, a large amount nearly enough to represent the total no.of compounds in the entire chemical space.

Dataset's data preparation scheme - multiple experimental values were available for the same compound-target pair the maximum value was used. Other studies used median and mean values, implying that a different data pre-processing method could improve the model's prediction performance.

Using statistical measures and stepwise elimination relies on the statistical method to determine which compound is relevant for training the possibility of certain essential compounds being removed. This would have provided ample information to the model.

4.3 Future Scope

1. Predicting unknown binding affinity and processing molecular dynamic simulations
2. Using protein sequences to implement these pipelines for BACE-1 data.
3. Wet lab validation

Bibliography

[1] R. Vassar and S. Cole, "The Basic Biology of BACE1: A Key Therapeutic Target for Alzheimers Disease," *CG*, vol. 8, no. 8, pp. 509–530, Dec. 2007. [Online]. Available: http://www.eurekaselect.com/openurl/content.php?genre=article&issn=1389-2029&volume=8&issue=8&spage=509

[2] A. K. Ghosh and H. L. Osswald, "BACE1 (-secretase) inhibitors for the treatment of Alzheimer's disease," *Chem. Soc. Rev.*, vol. 43, no. 19, pp. 6765–6813, Apr. 2014. [Online]. Available: http://xlink.rsc.org/?DOI=C3CS60460H

[3] S. L. Cole and R. Vassar, "The Alzheimer's disease Beta-secretase enzyme, BACE1," *Mol Neurodegeneration*, vol. 2, no. 1, p. 22, 2007. [Online]. Available: http://molecularneurodegeneration.biomedcentral.com/articles/10.1186/1750-1326-2-22

[4] G. Subramanian, B. Ramsundar, V. Pande, and R. A. Denny, "Computational Modeling of -Secretase 1 (BACE-1) Inhibitors Using Ligand Based Approaches," *J. Chem. Inf. Model.*, vol. 56, no. 10, pp. 1936–1949, Oct. 2016. [Online]. Available: https://pubs.acs.org/doi/10.1021/acs.jcim.6b00290

[5] G. Dhamodharan and C. G. Mohan, "Machine learning models for predicting the activity of AChE and BACE1 dual inhibitors for the treatment of Alzheimer's disease," *Mol Divers*, vol. 26, no. 3, pp. 1501–1517, Jun. 2022. [Online]. Available: https://link.springer.com/10.1007/s11030-021-10282-8

[6] R. Singh, A. Ganeshpurkar, P. Ghosh, A. V. Pokle, D. Kumar, R. b. Singh, S. K. Singh, and A. Kumar, "Classification of beta-site amyloid precursor protein cleaving enzyme 1 inhibitors by using machine learning methods," *Chem Biol Drug Des*, vol. 98, no. 6, pp. 1079–1097, Dec. 2021. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1111/cbdd.13965

[7] S. John, S. Thangapandian, S. Sakkiah, and K. W. Lee, "Potent bace-1 inhibitor design using pharmacophore modeling, in silico screening and molecular docking studies," *BMC Bioinformatics*, vol. 12, no. S1, p. S28, Dec. 2011. [Online]. Available: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-12-S1-S28

[8] M. B. Kursa, A. Jankowski, and W. R. Rudnicki, "Boruta – A System for Feature Selection," *Fundamenta Informaticae*, vol. 101, no. 4, pp. 271–285, 2010. [Online]. Available: https://www.medra.org/servlet/aliasResolver?alias=iospress&doi=10.3233/FI-2010-288

[9] C.-H. Chang, C.-H. Lin, and H.-Y. Lane, "Machine Learning and Novel Biomarkers for the Diagnosis of Alzheimer's Disease," *IJMS*, vol. 22, no. 5, p. 2761, Mar. 2021. [Online]. Available: https://www.mdpi.com/1422-0067/22/5/2761

[10] A. Singh, "DeepChem and its importance in Drug Discovery and Biology," vol. 9, no. 4, 2012.

[11] M. Sakai, K. Nagayasu, N. Shibui, C. Andoh, K. Takayama, H. Shirakawa, and S. Kaneko, "Prediction of pharmacological activities from chemical structures with graph convolutional neural networks," *Sci Rep*, vol. 11, no. 1, p. 525, Jan. 2021. [Online]. Available: https://www.nature.com/articles/s41598-020-80113-7

[12] M. García-Ortegón, G. N. C. Simm, A. J. Tripp, J. M. Hernández-Lobato, A. Bender, and S. Bacallado, "DOCKSTRING: easy molecular docking yields better benchmarks for ligand design," Oct. 2021, number: arXiv:2110.15486 arXiv:2110.15486 [cs, q-bio, stat]. [Online]. Available: http://arxiv.org/abs/2110.15486