Machine Learning Based Design Of Multi-targeting Inhibitors for Alzheimer Disease and Structure Based Validation

By
Virendra Kumar

Under the supervision of
Dr. N. Arul Murugan

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

Center for Computational Biology
IndraprasthaInstitute of Information Technology - Delhi June, 2024

Certificate

This is to certify that the thesis titled *"ML based design of multi-targeting inhibitors for Alzheimer Disease and structure based validation"* being submitted by **Virendra Kumar** 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 anyother university or institute for the award of any degree/diploma.

June, 2024

Dr. N. Arul Murugan

Department of Computational Biology
Indraprastha Institute of Information Technology Delhi

New Delhi 110 020

Acknowledgements

I would like to thank Prof. N. Arul Murugan for his continuous guidance and mentoring, without him this work would not have been possible. From day one of starting mythesis, 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 wasstruck 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 father, 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 and PhD specially to Prateek Paul, 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 staff and faculty members of the Department of Computational Biology and IIIT Delhi for always helping us throughout our college journey.

Abstract

Abstract
Neurodegenerative diseases are one of the major problems in today's world, and Alzheimer's is one of them. The disease-associated symptoms are not diagnosed earlier, and also very few FDA-approved disease-reversing drugs are available on the market for the disease. The Machine learning approach in the field of drug discovery and design is increasing rapidly. The objective of this study is creating a Machine Learning models to predict the pIC50 value of chemical compounds and find out the multi-targeting drugs against the promising targets of Alzheimer's disease, i.e., AChE, CDK5, GS, and GSK-3. To build a machine learning model, the dataset we use contains the SMILES of chemical compounds and their corresponding IC50 values (for aforementioned targets) and was downloaded from ChEMBL. We generate the features or descriptors using RDkit. Regression models like multiple linear regression, linear regression, SVR, polynomial regression, and RF on 2D and all descriptors (1D, 2D, and 3D) have been used to predict the pIC50 value. By testing all the models, we see the best performance is given by the RF with an R- squared value of 0.61 for targets GS and GSK3 and SVR with an R- squared value of 0.65 and 0.72 for targets AChE and CDK5 respectively for 2D descriptors. For all descriptors, the SVR shows the best performance with an R-squared value of 0.68, 0.69, 0.62 and 0.67 for targets AChE, CDK5, GS and GSK3 respectively. Further prediction is done on drug databases like DrugBank. The prediction shows that Arundic acid, Colocoxib, Chlorphenoxamine, Clomocycline and Raloxifene from DrugBank may be the best candidates for the selected targets, and the structure-based validation is done by using molecular docking and the computed binding affinity shows that indeed these compounds can potentially inhibit these targets.

Keywords: Alzheimer's disease, AChE, CDK5, GS, GSK-3, Arundic acid, Colocoxib, Chlorphenoxamine, Clomocycline and Raloxifene.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Alzheimer’s disease

Neurodegenerative diseases encompass a range of conditions resulting in the death of or degeneration nerve cells, leading to issues with movement (ataxia) or mental functioning (dementia). Dementia entails a decline in cognitive abilities such as reasoning and memory, ultimately affecting daily activities. The different types of dementia are:

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

Alzheimer's disease (AD) has the highest frequency of all neurological illnesses. It is an unrelenting thief of brain function, characterized by a slow but steady decrease. The subtle accumulation of plaques and tangles in the brain, as well as the loss of vital synapses—connections between nerve cells—and persistent inflammation—are the hallmarks of this deterioration. In Alzheimer's, damage to brain tissue inhibits the transmission of messages through chemical transmitters, resulting in disease symptoms. Early signs typically include a gradual decline in cognitive function coupled with short-term memory loss, neuronal loss, and degeneration in cerebral cortex synapses and certain subcortical regions, leading to gross atrophy in affected areas such as the temporal, parietal, cingulate, and frontal cortex (1).

Amnestic cognitive impairment is a common feature of AD presentations, but non-amnestic cognitive impairment can also occur and impact cognitive abilities including executive function, visuospatial processing, and expressive speech. The development of AD is complicatedly influenced by genetics, even though the majority of cases are not hereditary. (2).

Despite the extensive research on Alzheimer's, the focus has primarily been on a limited range of brain areas, particularly concerning neurofibrillary tangle (NFT) pathology. Despite the broad scope of research efforts across various institutions and disciplines, there is still a narrow focus on investigating brain regions affected by NFT pathology (3).

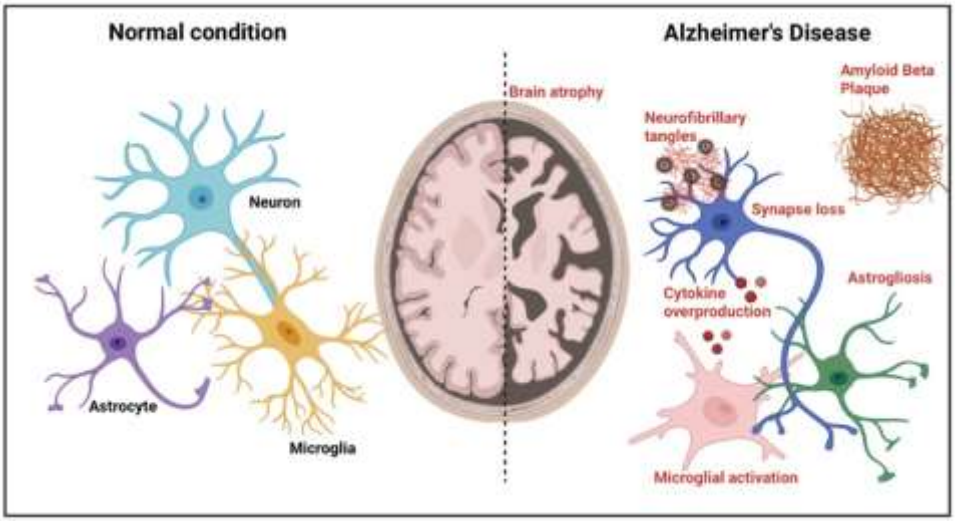


Figure_1: A diagram illustrating the pathological changes in AD brains instead of healthy brains reveals evident brain atrophy resulting from neuronal loss at the macroscopic level. Microscopically, key observations include the presence of harmful amyloid-β, the formation of intra neuronal neurofibrillary tangles, synaptic depletion, microglial activation, astrogliosis, excessive cytokine production, and neurite dystrophy (4).

1.2 Identifying the different Targets

By studying different research articles and paper the selected target for Alzheimer's disease are as follows:

PROTEINS
1. Apolipoprotein E4 (APOE ε4) (5)
2. Tau (6)
3. Presenilin 2 (PSEN2) (7)
4. Presenilin 1 (PSEN1) (8)
5. Presenilin enhancer 2 (9)
6. Nicastrin (NCSTN) (10)
7. Amyloid precursor protein (APP) (11)
8. ATP-binding cassette subfamily A member 7 (ABCA7) (12)
9. Encodes Synaptophysin (p38)/ synaptic vesicle protein p38(SYP) (13)
10. Myelin basic Protein (MBP) (14)
11. Neurogranin/Ras-specific guanine nucleotide-releasing factor 3 (RC3)/ p17 (15)
12. Ubiquitin (16)
13. Calmodulin (17)
14. α-synuclein (18)
15. Calbindin D28K (17)
16. Synaptosomal-associated protein 25 (SNAP-25) (19)
17. Postsynaptic density protein (PSD-95) (20)
18. Proteasome (21)
19. Peroxiredoxins (PRDX1, PRDX2, PRDX3, PRDX4, and PRDX5) (22)
20. Hsp90, Hsp40, Hsp70 (HSP: Heat Shock Protein) (23)
NEUROFILAMENT PROTEINS

21. NF-L and NF-M (24)
RECEPTORS
22. Receptor for advanced glycation end products (RAGE) (25)
23. Triggering Receptor Expressed on Myeloid Cells 2 (TREM2) (26)
24. Alpha-7 nicotinic acetylcholine receptor (α7nAChR) (27)
25. M1 muscarinic acetylcholine receptor (28)
26. Glutamate Receptor [**(α-amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid receptor (**AMPAR Receptor**)** and N-methyl-D-aspartate Receptor(NMDAR)] (29)
27. Sortilin-related receptor (SorLA 1) (30)
28. p75 neurotrophin receptor (p75NTR) (31)
29. Low-density lipoprotein receptor-related protein (LRP) (32)
30. Seratonine Receptors (33)
TRANSPORTERS
31. Glutamate transporter EAAT2 (34)
32. Serotonin binding (35)
33. Imipramine binding (associated with serotonin transporter) (36)
NEUROTRANSMITTERS
34. Glutamate (37)
35. Acetylcholine (38)
ENZYMES
36. Choline acetyltransferase (39)
37. Glucocerebrosidase (GBA) (40)
38. Metallothionein-I, -II (41)
39. Cystatin C (42)
40. Cyclin dependent Kinase-5 (CDK-5) (43)
41. p38 MAPK (Mitogen-activated protein kinase) (44)
42. Ubiquitin carboxy-terminal hydrolase L1 (UCH-L1) (45)
43. Superoxide dismutase 2 (SOD2) (46)
44. Insulin-degrading enzyme (IDE) (47)
45. Glycogen synthase kinase 3 (GSK-3) (48)
46. Glycogen synthase kinase 3 beta (49)
GLYCOPROTEIN
47. Clusterin (CLU) or APOJ (50)
INFLAMATORY FACTORS
48. IL-1β (51)
49. IL-6 (52)
50. TNF-α (53)
51. C-reactive protein (CRP) (52)
GROWTH FACTORS
52. Vascular endothelial growth factor (VEGF) (53)

Table-1: Targets for Alzheimer's disease

1.3 Selected targets for our study

After studying different research paper we selected main four inhibitors or targets for our research work. The inhibitors or targets are as follows:

1. Acetylcholinesterase (AChE)
2. Cyclin dependent Kinase-5 (CDK-5)
3. Gamma Secretase (γ-secretase)
4. Glycogen Synthase Kinase-3 (GSK-3)

1.3.1 Acetylcholinesterase (AChE)

Acetylcholinesterase is part of the alpha/beta hydrolase fold protein superfamily, which includes cholinesterase, carboxylesterases, and lipases, sharing structural homology. Its primary role involves swiftly breaking down acetylcholine at the synapse and neuromuscular junction, thus halting nerve impulses. Scientists first figured out the 3D shape of Acetylcholinesterase (AChE) in 1991. They studied an enzyme from the electric ray, Torpedo californica (Tc), using a combination of techniques that looked closely at its structure and made specific changes to its amino acids (site-directed mutagenesis). This revealed that AChE has a deep, narrow trench with two separate areas where molecules can attach (54).

Medically, Acetylcholinesterase (AChE) inhibitors are primarily used to treat myasthenia gravis, glaucoma, Alzheimer's disease, and muscular weakness after surgery. The two kinds of enzymes that degrade acetylcholine in vertebrates are butyrylcholinesterase (BuChE) and acetylcholinesterase (AChE). Both function quite well, however compared to BuChE, AChE is more sensitive to medication blockade (inhibitors), is present in distinct organs, and has a predilection for particular molecules (substrate specificities) (**55**).

Research on the function of acetylcholine (Ach) in Alzheimer's dementia (AD) has been extensive. The cholinergic hypothesis postulates that AD brains have an overactive form of the enzyme acetylcholinesterase (AChE). AChE degrades Ach, an essential neurotransmitter. It is believed that this collapse has a role in the cognitive and memory loss associated with AD. Surprisingly, research indicates that AChE may have a more complicated function than only degrading Ach. In addition, it may have a role in the development of the aberrant protein cluster in Alzheimer's patient. Research indicates that AChE facilitates the interaction between beta-amyloid fragments and promotes their clumping, which may lead to more damage to brain cells in AD (**56,57**).

Different functions of AChE are as follows:

- **Cholinergic Hypothesis:** Acetylcholinesterase (AChE) is a pivotal element of the cholinergic hypothesis regarding Alzheimer's disease (AD), positing that cognitive decline is linked to the degeneration of cholinergic neurons. Acetylcholine is broken

down by AChE and is used by these neurons as a neurotransmitter and in post-synaptic transmission**.**

- **An imbalance of neurotransmitters**: Acetylcholine, a neurotransmitter that is critical for memory and learning, is lacking in Alzheimer's disease (AD). The breakdown of acetylcholine by acetylcholinesterase (AChE) is thought to be the cause of this deficiency.
- **Cholinesterase:** Drugs used to treat Alzheimer's disease (AD) include galantamine, rivastigmine, and donepezil. These drugs operate by blocking the activity of AChE, temporarily elevating acetylcholine levels within the brain. For specific individuals with AD, this mechanism assists in enhancing or maintaining cognitive functions.
- **Cognitive Function:** The enzymatic breakdown of acetylcholine by AChE affects cognitive functions, and this enzyme's activity is linked to the memory and learning impairments observed in Alzheimer's disease (AD).

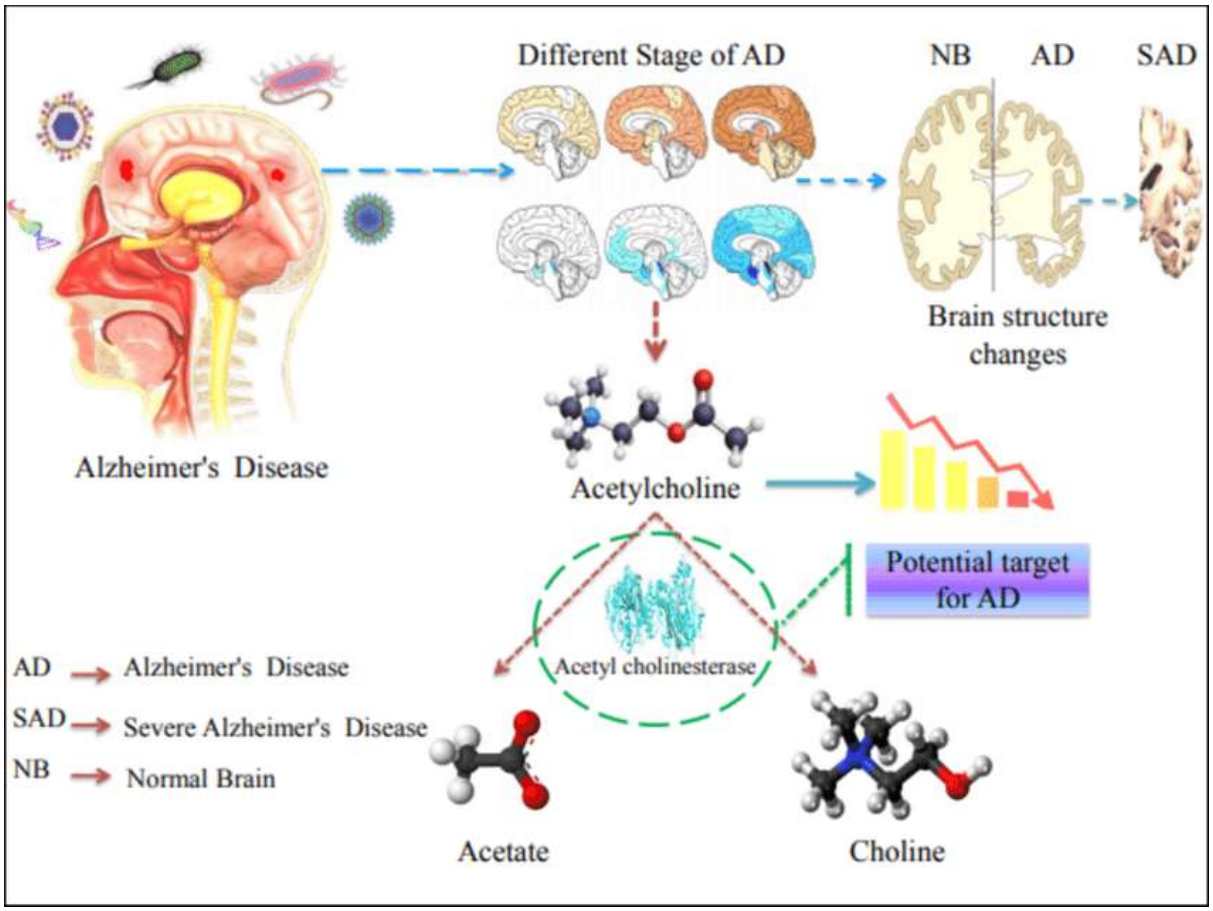


Figure_2: Brief description of Alzheimer's disease (AD) (58).

1.3.2 Cyclin dependent Kinase-5 (CDK5)

The multifunctional kinase Cdk5 is essential for both apoptosis and neuron survival. It needs its activator, p35, in psychological circumstances to perform various tasks, including memory consolidation, learning, and cerebellar development, as well as neuronal activities like migration, differentiation, and neurotransmission. Nonetheless, increased calcium levels in neurotoxic environments can cause the cleavage of p35 into p25 and p10 mediated by calpain.

The resultant p25 may active Cdk5 independently and is more stable but neurotoxic. This creates a stable complex of p25/Cdk5, which triggers pathological events, including the build-up of Aβ, tau phosphorylation, dysfunctional synapses, apoptosis, reactivation of the cell cycle, and oxidative stress, ultimately leading to the death of neurons (59).

Research shows that a biological mechanism known as calpain-mediated cleavage is capable of cleaving the protein p35 into a smaller component termed p25. This disintegration is seen in a variety of neurodegenerative situations, such as ischemic brain injury, oxidative stress, excitotoxicity, overstimulation of brain cells, and beta-amyloid peptide therapy of cultured neurons. Fascinatingly, studies indicate that phosphorylating p35 to connect phosphate group's functions as a barrier against calpain cleavage. It appears that this phosphorylation redirects p35 toward degradation via an alternative cellular process involving entities known as proteasomes (60).

Different functions of CDK 5 are as follows:

- **Hyperphosphorylation of Tau:** CDK5 has the ability to phosphorylate the Neurofibrillary tangles, tau protein., a defining feature of Alzheimer's disease, are produced when tau is abnormally hyperphosphorylated.

- **Synaptic dysfunction:** It is possible for CDK5 to phosphorylate the tau protein. When tau is inappropriately hyperphosphorylated, neurofibrillary tangles, a characteristic of AD, are created.

- **Amyloid Precursor Protein (APP):** CDK5 affects how APP is processed and how beta-amyloid peptides are made as a result. The accumulation of beta-amyloid is a characteristic that distinguishes Alzheimer's disease.

- **Neuronal cell death:** Neuronal cell death can be a consequence of dysregulated CDK5 activity, which is commonly seen in other neurodegenerative disorders and Alzheimer's disease.

- **Therapeutic target:** Scientists are looking at CDK5 as an important target for treatment. Reduced tau hyperphosphorylation and synaptic dysfunction may be achieved by altering CDK5 activity as a possible treatment for Alzheimer disease.

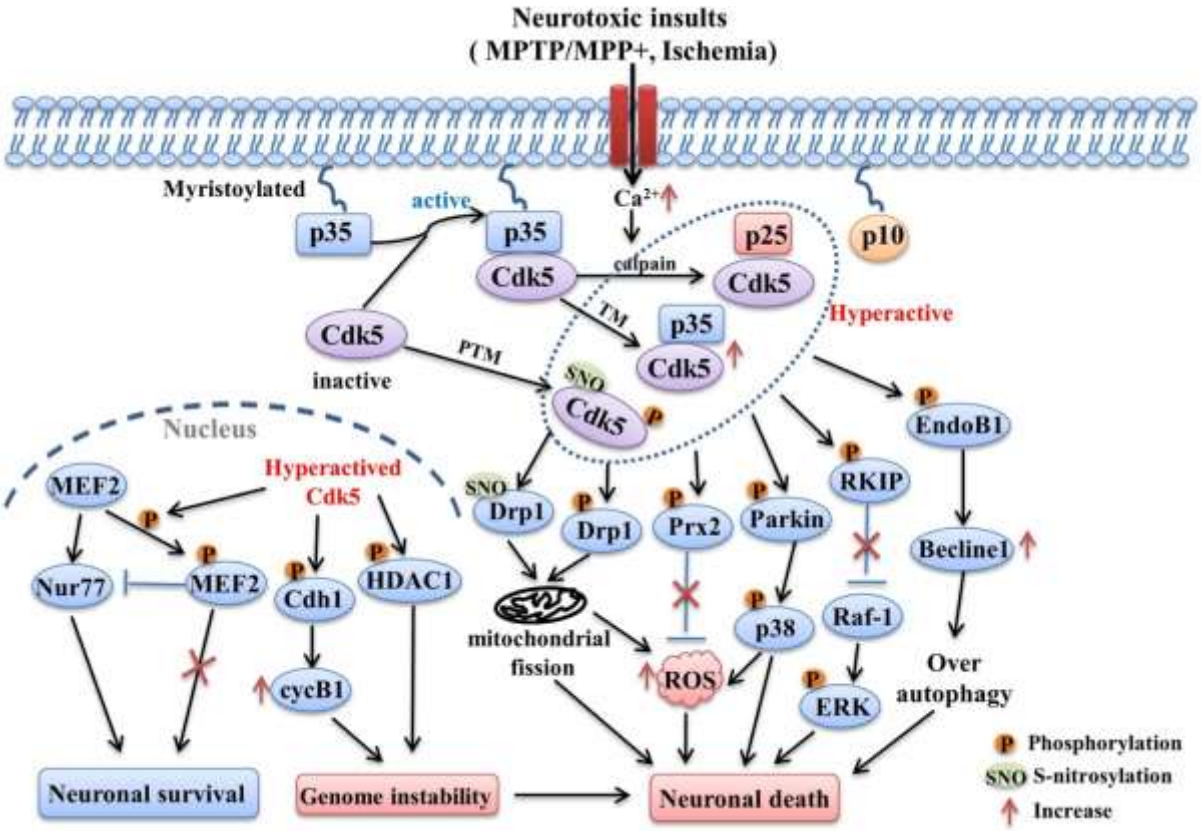


Figure_3: The pathways linked to Cdk5 are implicated in both stroke and Alzheimer's disease. When exposed to neurotoxic substances, the extended rise of calcium levels causes an overabundance of calpain activity, which causes p35 to break into fragments of p25 and p10. Due to p25's excellent stability over p35, it may form a more durable complex with Cdk5, raising Cdk5 activity. Additionally, its activity may be impacted by post-translational changes that further augment Cdk5 activation, such as S-nitrosylation at Cys83 and Cys157 and phosphorylation at certain locations like Ser159, furthermore, through specific transcription factors like Fos and CREB, transcriptional changes can increase the expression and activity of Cdk5 and p35 respectively (60).

1.3.3 Gamma Secretase (γ-secretase)

An enzyme complex called gamma secretase plays a critical role in chopping up a protein called amyloid precursor protein (APP). This chopping process is linked to the development of Alzheimer's disease. It leads to the formation of short protein fragments called amyloid-beta peptides, including the particularly harmful Aβ42 variant.

The C-terminus (end) of a protein known as amyloid precursor protein (APP) is cleaved by the enzyme gamma-secretase. After prior processing by alpha- or beta-secretase, this cleavage releases a fragment of APP into the cell. Prior to gamma-secretase's final identification, researchers noticed that it could cleave amyloid-beta peptide (APP) at different C-terminal sites, producing distinct forms of the protein. Remarkably, presenilin 1 and 2 (PS-1 and PS-2) proteins appear to control its function. These proteins affect the precise location on APP where gamma-secretase is cut (61).

A possible target for Alzheimer's disease (AD) therapy is the enzyme γ-secretase, which cleaves the C99 substrate sequentially to create Aβ peptides. Nevertheless, γ-secretase targeting has been difficult, and further study is needed to understand its processes completely. Nicastrin, Pen-2, presenilin enhancer-2 (PS) and anterior pharynx defective-1 (Aph-1), comprise the transmembrane protein complex known as γ-Secretase. Membrane lipid bilayers are where γ-secretase, an intramembrane-cleaving protease (I-CliP), performs its specific cleavage (62).

Different functions of Gamma secretase are as follows:

- **Amyloid-beta Accumulation:** Beta-amyloid peptides abnormally build up in Alzheimer's disease patients' brains to form plaques. Gamma-secretase plays a critical part in this process by cleaving the amyloid precursor protein (APP) and releasing amyloid-beta peptides, which can collect and aid in developing these dangerous plaques.

- **Genetic Associations:** Alzheimer's disease is linked to mutations in gamma-secretase components, including PSEN1 and PSEN2. These mutations cause the overproduction of longer, more hazardous amyloid-beta peptides.

- **Therapeutic Focus:** The development of drugs to treat Alzheimer's disease has targeted gamma-secretase. Nevertheless, several obstacles have been overcome in creating safe and efficient gamma-secretase inhibitors for clinical use

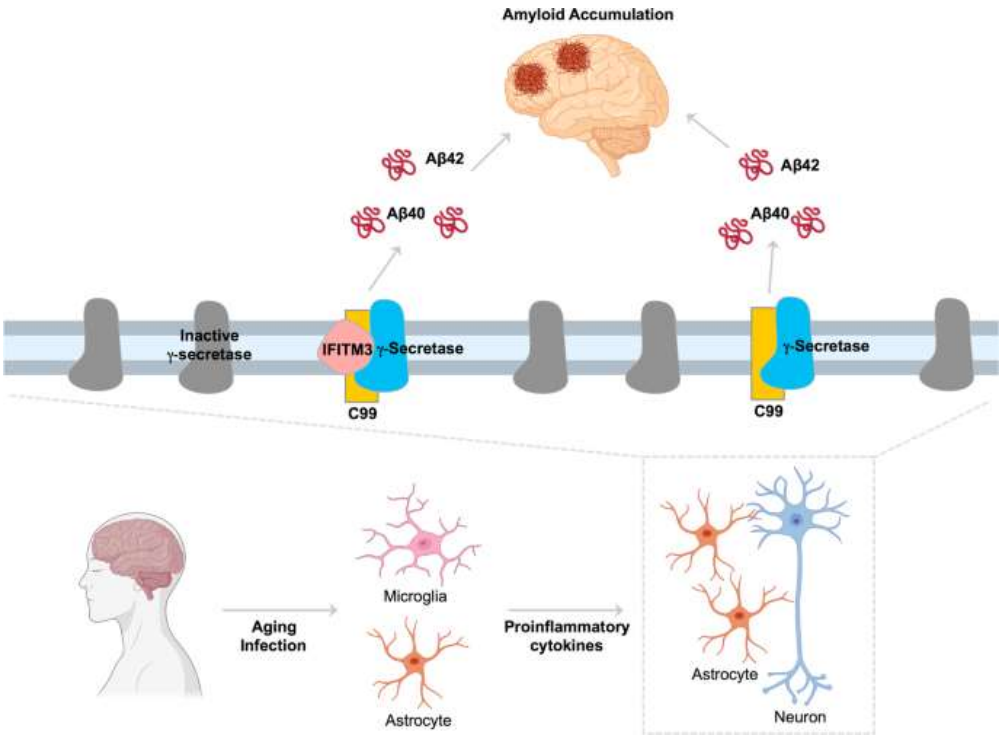


Figure_4: Aβ production by IFITM3-γ-secretase complexes (61).

1.3.4 Glycogen Synthase Kinase-3 (GSK-3)

GSK-3 is an enzyme that is universal in protein kinases and is present in all eukaryotic (cell-containing) species. GSK-3 occurs in two primary flavors: GSK-3beta and GSK-3alpha. GSK-3 was first discovered as one of multiple protein kinase activity in muscle cell extract samples. Notwithstanding their differences in their N- and C-terminal sections, these isoforms which are encoded by distinct genes on separate chromosomes share many characteristics. What's interesting is that GSK-3b2, a distinct variation of GSK-3beta can be produced by an alternate splicing event (63).

Tau proteins that have been hyperphosphorylated, forming intracellular neurofibrillary tangles and the existence of extracellular senile plaques made of β-amyloid have been proposed as Alzheimer's disease has two main pathological features, and GSK-3 plays a critical role in connecting them. Tau protein in neurofibrillary tangles is modified at several locations by GSK-3, one of the central tau kinases (64).

GSK-3 can also react to this peptide and affect how much amyloid-β is produced. Higher levels of GSK-3 have been linked to tau hyperphosphorylation, neuronal death, and cognitive impairment in several transgenic animals. Tau phosphorylation has been demonstrated to be prevented by clinically relevant amounts of lithium, a drug that is frequently administered for mood disorders. This prescription inhibits GSK-3 (64).

Different functions of Gamma secretase are as follows:

- **Tau Phosphorylation:** One prominent feature of AD is the aberrant build-up of tau protein into neurofibrillary tangles. Tau protein is phosphorylated by GSK-3, which causes it to aggregate into tangles. The loss of normal function in hyperphosphorylated tau accelerates neuronal structure deterioration and slows the aging process.

- **Amyloid Precursor Protein (APP):** GSK-3 is involved in APP processing. Another characteristic of AD is the development of beta-amyloid plaques, which can result from dysregulated processing of APP. GSK-3, affects the synthesis of beta-amyloid by phosphorylating proteins involved in APP processing.

- **Neuroinflammatory and Oxidative Stress:** Research suggests GSK-3 plays a role in regulating how the nervous system responds to inflammation and oxidative stress. These responses are unfortunately linked to the development of neurodegenerative diseases, including AD.

- **Synaptic Dysfunction:** GSK-3 activity is associated with synaptic dysfunction, a crucial factor in the development of AD. Memory loss is a frequent sign of AD caused by disruption of synaptic connections between neurons.

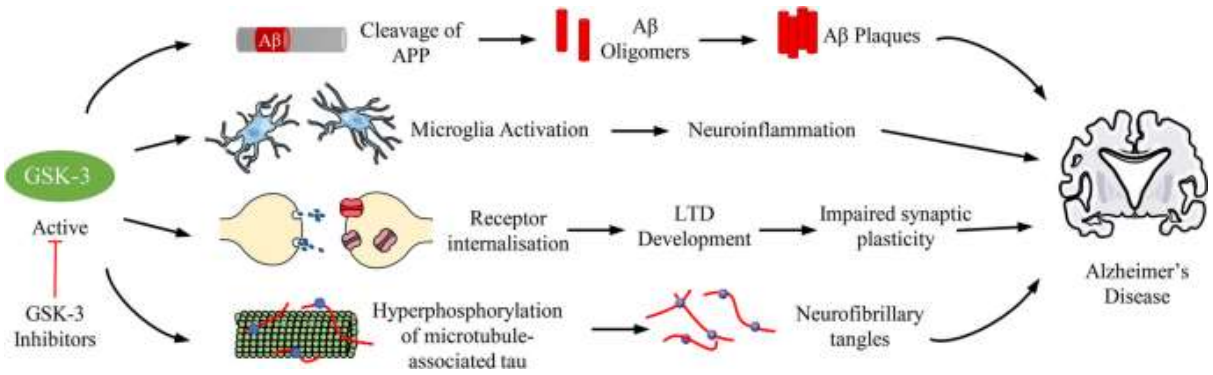
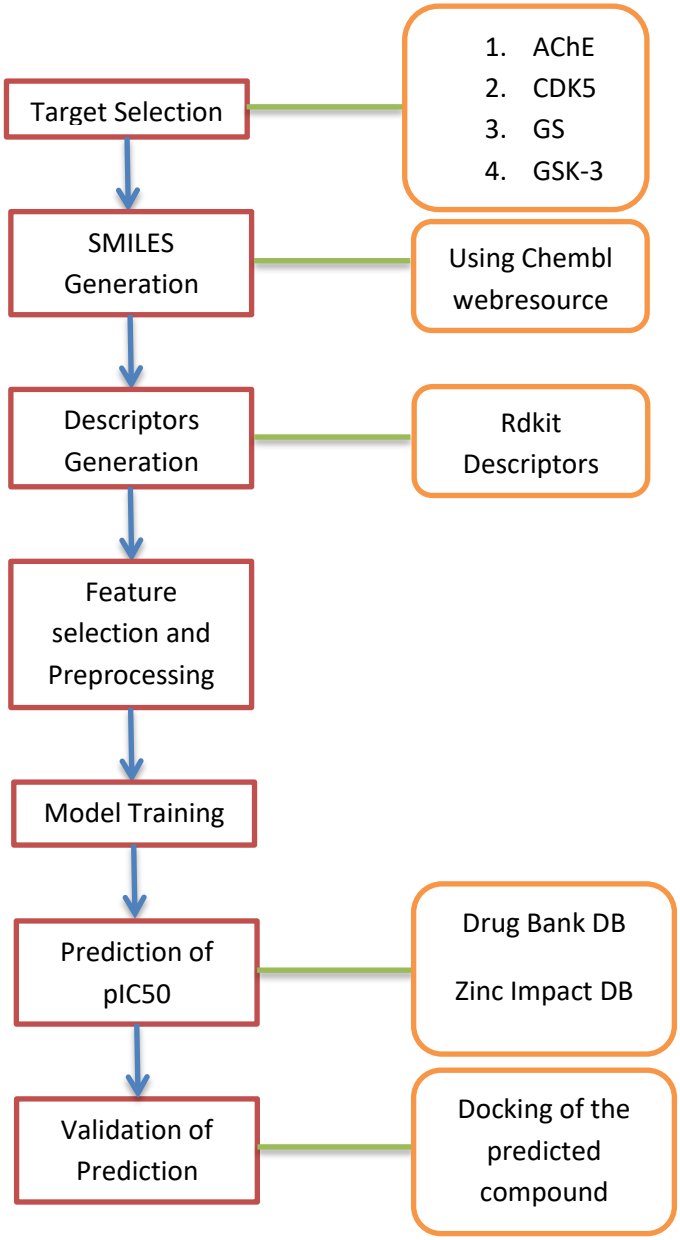


Figure_5: Graphical representation of GSK-3 Inhibitors functions in Alzheimer's disease (65).

1.4 Workflow


Target Selection
1. AChE
2. CDK5
3. GS
4. GSK-3
SMILES Generation
Using Chembl webresource
Descriptors Generation
Rdkit Descriptors
Feature selection and Preprocessing
Model Training
Prediction of pIC50
Drug Bank DB
Zinc Impact DB
Validation of Prediction
Docking of the predicted compound

Chapter 2

Data Collection

2.1 Generating SMILES

2.1.2 SMILES

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. This notation is commonly employed in entering chemical structures into estimation programs like EPI Suite™ and ECOSAR. Further examples and guidance on SMILES notation can be found in the HELP files of these programs. Various software tools exist to translate chemical structures into SMILES format (66).

Examples:

- Aspirin: SMILES: O=C(C)Oc1ccccc1C(=O)O
- Paracetamol (Acetaminophen): SMILES: CC(=O)Nc1ccc(O)cc1
- Morphine: SMILES: CN1CC[C@]23c4c5ccc(O)c4O[C@H]2C@@HO

2.1.2 Steps for generating the SMILES (Inhibitors)

SMILES are generated by using ChEMBL web resource.

- !pip install chembl-webresource-client

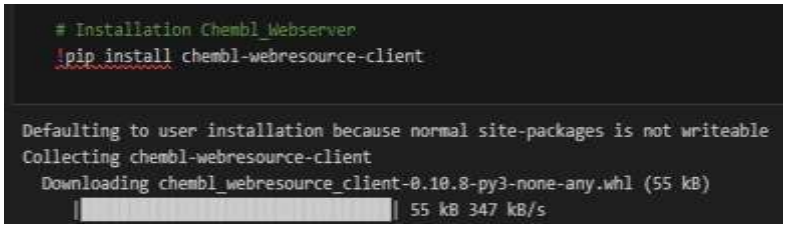


Figure_6: Installation of chembl_webresource

- Importing the libraries
- Different target inhibitors search

Figure_7: Target inhibitor search script

- Filter targets for Homo sapiens

```
# Filter targets for Homo sapiens
selected_targets = targets[targets['organism'] == 'Homo sapiens']
selected_targets
```

	cross_references	organism	pref_name	score	species_group_flag	target_chembl_id
1	[{'xref_id': 'Q00535', 'xref_name': None, 'xre...	Homo sapiens	Cyclin-dependent kinase 5	10.0	False	CHEMBL4036

Figure_8: Filter targets for *Homo sapiens*

- Retrieve the ChEMBL ID of the selected target
- Filter activities for the selected target and IC50 values

```
# Filter activities for the selected target and IC50 values
activity = new_client.activity
res = activity.filter(target_chembl_id=selected_target).filter(standard_type="IC50")
res
```

Figure_9: Script for IC50 generation

- Convert the response to a pandas DataFrame
- Filter for inhibitors

```
# Filter for inhibitors
inhibitor_df = df[df['assay_description'].str.contains('nhibit')]
inhibitor_df
```

	action_type	activity_comment	activity_id	activity_properties	assay_chembl_id	assay_description
0	None	None	72750	[]	CHEMBL663602	Inhibitory activity against Cyclin-dependent k...

Figure_10: Filtering the data on the basis of Inhibitors

- Rename the columns
- Save the data to a CSV file

The above steps are followed for all the targets, i.e., AChE, CDK5, GS, and GSK-3, for generating the SMILES, and only inhibitors SMILES of *Homo sapiens* are taken for developing the descriptors in the next step.

2.1.3 IC50 and pIC50

IC50:

The IC50 represents the drug concentration needed to inhibit 50% of a particular biological activity, and the assay conditions influence its value. In cases where complete inhibition is desired, IC90 or IC99 may be used. Calculations of fractional occupancy reveal that the IC90 concentration is roughly ten times higher than the IC50 concentration when assuming equilibrium binding at a single site with a Hill coefficient of 1. Similarly, the IC99 concentration is approximately 100 times greater than the IC50 concentration (67).

pIC50:

The pIC50 factor, or negative logarithm of the half maximum inhibitory concentration (IC50), is critical in medical assessment (68).

2.1.4 Converting IC50 to pIC50

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

pIC50_value = 9 - math.log10(IC50_value)

2.1.5 Final Targets table for descriptors generation

S. No	Targets Name	SMILES Count
1	Acetylcholinesterase	8728
2	Cyclin dependent Kinase-5 (CDK-5)	2487
3	Gamma Secretase	3018
4	Glycogen Synthase Kinase-3	4481

Table-2: Targets after SMILES generation

2.1.6 Tables of Isomeric SMILES and pIC50 value for different Targets

ChEMBL_ID	Isomeric_SMILES	IC50	pIC50
CHEMBL133897	CCOc1nn(-c2cccc(OCc3ccccc3)c2)c(=O)o1	750	6.12494
CHEMBL336398	O=C(N1CCCCC1)n1nc(-c2ccc(Cl)cc2)nc1SCC1CC1	100	7
CHEMBL131588	CN(C(=O)n1nc(-c2ccc(Cl)cc2)nc1SCC(F)(F)F)c1ccccc1	50000	4.30103
CHEMBL130628	O=C(N1CCCCC1)n1nc(-c2ccc(Cl)cc2)nc1SCC(F)(F)F	300	6.52288
CHEMBL130478	CSc1nc(-c2ccc(OC(F)(F)F)cc2)nn1C(=O)N(C)C	800	6.09691
CHEMBL130112	CSc1nc(-c2ccc(C)cc2)nn1C(=O)N(C)c1ccccc1	2400	5.61979
CHEMBL130098	CSc1nc(-c2ccc(Cl)cc2)nn1C(=O)N(C)C	100	7
CHEMBL337486	CCCCCCSc1nc(-c2ccc(Cl)cc2)nn1C(=O)N1CCOCC1	50000	4.30103
CHEMBL336538	COc1ccc(-c2nc(SC)n(C(=O)N(C)C)n2)cc1	800	6.09691
CHEMBL131051	CSc1nc(-c2ccc(OC(F)(F)F)cc2)nn1C(=O)N(C)c1ccccc1	50000	4.30103
CHEMBL341437	CCSc1nc(-c2ccc(OC)cc2)nn1C(=O)N1CCOCC1	50000	4.30103
CHEMBL335033	CSc1nc(-c2ccc3ccccc3c2)nn1C(=O)N(C)C	50	7.30103
CHEMBL122983	C[C@H]1C(=O)N(C(=O)NCc2ccccc2)[C@@H]1Oc1ccc(C	100000	4
CHEMBL338720	CSc1nc(-c2ccc(-c3ccccc3)cc2)nn1C(=O)N(C)C	560	6.25181
CHEMBL339995	CSc1nc(/C=C/c2ccccc2)nn1C(=O)N(C)C	10	8
CHEMBL335158	CCCCCCSc1nc(-c2ccc(Cl)cc2)nn1C(=O)N1CCCCC1	10	8
CHEMBL131536	CSc1nc(-c2ccc(Cl)cc2)nn1C(=O)N(C)c1ccccc1	1400	5.85387
CHEMBL106126	Cc1c(C(C)C)c(=O)on1C(=O)N1CCC[C@H](C)C1	17000	4.76955
CHEMBL334971	CCSc1nc(-c2ccc(OC)cc2)nn1C(=O)N(C)c1ccccc1	1100	5.95861

Table-3: AChE (SMILES, pIC50)

ChEMBL_ID	Isomeric SMILES	IC50	pIC50
CHEMBL440411	O=C1Nc2ccc(S(=O)(=O)O)cc2/C1=C1/Nc2ccccc2/C1=N\O	7	8.154902
CHEMBL337300	O=Nc1c(-c2c(O)[nH]c3ccc(I)cc23)[nH]c2ccccc12	20	7.69897
CHEMBL261425	O=C1Nc2ccccc2/C1=C\c1c(Cl)[nH]c2ccccc12	25000	4.60206
CHEMBL259850	COc1ccc(CNC(=O)Nc2ncc([N+](=O)[O-])s2)cc1	100000	4
CHEMBL126077	O=Nc1c(-c2c(O)[nH]c3ccccc23)[nH]c2ccccc12	100	7
CHEMBL1933576	CN1CCN(c2ccc(OC(F)(F)F)c(Nc3nccc(-c4cc(C(N)=O)cn4C)n3)c	10000	5
CHEMBL1933582	CN1CCN(c2ccc(OC(F)(F)F)c(Nc3nccc(-c4cc5c(n4C)CCNC5=O)n	10000	5
CHEMBL174068	CC(=O)O/N=C1/C(c2c(O)[nH]c3cc(Cl)ccc23)=Nc2ccccc21	200	6.69897
CHEMBL435707	Cn1c(O)c(C2=Nc3ccccc3C2=O)c2ccccc21	100000	4
CHEMBL176904	O=C1C(c2c(O)[nH]c3cc(Br)ccc23)=Nc2ccccc21	53000	4.275724
CHEMBL173040	O=Nc1c(-c2c(O)[nH]c3cc(Cl)ccc23)[nH]c2ccccc12	100	7
CHEMBL176896	CC(=O)O/N=C1/C(c2c(O)[nH]c3ccccc23)=Nc2ccccc21	700	6.154902
CHEMBL173171	Cn1c(O)c(-c2[nH]c3ccccc3c2N=O)c2ccc(Br)cc21	100000	4
CHEMBL354276	CO/N=C1/C(c2c(O)[nH]c3cc(Br)ccc23)=Nc2ccccc21	2200	5.657577
CHEMBL368954	C=Cc1ccc2c(C3=Nc4ccccc4/C3=N\OC(C)=O)c(O)[nH]c2c1	400	6.39794
CHEMBL172547	CC(=O)O/N=C1/C(c2c(O)[nH]c3cc(Br)c([N+](=O)[O-])cc23)=N	31000	4.508638
CHEMBL353554	CC(=O)O/N=C1/C(c2c(O)[nH]c3cc(I)ccc23)=Nc2ccccc21	1300	5.886057
CHEMBL174292	O=Nc1c(-c2c(O)[nH]c3cc(F)ccc23)[nH]c2ccccc12	150	6.823909

Table-4: CDK5 (SMILES, pIC50)

ChEMBL_ID	Isomeric_SMILES	IC50	pIC50
CHEMBL336071	C[C@@H](Cc1ccc(F)c(F)c1)C(=O)N[C@H]1N=C(c2	1.8	8.744727
CHEMBL336813	C[C@@H](Cc1ccc(F)c(F)c1)C(=O)N[C@H]1N=C(c2	0.9	9.045757
CHEMBL334594	C[C@@H](Cc1ccc(F)c(F)c1)C(=O)N[C@H]1N=C(c2	13.2	7.879426
CHEMBL133237	C[C@@H](Cc1ccc(F)c(F)c1)C(=O)N[C@H]1N=C(c2	1.2	8.920819
CHEMBL132900	C[C@@H](Cc1ccc(F)c(F)c1)C(=O)N[C@H]1N=C(c2	1884	5.724919
CHEMBL423533	C[C@@H](Cc1ccc(F)c(F)c1)C(=O)N[C@H]1N=C(c2	3.6	8.443697
CHEMBL133857	C[C@@H](Cc1ccc(F)c(F)c1)C(=O)N[C@H]1N=C(c2	0.2	9.69897
CHEMBL133235	C[C@@H](Cc1ccc(F)c(F)c1)C(=O)N[C@H]1N=C(c2	28	7.552842
CHEMBL130093	C[C@@H](Cc1ccc(F)c(F)c1)C(=O)NC1N=C(c2ccc(C	0.49	9.309804
CHEMBL133746	C[C@@H](Cc1ccc(F)c(F)c1)C(=O)N[C@H]1N=C(c2	13.2	7.879426
CHEMBL335184	CC(=O)CCN1C(=O)[C@@H](NC(=O)[C@@H](C)C(	3.7	8.431798
CHEMBL341256	CNC(=O)CN1C(=O)[C@@H](NC(=O)[C@@H](C)C	1.8	8.744727
CHEMBL134124	C[C@@H](Cc1ccc(F)c(F)c1)C(=O)N[C@H]1N=C(c2	5.5	8.259637
CHEMBL337670	C[C@@H](Cc1ccc(F)c(F)c1)C(=O)N[C@H]1N=C(c2	100.6	6.997402
CHEMBL132670	C[C@@H](Cc1ccc(F)c(F)c1)C(=O)N[C@H]1N=C(c2	2.2	8.657577
CHEMBL133820	C[C@@H](Cc1ccc(F)c(F)c1)C(=O)N[C@H]1N=C(c2	0.3	9.522879
CHEMBL335994	COC(=O)CN1C(=O)[C@@H](NC(=O)[C@@H](C)C	12.7	7.896196
CHEMBL341257	C[C@@H](Cc1ccc(F)c(F)c1)C(=O)N[C@H]1N=C(c2	4.7	8.327902
CHEMBL18997	CCOC(=O)[C@H](O)[C@H](CC1CCCCC1)NC(=O)[(	5000	5.30103

Table-5: Gamma Secretase (SMILES, pIC50)

ChEMBL_ID	Isomeric_SMILES	IC50	pIC50
CHEMBL174068	CC(=O)O/N=C1/C(c2c(O)[nH]c3cc(Cl)ccc23)=Nc2cc(	17	7.769551
CHEMBL435707	Cn1c(O)c(C2=Nc3ccccc3C2=O)c2ccccc21	100000	4
CHEMBL176904	O=C1C(c2c(O)[nH]c3cc(Br)ccc23)=Nc2ccccc21	45	7.346787
CHEMBL173040	O=Nc1c(-c2c(O)[nH]c3cc(Cl)ccc23)[nH]c2ccccc12	20	7.69897
CHEMBL367345	O=C1C(c2c(O)[nH]c3ccc(F)cc23)=Nc2ccccc21	78	7.107905
CHEMBL176896	CC(=O)O/N=C1/C(c2c(O)[nH]c3ccccc23)=Nc2ccccc2	200	6.69897
CHEMBL173171	Cn1c(O)c(-c2[nH]c3ccccc3c2N=O)c2ccc(Br)cc21	100000	4
CHEMBL354276	CO/N=C1/C(c2c(O)[nH]c3cc(Br)ccc23)=Nc2ccccc21	30	7.522879
CHEMBL368954	C=Cc1ccc2c(C3=Nc4ccccc4/C3=N\OC(C)=O)c(O)[nH	65	7.187087
CHEMBL173445	O=C1C(c2c(O)[nH]c3ccc(I)cc23)=Nc2ccccc21	68	7.167491
CHEMBL172547	CC(=O)O/N=C1/C(c2c(O)[nH]c3cc(Br)c([N+](=O)[O	6	8.221849
CHEMBL353554	CC(=O)O/N=C1/C(c2c(O)[nH]c3cc(I)ccc23)=Nc2cccc	13	7.886057
CHEMBL174292	O=Nc1c(-c2c(O)[nH]c3cc(F)ccc23)[nH]c2ccccc12	130	6.886057
CHEMBL352781	CC(=O)O/N=C1/C(c2c(O)[nH]c3cc(F)ccc23)=Nc2ccc(	90	7.045757
CHEMBL367113	O=C1C(c2c(O)[nH]c3ccc(Cl)cc23)=Nc2ccccc21	50	7.30103
CHEMBL369823	C=Cc1ccc2c(-c3[nH]c4ccccc4c3N=O)c(O)[nH]c2c1	60	7.221849
CHEMBL176306	O=C1C(c2c(O)[nH]c3cccc(Cl)c23)=Nc2ccccc21	100000	4
CHEMBL111433	Cc1ccc2[nH]c(O)c(C3=Nc4ccccc4C3=O)c2c1	62	7.207608
CHEMBL170347	O=C1C(c2c(O)[nH]c3cc(Br)ccc23)=Nc2cc(Br)ccc21	4500	5.346787

Table-6: GSK-3 (SMILES, pIC50)

2.2 Generating Molecular descriptors

2.2.1 Molecular descriptors

Molecular descriptors are crucial in many domains, including toxicology, ecotoxicology, pharmaceutical sciences, chemistry, quality control, and environmental protection regulations. Given that there are already over 5000 descriptors accessible, it is clear that the scientific community is quite interested in molecular descriptors. Specialized software intended for chemical structure analysis may be used to calculate these descriptors generated from various theories and approaches (69).

Molecular descriptors are essential in studying chemical compounds, enabling scientists to gain insights into their properties, predict their behavior, 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.

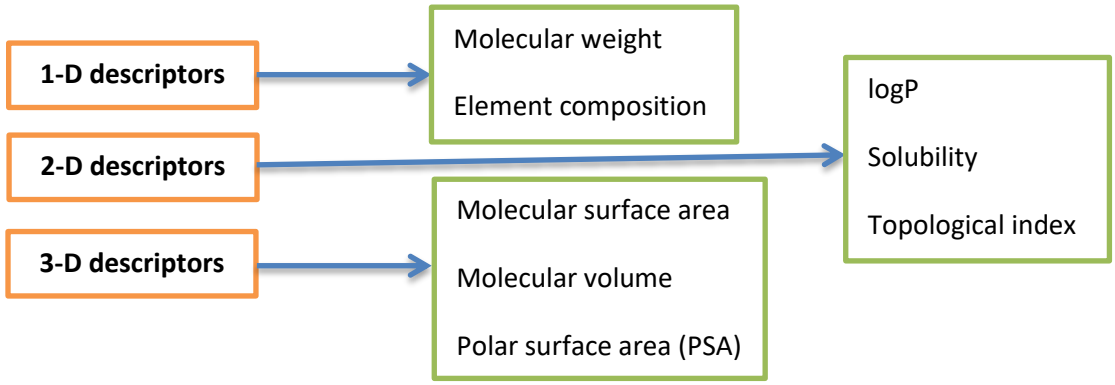


Figure_11: Example of different types of descriptors

2.2.2 RDKit Molecular descriptors

The RDKit is a powerful toolkit for scientists working with molecules. It combines cheminformatics, the computer science of handling chemical information, with machine learning capabilities. Available in both C++ and Python programming languages, RDKit offers researchers a versatile platform to analyze, manipulate, and predict properties of molecules. 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.

RDKit can calculate molecular descriptor and fingerprints and RDKit currently calculate 1826 descriptors.

2.2.3 Steps to generate Molecular descriptors

Descriptors are generated by using Rdkit module.

- Importing the Rdkit libraries
- Importing FeatureGenerator from utility
- Importing other important libraries
- Importing IPythonConsole from rdkit.Chem.Draw
- Saving file to csv after descriptors generation

Rdkit:
https://www.rdkit.org/docs/source/rdkit.Chem.rdMolDescriptors.html#

2.2.4 Tables of molecular descriptors of different targets and drug database

nAcid	nBase	SpAbs_A	SpMax_A	SpDiam_A	SpAD_A	SpMAD_A	LogEE_A	VE1_A
0	0	30.23246	2.405582	4.717447	30.23246	1.314455	4.058594	3.840743
0	0	33.49096	2.451292	4.758891	33.49096	1.339638	4.186356	4.287669
0	0	35.26661	2.452221	4.783405	35.26661	1.259522	4.253862	4.356057
0	0	32.83482	2.449436	4.761983	32.83482	1.262878	4.18349	4.24282
0	0	27.69004	2.434744	4.742345	27.69004	1.203915	4.04731	4.021272
0	0	31.42423	2.445495	4.770803	31.42423	1.309343	4.104776	4.163159
0	0	23.79813	2.430991	4.699082	23.79813	1.252533	3.860423	3.794519
0	0	35.21922	2.445971	4.753489	35.21922	1.304416	4.208342	4.140251
0	0	25.40047	2.432254	4.710031	25.40047	1.270023	3.90751	3.845493
0	0	35.31613	2.448188	4.776706	35.31613	1.26129	4.254008	4.349996
0	0	31.73798	2.445993	4.753653	31.73798	1.322416	4.101255	4.124946
0	0	28.65426	2.451456	4.820839	28.65426	1.302466	4.026621	4.194205
0	0	36.34823	2.502841	5.005681	36.34823	1.253387	4.294303	3.786712
0	0	31.35765	2.438022	4.770841	31.35765	1.306569	4.104776	4.15296
0	0	25.47317	2.397238	4.632981	25.47317	1.273658	3.902803	3.618517
0	0	35.21922	2.445971	4.753489	35.21922	1.304416	4.208342	4.140251
0	0	31.42423	2.445495	4.770803	31.42423	1.309343	4.104776	4.163159
0	0	23.1561	2.462371	4.788237	23.1561	1.218742	3.865205	3.755832
0	0	34.1697	2.448906	4.776176	34.1697	1.314219	4.177435	4.242876

Table-7: Some Rdkit descriptors AChE

nAcid	nBase	SpAbs_A	SpMax_A	SpDiam_A	SpAD_A	SpMAD_A	LogEE_A	VE1_A
1	0	31.9719	2.5584	4.95201	31.9719	1.27888	4.18088	4.40895
0	0	29.111	2.55062	4.89944	29.111	1.32323	4.05914	4.20315
0	0	27.9737	2.49715	4.81302	27.9737	1.33208	4.01476	4.24941
0	0	26.5575	2.31055	4.54899	26.5575	1.26464	3.94733	3.79527
0	0	28.1764	2.54551	4.87798	28.1764	1.34173	4.01515	4.12985
0	1	42.605	2.43288	4.84073	42.605	1.25309	4.45687	5.12213
0	1	46.0895	2.49757	4.9058	46.0895	1.28026	4.52762	4.30964
0	0	32.2903	2.55232	4.90247	32.2903	1.29161	4.17387	4.28794
0	0	28.1421	2.56185	4.88389	28.1421	1.3401	4.01923	4.13119
0	0	27.735	2.54375	4.88071	27.735	1.32072	4.01881	4.14806
0	0	29.0225	2.54998	4.89504	29.0225	1.3192	4.05914	4.20173
0	0	31.4452	2.54802	4.88709	31.4452	1.31022	4.13474	4.21902
0	0	30.4638	2.57206	4.91792	30.4638	1.32451	4.10167	4.25431
0	0	30.2869	2.55144	4.89916	30.2869	1.31682	4.09775	4.23658
0	0	33.8807	2.55342	4.90781	33.8807	1.3031	4.20851	4.32471
0	0	35.508	2.5635	4.95969	35.508	1.26814	4.28017	4.4768
0	0	32.2903	2.55232	4.90247	32.2903	1.29161	4.17387	4.28794
0	0	29.0225	2.54998	4.89504	29.0225	1.3192	4.05914	4.20173
0	0	32.2903	2.55232	4.90247	32.2903	1.29161	4.17387	4.28794

Table-8: Some Rdkit descriptors CDK5

nAcid	nBase	SpAbs_A	SpMax_A	SpDiam_A	SpAD_A	SpMAD_A	LogEE_A	VE1_A
0	0	45.6999	2.47028	4.86578	45.6999	1.30571	4.49767	4.74659
0	0	49.6829	2.49947	4.9085	49.6829	1.30744	4.57358	4.82968
0	0	50.4133	2.50194	4.91332	50.4133	1.29265	4.59888	4.87576
0	0	49.4526	2.49406	4.93326	49.4526	1.30138	4.57491	4.88869
0	0	50.7634	2.49826	4.95101	50.7634	1.30162	4.60038	4.99006
0	0	48.296	2.49873	4.90645	48.296	1.3053	4.55206	4.86609
0	0	49.4526	2.49406	4.93326	49.4526	1.30138	4.57491	4.88869
0	0	48.6521	2.50808	4.91212	48.6521	1.28032	4.58023	4.94629
0	0	46.152	2.48035	4.87732	46.152	1.282	4.50769	4.58876
0	0	47.1405	2.49208	4.89605	47.1405	1.30946	4.52527	4.76585
0	0	51.5279	2.50046	4.9105	51.5279	1.2882	4.62156	4.88188
0	0	52.1013	2.5026	4.91476	52.1013	1.30253	4.62166	4.90385
0	0	52.8681	2.50374	4.91744	52.8681	1.28947	4.64588	4.94974
0	0	49.4174	2.49157	4.92325	49.4174	1.30046	4.5748	4.85648
0	0	49.4526	2.49406	4.93326	49.4526	1.30138	4.57491	4.88869
0	0	49.4526	2.49406	4.93326	49.4526	1.30138	4.57491	4.88869
0	0	52.1013	2.5026	4.91476	52.1013	1.30253	4.62166	4.90385
0	0	48.6436	2.48952	4.91365	48.6436	1.31469	4.54877	4.7785
0	0	41.2741	2.33607	4.67213	41.2741	1.25073	4.37872	4.68189

Table-9: Some Rdkit descriptors GS

nAcid	nBase	SpAbs_A	SpMax_A	SpDiam_A	SpAD_A	SpMAD_A	LogEE_A	VE1_A
0	0	32.2903	2.55232	4.90247	32.2903	1.29161	4.17387	4.28794
0	0	28.1421	2.56185	4.88389	28.1421	1.3401	4.01923	4.13119
0	0	27.735	2.54375	4.88071	27.735	1.32072	4.01881	4.14806
0	0	29.0225	2.54998	4.89504	29.0225	1.3192	4.05914	4.20173
0	0	27.8	2.54442	4.88544	27.8	1.32381	4.01881	4.14928
0	0	31.4452	2.54802	4.88709	31.4452	1.31022	4.13474	4.21902
0	0	30.4638	2.57206	4.91792	30.4638	1.32451	4.10167	4.25431
0	0	30.2869	2.55144	4.89916	30.2869	1.31682	4.09775	4.23658
0	0	33.8807	2.55342	4.90781	33.8807	1.3031	4.20851	4.32471
0	0	27.8	2.54442	4.88544	27.8	1.32381	4.01881	4.14928
0	0	35.508	2.5635	4.95969	35.508	1.26814	4.28017	4.4768
0	0	32.2903	2.55232	4.90247	32.2903	1.29161	4.17387	4.28794
0	0	29.0225	2.54998	4.89504	29.0225	1.3192	4.05914	4.20173
0	0	32.2903	2.55232	4.90247	32.2903	1.29161	4.17387	4.28794
0	0	27.8	2.54442	4.88544	27.8	1.32381	4.01881	4.14928
0	0	30.612	2.55113	4.90096	30.612	1.33096	4.0979	4.23969
0	0	27.7962	2.55299	4.91267	27.7962	1.32363	4.019	4.15371
0	0	27.8	2.54442	4.88544	27.8	1.32381	4.01881	4.14928
0	0	28.4882	2.54777	4.88557	28.4882	1.29492	4.06264	4.22129

Table-10: Some Rdkit descriptors GSK-3

ABC	ABCGG	nAcid	nBase	SpAbs_A	SpMax_A	SpDiam_A	SpAD_A	SpMAD_A
10.4853	10.0164	2	0	17.5059	2.45545	4.65773	17.5059	1.34661
25.8981	20.222	0	0	42.0994	2.45924	4.84859	42.0994	1.27574
10.6066	8.73114	0	0	18.8778	2.21051	4.42102	18.8778	1.34842
6.65169	6.74143	1	0	10.668	2.26717	4.34256	10.668	1.18534
13.0727	11.5415	1	0	20.9029	2.45669	4.91337	20.9029	1.22958
13.504	10.6206	0	2	23.3076	2.32516	4.65031	23.3076	1.29486
3.93265	4.24438	1	0	6	2	4	6	1
10.5619	10.3813	0	0	15.909	2.38552	4.62776	15.909	1.13635
3.93265	4.24438	0	0	6	2	4	6	1
19.4081	15.766	0	0	32.2323	2.38576	4.77152	32.2323	1.2397
15.714	13.9636	0	0	24.507	2.18598	4.37195	24.507	1.11396
21.119	16.6941	2	0	34.2075	2.43628	4.87257	34.2075	1.2217
22.8561	17.9086	0	0	36.9009	2.62671	5.21306	36.9009	1.27244
12.9633	12.2483	0	2	21.175	2.45705	4.66615	21.175	1.24559
18.1903	14.1879	0	0	28.3454	2.44302	4.88604	28.3454	1.23241
12.2277	11.1149	0	0	20.1472	2.40696	4.76699	20.1472	1.2592
17.913	14.6405	0	0	29.3454	2.48445	4.9689	29.3454	1.27589
13.9804	10.3246	0	0	23.5427	2.50191	5.00382	23.5427	1.38486
11.5206	10.5907	1	0	19.0656	2.39027	4.72855	19.0656	1.27104

Table-11: Some Rdkit descriptors of Drug Bank database

Chapter 3

Machine Learning Approach

3.1 Data pre-processing and 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 which has no use, like duplicates, character values, etc.

- NAN value handling

- Using Lipinski's rule
  The dataset is filtered according to the Lipinski rule of 5.

  Lipinski Rule of Five:

  In drug research and development, Lipinski's Rule of Five (RO5) is widely utilized and performs several crucial tasks. First, it is a tool for confirming forecasts and determining whether chemicals need to be registered with regulatory bodies.

  Second, it makes it easier to screen small-molecule medications at first, which helps with the first evaluation of drug candidates who might only partially adhere to the regulation. Drug research and development projects can save money by using this first screening to reduce the number of potential therapeutic options.

  Additionally, by offering a standard for chemical synthesis, the rule is essential in establishing drug databases. Lastly, it provides direction for locating clinical backup medications and improving clinical development pathways' decision-making processes (LP5).

  Rules are as follows:

  1. A maximum of five hydrogen bond donors, restricted to the combination of nitrogen- and oxygen-hydrogen bonds.

2. Ten hydrogen bond acceptors at most, all of which are made up of oxygen or nitrogen atoms.
3. A molecular mass of no more than 500 daltons; and
4. An octanol-water partition coefficient (Clog P) of no more than five.

- Splitting the dataset into train and test data i.e., 80:20

- Standard scaling using standard scaler method

StandardScaler is a sci-kit-learn preprocessing approach that removes the mean and scales to unit variance to standardize features.

- Feature Selection by using SelectKBest

SelectKBest uses a scoring method such as chi-squared, ANOVA F-value, mutual information, or regression coefficients to determine each feature's importance separately.

After feature selection, approx. 500 features are selected for modeling.

	SpAbs_A	SpMax_A	SpDiam_A	SpAD_A	LogEE_A	VE1_A	VE3_A	VR1_A	VR2_A
0	30.23	2.41	4.72	30.23	4.06	3.84	2.18	342.14	14.88
1	33.49	2.45	4.76	33.49	4.19	4.29	2.37	296.57	8.26
2	35.27	2.45	4.78	35.27	4.25	4.36	2.50	257.19	9.19
3	32.83	2.45	4.76	32.83	4.18	4.24	2.40	222.32	8.55
4	27.69	2.43	4.74	27.69	4.05	4.02	2.22	178.66	7.77
...	...	...	...	...	...	...	...	...	...
7278	33.46	2.44	4.85	33.46	4.18	3.98	2.34	498.56	19.18
7279	27.08	2.41	4.82	27.08	3.98	3.75	2.06	265.49	12.64
7280	29.33	2.41	4.82	29.33	4.06	3.83	2.17	330.87	14.39
7281	20.26	2.46	4.92	20.26	3.67	3.61	1.69	74.56	4.97
7282	40.69	2.46	4.92	40.69	4.36	4.91	2.69	289.50	9.65

Figure_12: Some Features after feature selection step

3.2 Regression Models

Regression analysis is categorized as supervised learning, which is a type of analysis in which predictions are made or data is analyzed using preexisting or previous data. Both training and testing datasets are used in supervised learning. A statistical method called regression analysis is used to analyze and forecast data, especially numerical or predictive data (70).

Regression analysis is used in many different areas, including machine learning, artificial intelligence, data science, economics, finance, real estate, healthcare, marketing, business, science, education, psychology, sports analysis, and agriculture. Its primary objective is to elucidate the kind, strength, and interactions among variables so that predictions based on the created model may be made.

By applying different Regression model we get best score in Random Forest Regressor (RF) and Support Vector Regressor (SVR).

3.2.1 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 favored research approach within the realms of data mining and have found extensive application in diverse fields, including biological research (70).

3.2.2 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 (70).

3.2.3 Assessing Performance: Scoring Metrics for Regression Models

Mean Squared Error (MSE): MSE is anticipated and actual values average squared difference is often assessed using regression analysis, and the 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 (70).

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 (70).

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 (70).

Chapter 4

Validation of Model

4.1 Prediction of pIC50 value

The model is saved and then we predicted the pIC50 value of different Drug database (Drug Bank and Zinc Natural product) for validation of the model.

4.1.1 Drug Bank Database

DrugBank Online stands as a treasure trove of information on medications and their targets, freely available to everyone. This bioinformatics and cheminformatics powerhouse brings together detailed data on drugs (chemical, pharmacological, and pharmaceutical aspects) with comprehensive information about their targets (including sequences, structures, and pathways).
No wonder DrugBank Online is a go-to resource for everyone from the pharmaceutical industry and medicinal chemists to pharmacists, doctors, students, and even the general public. With its extensive coverage, meticulous referencing, and in-depth data descriptions, DrugBank Online is a driving force in the data-driven revolution of medicine.

DrugBank till now having **9137** drugs molecules.

4.1.2 Zinc Natural Product (ZincNP) Database

Natural products are a popular starting point for discovering new drugs. These molecules have often evolved to interact with specific proteins in living organisms, making them good candidates for targeting human diseases. While most natural products can be bought, some require collaboration to access. Notably, over a third of all drugs are either natural products themselves or based on their structures. However, natural products can be expensive. Thankfully, online resources like the Special Subsets within DrugBank can help researchers explore the vast potential of natural products and their metabolites in drug discovery.

ZINC provides access to various chemical databases thanks to collaborations with their owners. However, to ensure the data's relevance to drug discovery, ZINC curates these catalogs, meaning some original compounds might be excluded. Despite this, extensive documentation exists for ZINC's subsets. Importantly, anyone can use ZINC for research at no cost, individuals and institutions alike. While sharing search results or screened

molecules from ZINC is allowed, redistributing large portions of the database requires written permission from John Irwin.

4.2 Molecular Docking

4.2.1 Protein Preparation

The Protein Data Bank critical resource for researchers by providing a massive archive of 3D structures for biological molecules like proteins and nucleic acids. These structures are determined using techniques such as X-ray crystallography, NMR spectroscopy, and cryo-electron microscopy. Scientists can access the PDB, like the RCSB PDB you mentioned, to retrieve structures of specific proteins. This allows for a deeper understanding of a protein's function and physical properties, including how it interacts with other molecules (ligands) at the amino acid level. By analyzing these interactions, researchers can gain valuable insights into various biological processes and potentially develop new drugs.

AChE (PDB ID: 4ey7), CDK5 (PDB ID: 1h4l), and GSK-3 (PDB ID: 7sxf) were downloaded from RCSB-PDB in PDB format and for GS we need to model the structure by the use of SWISS Model tool, it is online tool available via the Expasy web server and is a fully automated protein structure homology-modelling service. This site is intended to provide protein modeling to all life science researchers worldwide.

To prepare the proteins, we downloaded AutoDock4 and the MGL tool and prepared it. Multiple steps were involved in the preparation of the protein.

Steps for Preparation of Protein:

1. We loaded protein molecules on AutoDock4, deleted water molecule heteroatoms, because they interfere with the binding of ligands to the protein molecule.
2. Adding the Polar hydrogen helps to find the hydrogen bond's interaction, making it more favorable for the ligand to bind with the protein.
3. Kollman charge (0.0) were also added to the protein.
4. The pdbqt format of the prepared protein was downloaded.

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

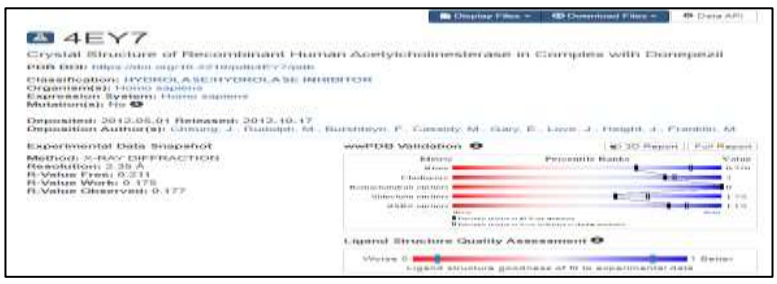


Figure_13: AChE PDB Structure

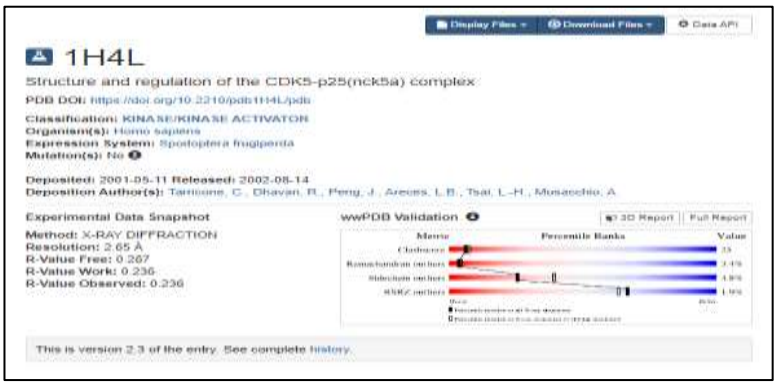


Figure_14: CDK5 PDB Structure

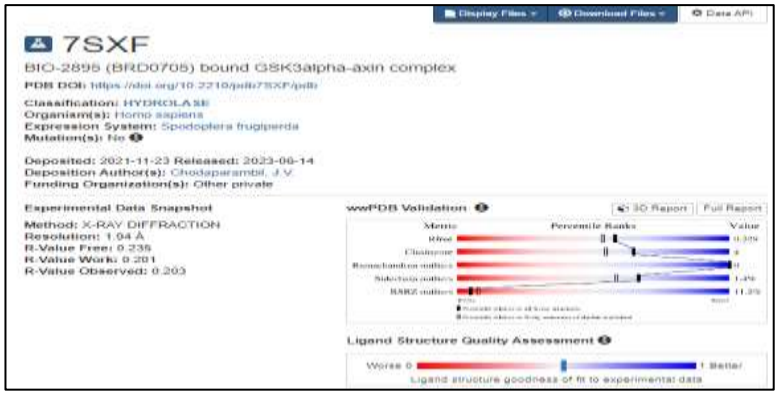


Figure_15: GSK-3 PDB Structure

4.2.2 Ligand Preparation

The ligand, in our case, is the drug that has more than 7 pIC50 values for all the targets and to prepare the ligand we downloaded the PyMol tool for converting .sdf file into .pdb file format.

Steps for Preparation of Ligand:

1. The ligand was downloaded in .sdf format from the PubChem chemical database.
2. The .sdf file is converted into .pdb format using the PyMol tool.

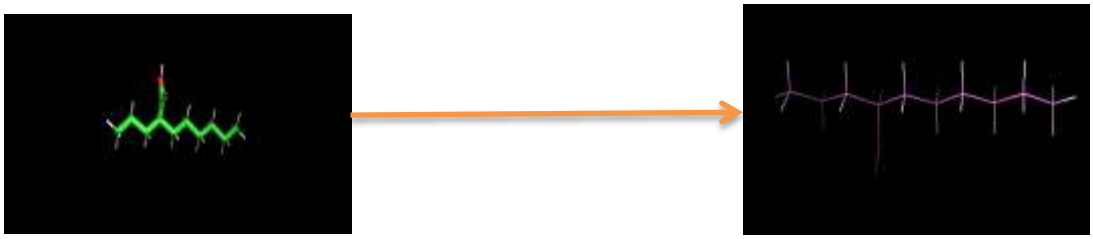


Figure_16: Ligand in .sdf file format

Figure_17: Ligand in .pdb file format

3. The .pdb file is then converted into .pdbqt format using the AutoDock4 tool and saved.

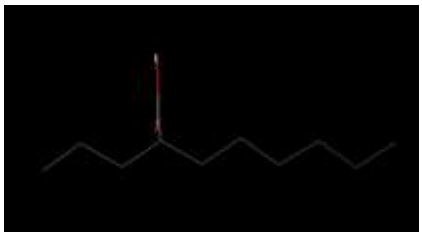


Figure_18: Ligand in .pdbqt file format

4.2.3 AutoDock Vina Docking

Molecular docking is a vital technique in computer-assisted drug design and structural molecular biology. The primary goal of ligand-protein docking is to forecast the ligand's dominant binding interaction with the selected target protein within the known three-dimensional structure. We utilized Vina (v.1.2.0), an open-source molecular docking program. In computational drug discovery, virtual screening software is used to examine the interactions between compounds and possible therapeutic targets. In order to provide information regarding the binding site's active center position, we carried out site-specific docking. It includes coverage for both the ligand and the site.

Targets	Centre			Size (Dimension)		
	X	Y	Z	X	Y	Z
AChE	-4.988	52.822	13.113	126	126	126
CDK5	37.705	-19.547	24.315	74	84	90
GS	5.519	0.249	-22.842	126	126	126
GSK-3	-18.057	-19.421	21.470	80	96	102

Table-12: Description of Config file.

The energy range is 4 and exhaustiveness is 8.

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

4.2.4 Molecular Visualization of Interacting Molecules

The docked structure was shown using PyMOL and LigPlot Plus, which is 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/

Chapter 5

Results and Discussion

5.1 Inhibitors SMILES and Descriptors for Targets

The Inhibitors SMILES is generated by using chembl web resources and already mention in the above chapter about total number of inhibitors SMILES is generated for different targets and same as for descriptors we use Rdkit descriptors for feature generation. The detail is already mentioned in above chapters.

5.2 Machine Learning Models

The separate models are prepared for the targets on two datasets i.e. 2D descriptors and all descriptors (1D, 2D and 3D)

The features are the descriptors (X_train and X_test) of the targets and the pIC50 value (y_train and y_test) will be the prediction data.

For 2D descriptors, the Support Vector Regressor (SVR) gives the best score, i.e. R^2 value for the targets AChE and CDK5; the accuracy score are 0.65 and 0.72 respectively. Similarly, the Random Forest Regressor gives the best score for the targets GS and GSK-3; the accuracy score are 0.61 for both respectively.

For all descriptors (1D, 2D and 3D), only the Support Vector Regressor (SVR) gives the best prediction score for all targets, i.e., AChE, CDK5, GS, and GSK-3 and the accuracy are 0.68, 0.69, 0.62 and 0.67 respectively.

Target / Model (R^2 Score)	Linear Regression	Polynomial Regression	Support Vector Machine	Decision Tree	Random Forest Regressor
AChE	19.7	20.2	**65**	7.3	7.2
CDK5	8.2	2.4	**72**	7.3	65
GS	6.1	26.8	57.1	5.1	**61**
GSK3	7.4	16.8	57.5	3.6	**61**

Table-13: Different Regression Model for 2D descriptors.

Target ↓ / Model → (R^2 Score)	Linear Regression	Polynomial Regression	Support Vector Machine	Decision Tree	Random Forest Regressor
AChE	1.7	23.4	**68**	9.2	7.2
CDK5	10	2.4	**69**	10.3	6.5
GS	7.6	30	**62**	34	5.9
GSK3	11.3	10.5	**67**	1.9	3.5

Table-14: Different Regression Model for all descriptors (1D, 2D and 3D).

5.2.1 Scatter Plots

1. Scatter Plot for AChE: Predicted Values vs True Values

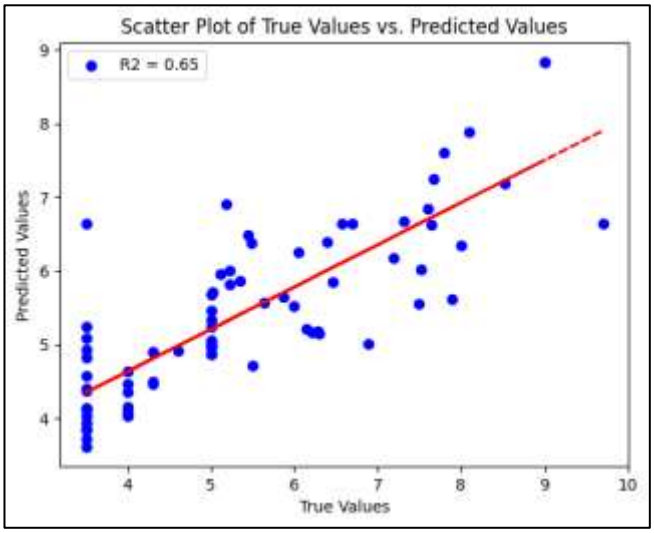


Figure_19: 2D descriptors

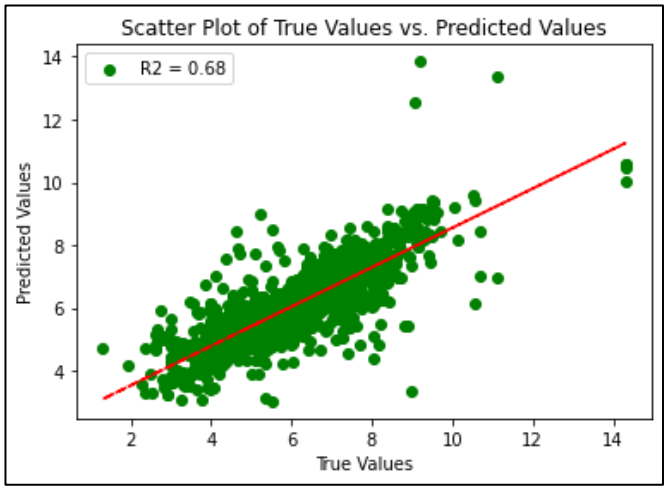


Figure_20: 3D descriptors

2. Scatter Plot for CDK5: Predicted Values vs True Values

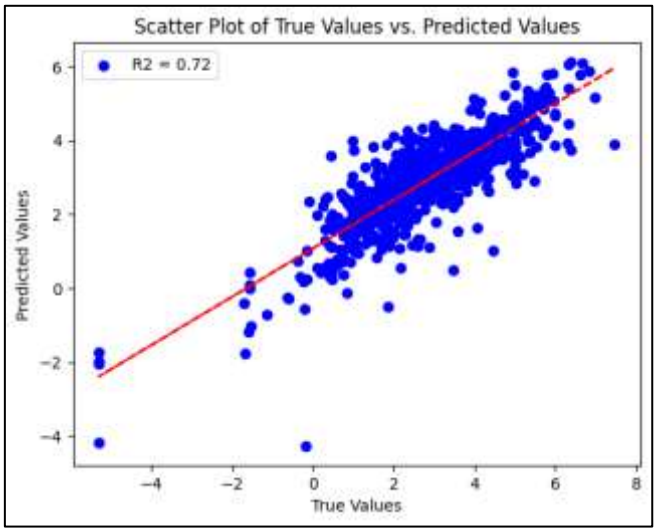


Figure_21: 2D descriptors

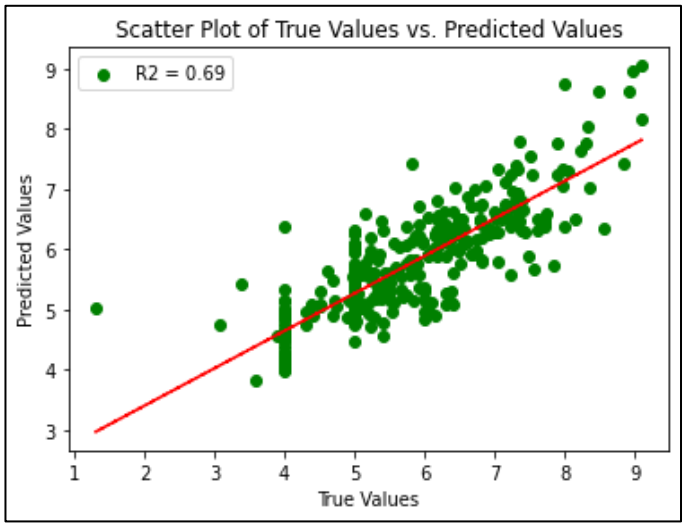


Figure_22: 3D descriptors

3. Scatter Plot for GS: Predicted Values vs True Values

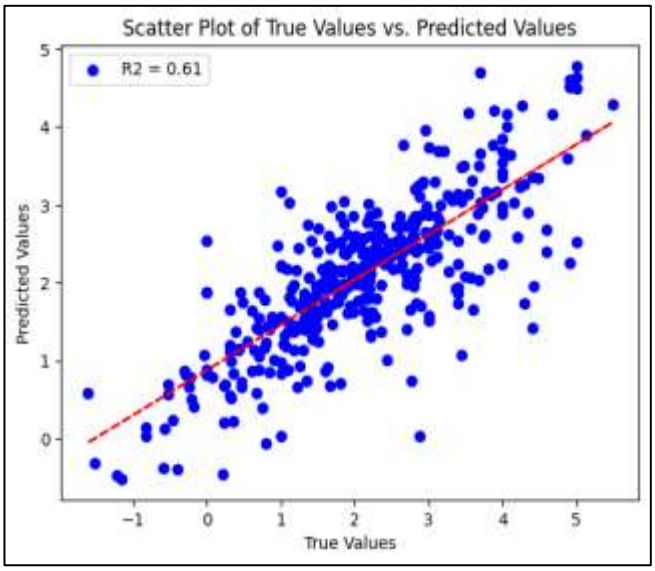


Figure_23: 2D descriptors

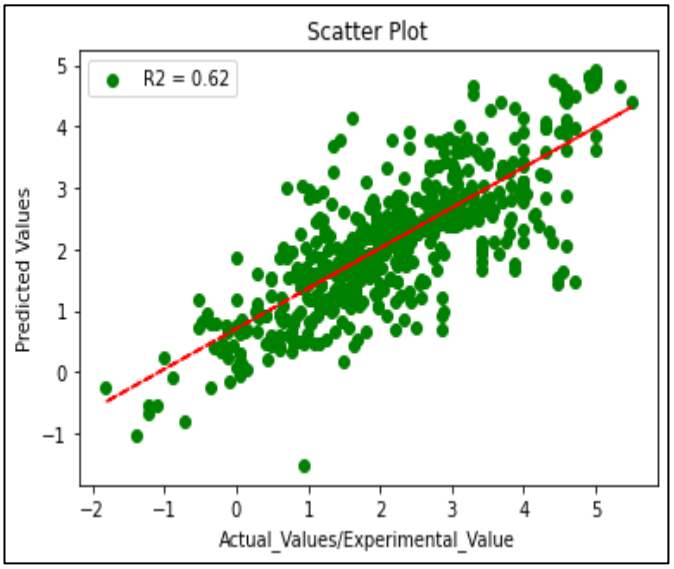


Figure_24: 3D descriptors

4. Scatter Plot for GSK-3: Predicted Values vs True Values

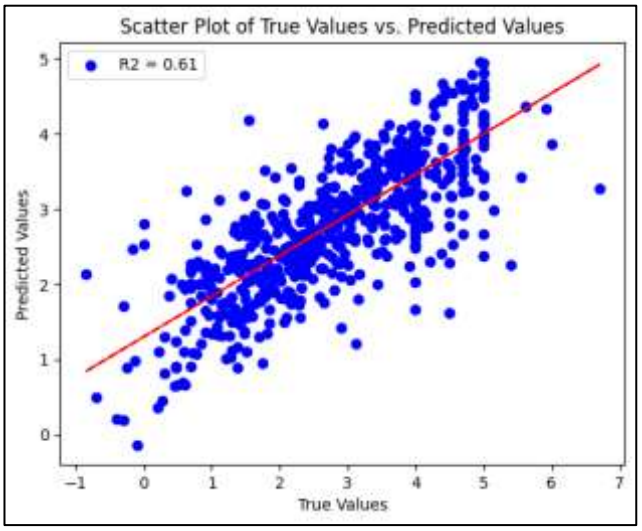


Figure_25: 2D descriptors

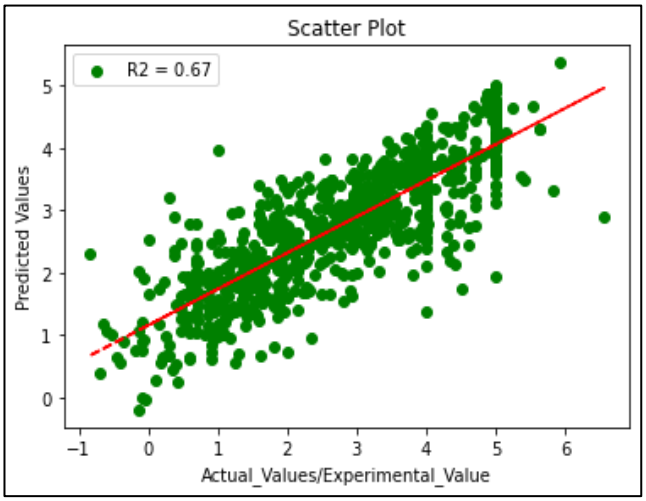


Figure_26: 3D descriptors

5.3 Prediction of pIC50

5.3.1 Drug Bank Database Prediction

By taking **11715 drugs** molecule in .mol2 file format as our Prediction dataset or Test dataset for validation of our model.

1. The first step is generating the descriptors for all molecules, so by using RDkit descriptors we generated all descriptors.

2. Second step is to save our model in .pkl format for validation.

The prediction table is as follows for Drug Bank:

AChE	CDK5	GS	GSK3
6.33	7.26	7.18	7.3
5.92	5.76	6.17	6.14
5.65	5.74	7.36	5.85
5.96	5.42	7.34	6.12
6.23	5.37	7.03	6.15
5.97	7.1	7.02	8.11
5.52	5.41	6.14	5.77
5.95	7.32	7.38	7.18
5.95	7.1	7.01	8.11
6.46	7.16	7.09	7.9
5.8	5.44	5.7	5.96
5.86	6.1	7.35	6.48
6.05	5.6	7.09	6.22
6.29	7.18	7.18	7.54
5.94	7.31	7.37	7.22
5.97	7.32	7.31	7.3
5.6	5.93	7.07	5.76
5.77	7.31	7.24	7.21
5.77	5.73	7.65	5.58

Table-15: Predicted pIC50 value for Drug Bank using different models.

- **After the prediction, we selected those drugs that gave a value of pIC50 above 7 for all targets, and we got a list of five medicines that could be the best candidates for multi-targeting.**

5.3.2 Zinc Natural Product Prediction

By taking **224205 natural products** molecule in .mol2 file format as our Prediction dataset or Test dataset for validation of our model.

3. The first step is generating the descriptors for all molecules, so by using RDkit descriptors we generated all descriptors.

4. Second step is to save our model in .pkl format for validation.

The prediction table is as follows for Zinc NP:

AChE	CDK5	GS	GSK3
6.18	5.62	7.14	6.34
5.76	5.59	6.71	5.52
5.88	5.37	5.28	6.03
6.2	5.23	8.1	6.5
5.65	5.15	5.34	4.82
6.31	5.68	5.92	5.71
5.53	5.51	6.55	5.37
6.21	5.5	5.75	5.31
4.81	5.28	5.73	5.42
6.49	5.14	5.28	6.04
5.49	5.48	5.43	5.16
5.69	5.52	5.69	5.53
5.17	5.44	5.65	5.56
6.89	5.52	5.71	5.46
5.37	5.54	5.74	5.96
5.79	5.48	6.83	5.74
6.17	5.85	6.97	6.07
5.6	5.54	5.27	4.95
6.31	5.71	7.48	6.12

Table-16: Predicted pIC50 value for Zinc NP using different models.

- **After predicting the pI50 value for Zinc Natural Products, we don't get the drugs that give the value above 7 for all targets.**
- **So, further processing will be done using Drug Bank prediction values.**

5.3.3 List of Drugs after prediction

After the prediction steps, we selected the drugs having a pIC50 value above 7 for all the targets, as mentioned earlier.

Name	AChE	CDK5	GS	GSK3
Clomocycline	7.28	7.32	7.28	7.18
Raloxifene	7.06	7.22	7.09	7.69
Celecoxib	7.12	7.29	7.27	7.2
Arundic acid	7.18	7.18	7.11	7.8
Chlorphenoxamine	7.21	7.31	7.31	7.19

Table-17: List of drugs from Drug Bank having pIC50 value above 7 for all targets.

The list of drugs is mentioned below in the table:

S.No	Drugs_SMILES	Drugs_Name
1.	CCCCCC[C@@H](CCC)C(=O)O	Arundic acid
2.	CC1=CC=C(C=C1)C2=CC(=NN2C3=CC=C(C=C3)S(=O)(=O)N)C(F)(F)F	Celecoxib
3.	CC(C1=CC=CC=C1)(C2=CC=C(C=C2)Cl)OCCN(C)C	Chlorphenoxamine
4.	C[C@@]1([C@H]2C[C@H]3[C@@H](C(=O)C(=C([C@]3(C(=O)C2=C(C4=C(C=CC(=C41)Cl)O)O)O)O)C(=O)NCO)N(C)C)O	Clomocycline
5.	C1CCN(CC1)CCOC2=CC=C(C=C2)C(=O)C3=C(SC4=C3C=CC(=C4)O)C5=CC=C(C=C5)O	Raloxifene

Table-18: List of drugs from Drug Bank after prediction.

- The next step is to validate predicted results by performing Molecular Docking of Targets as Proteins and the drugs mentioned above as Ligands.

5.4 Molecular Docking

Molecular docking 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 target scan easily bind with the specific ligand and attain stable conformation with significantly less energy expenditure.

The molecular docking steps are mentioning earlier i.e. Protein and Ligand preparation.

- **Proteins Structure:**

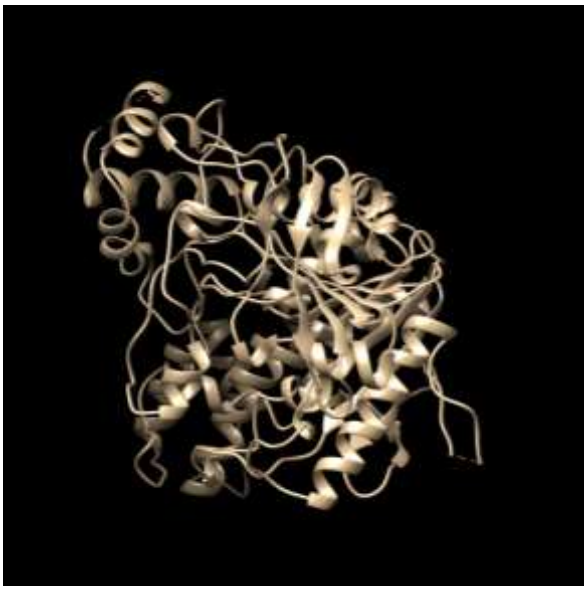


Figure_27: AChE structure

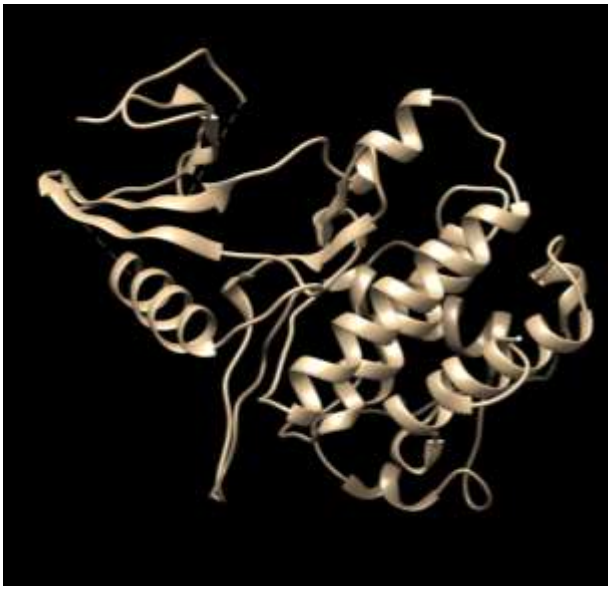


Figure_28: CDK5 structure

Figure_29: GS structure

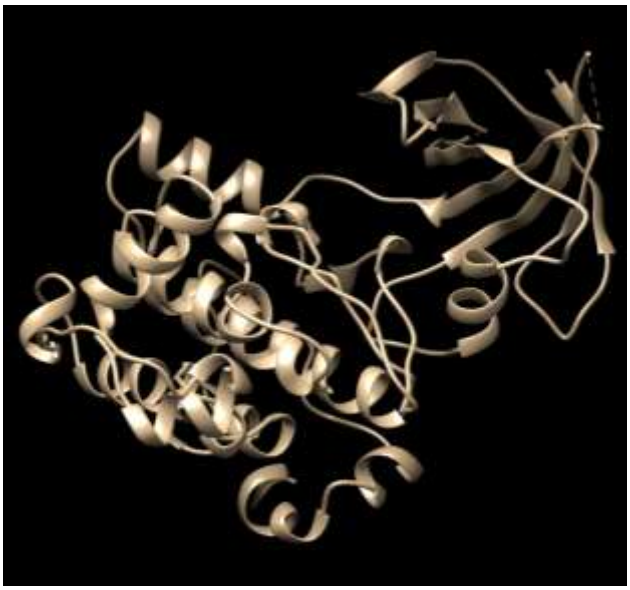


Figure_30: GSK-3 structure

- **Ligand Structure:**

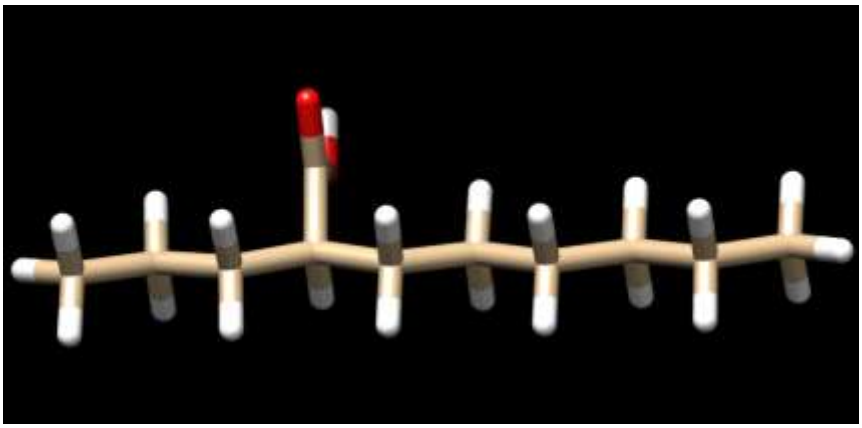


Figure_31: Arundic Acid structure

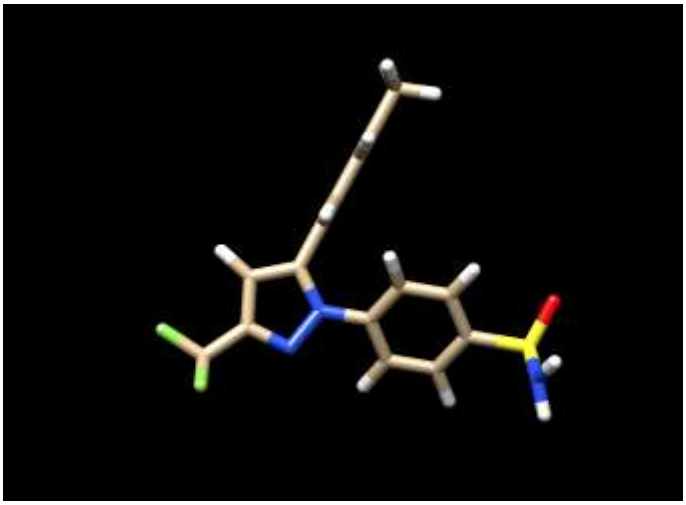


Figure_32: Celecoxib structure

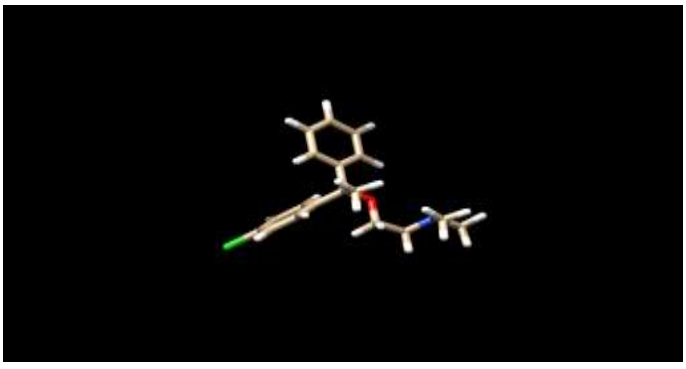


Figure_33: Chlorophenoxamine structure

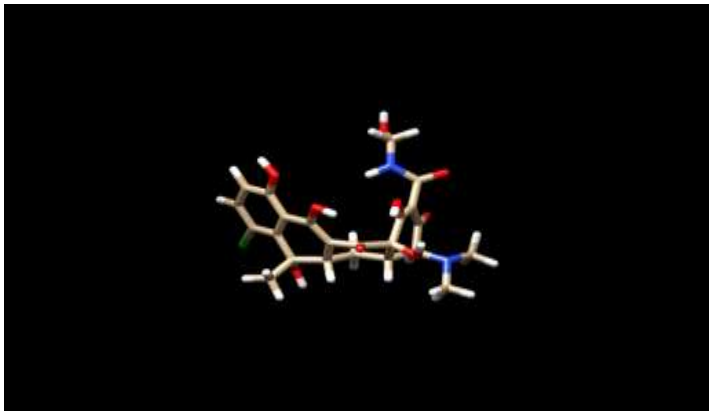


Figure_34: Clomocycline structure

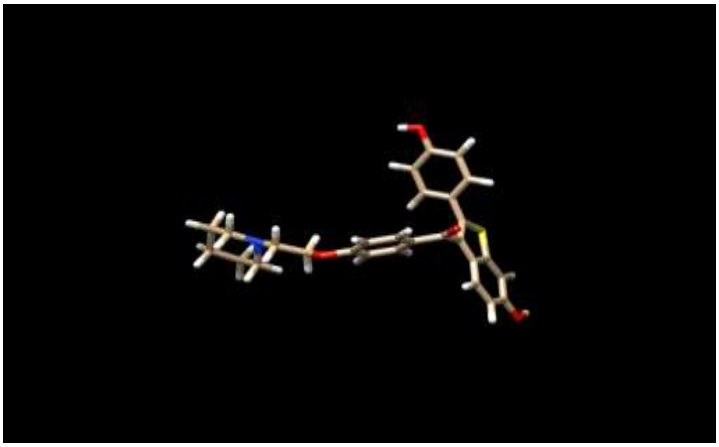


Figure_35: Raloxifene structure

The next step is the preparation of a grid box. We first define the protein's active site using PyMol or CastP tools.

The active sites for targets are:

1. **AChE**: ASP-390, GLU-389, ARG-393, HIS-387, ARG-356, LEU-394, TRP-385, ASP-384, ALA-397, TYR-382, ASN-350, GLU-351, LEU-353, SER-352, ASP-349, PHE-346, PRO-344, SER-347, ALA-526, CYS-529, ALA-530, GLN-508, ASN-533, ARG-534, ALA-505, ALA-343, TYR-337, TYR-341, PHE-338, SER-293, HIS-447, ASP-74, TRP-86, TRP-286, TYR-72, GLY-448, PHE-295, SER-208, GLY-120, GLU-202, TYR-133, TYR-124, GLY-121, LEU-289, ARG-296.

2. **CDK5**: GLU-8, ILE-10, VAL-18, ALA-31, LEU-32, LYS-33, GLU-51, VAL-64, PHE-80, GLU-81, PHE-82, CYS-83, ASP-84, GLN-85, ASP-86, LYS-89, GLN-130, LEU-133, ALA-143, ASP-144.

3. **GS**: We don't have any active site for the GS and did the blind docking as the GS structure is not available, so we modeled it using the SWISS model tool.

4. **GSK**-3: ILE-125, VAL-133, TYR-134, GLN-135, ALA-146, VAL-173, GLU196, TYR-197, VAL-198, PRO-199, GLU-200, THR-201, ARG-204, GLN-248, ASN-249, LEU-251, CYS-262, ASP-263, VAL-277, SER-278, ILE-280, CYS-281, ARG-283, ARG-286, SER-324, GLY-325

After grid box preparation, we save the file as a config.txt file for performing the docking.

The docking results are validated by seeing the binding affinity of protein-ligand interaction, and the more negative the value, the better the binding affinity will be present.

Target Protein / Ligands	Arundic Acid BA (kcal/mol)	Celecoxib BA (kcal/mol)	Chlorophenoxamine BA (kcal/mol)	Clomocycline BA (kcal/mol)	Raloxifene BA (kcal/mol)
AChE	-6.4	-8.5	-5.6	-8.0	-8.9
CDK5	-5.0	-6.8	-5.1	-8.1	-8.2
GS	-4.0	-7.4	-5.1	-7.6	-7.2
GSK-3	-4.4	-8.2	-5.1	-8.1	-8.9

Table-19: The table for the best binding affinity.

5.5 Molecular Visualization of Docked Molecule

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 and LigPlot Plus tools.

5.5.1 Using PyMol

PyMol visualization tool for visualize the molecule for docking.

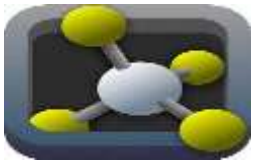


Figure_36: PyMol

- Molecular Visualization of the Docked Molecules for Arundic Acid with different Target Proteins.

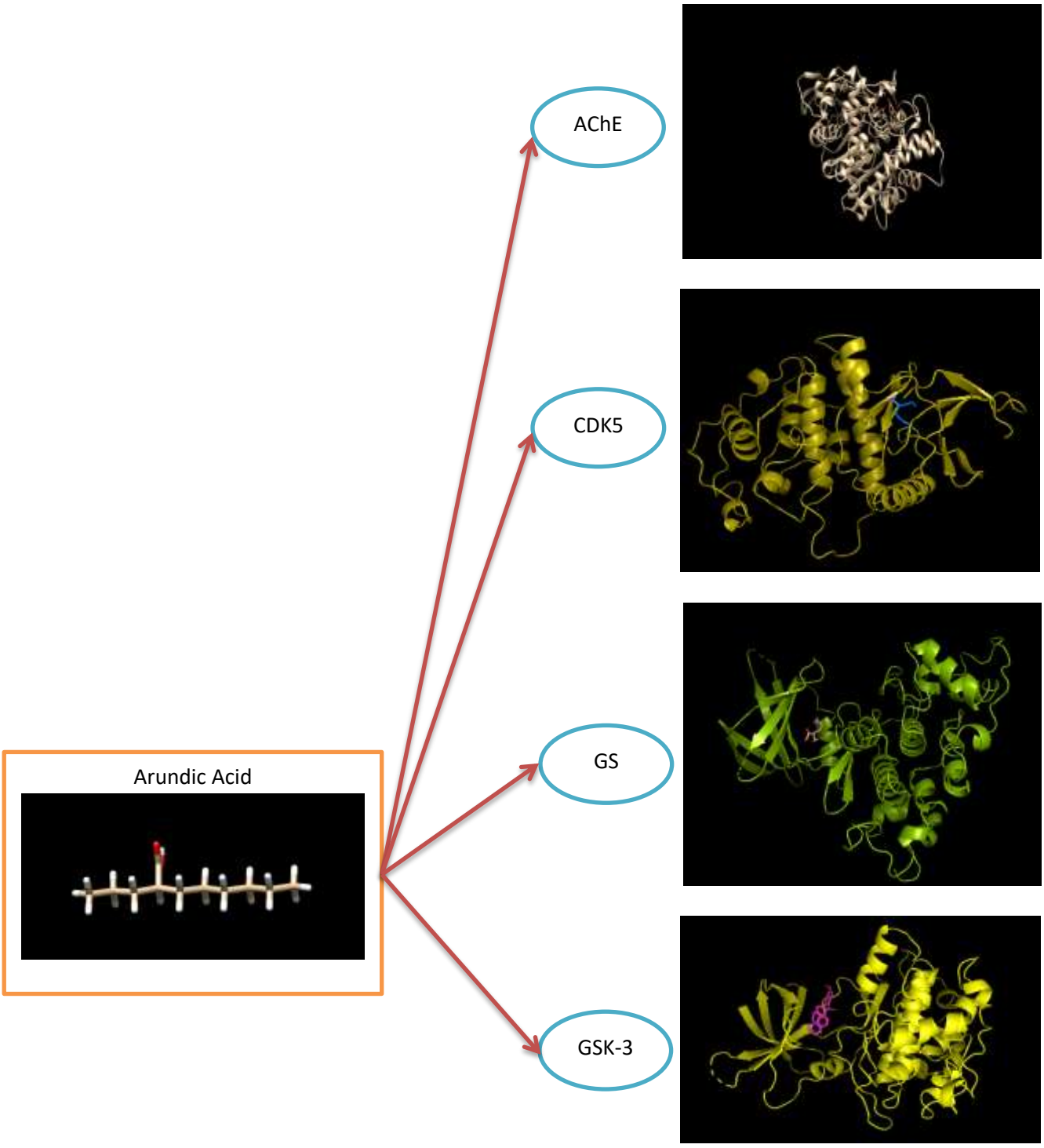


Figure_37: Graphical visualization of Arundic Acid with different target proteins.

- Molecular Visualization of the Docked Molecules for Celecoxib with different Target Proteins.

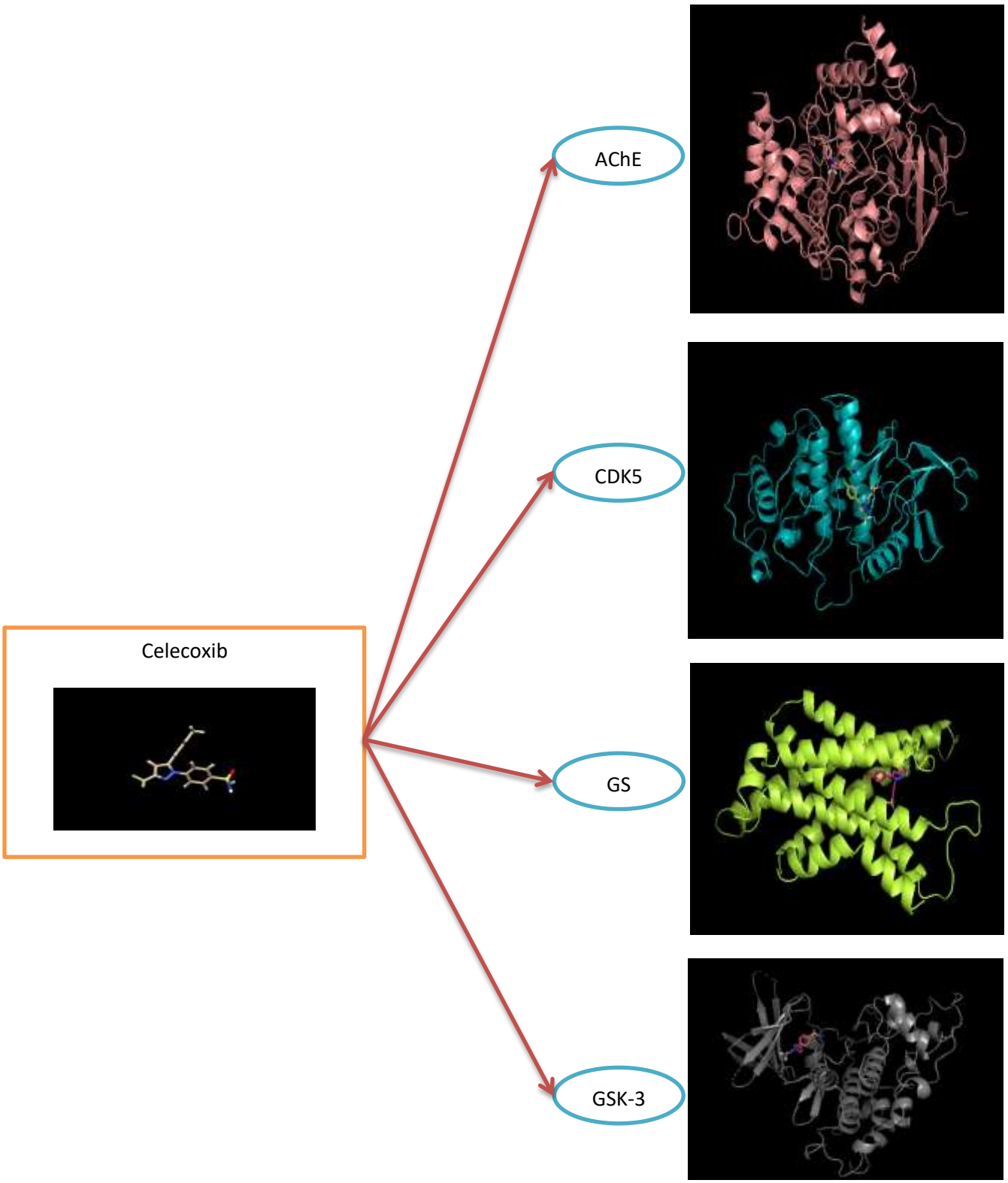


Figure_38: Graphical visualization of Celecoxib with different target proteins.

- Molecular Visualization of the Docked Molecules for Chlorphenoxamine with different Target Proteins.

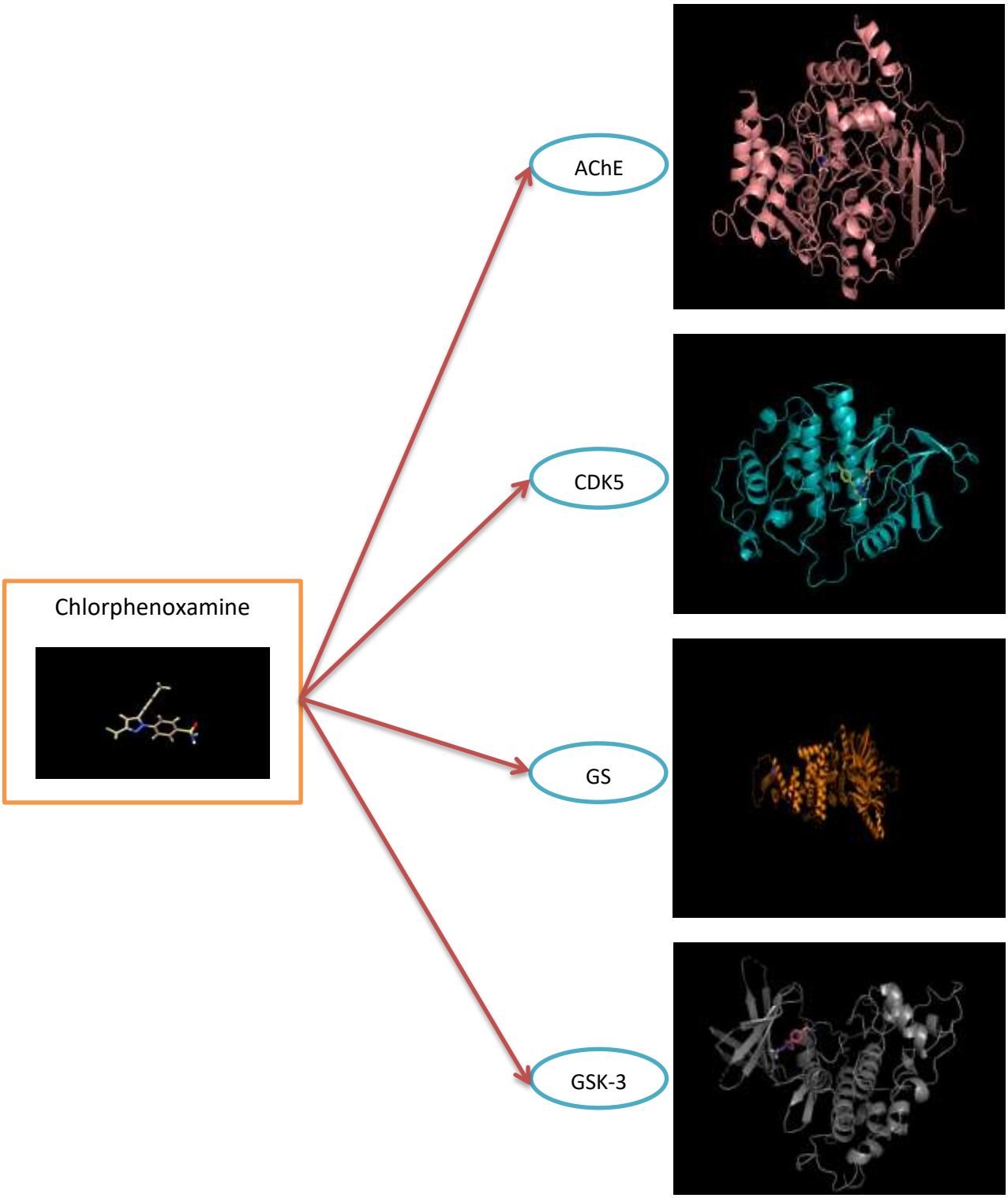


Figure_39: Graphical visualization of Chlorphenoxiamine with different target proteins.

- Molecular Visualization of the Docked Molecules for Clomocycline with different Target Proteins.

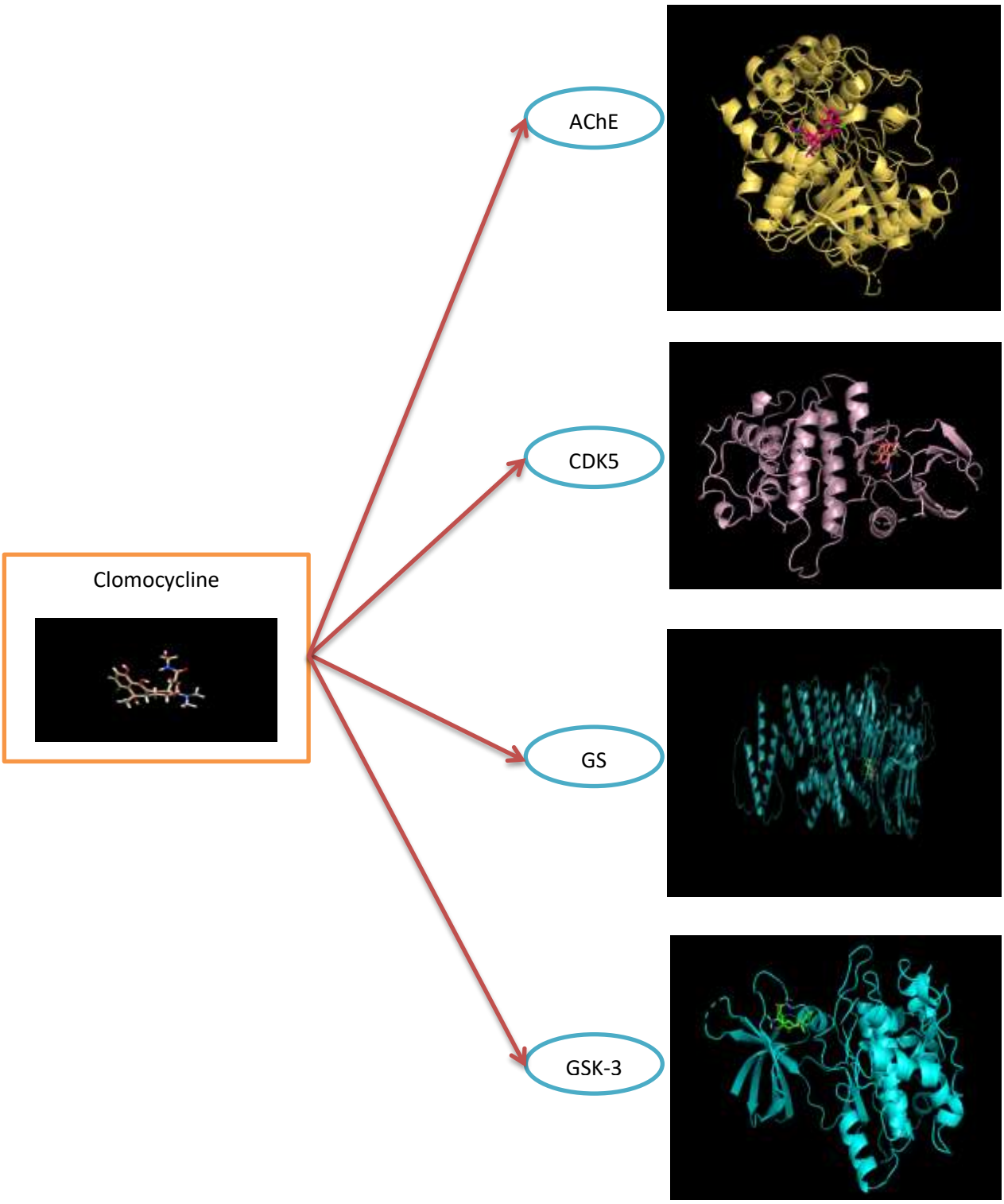


Figure_40: Graphical visualization of Clomocycline with different target proteins.

- Molecular Visualization of the Docked Molecules for Raloxifene with different Target Proteins.

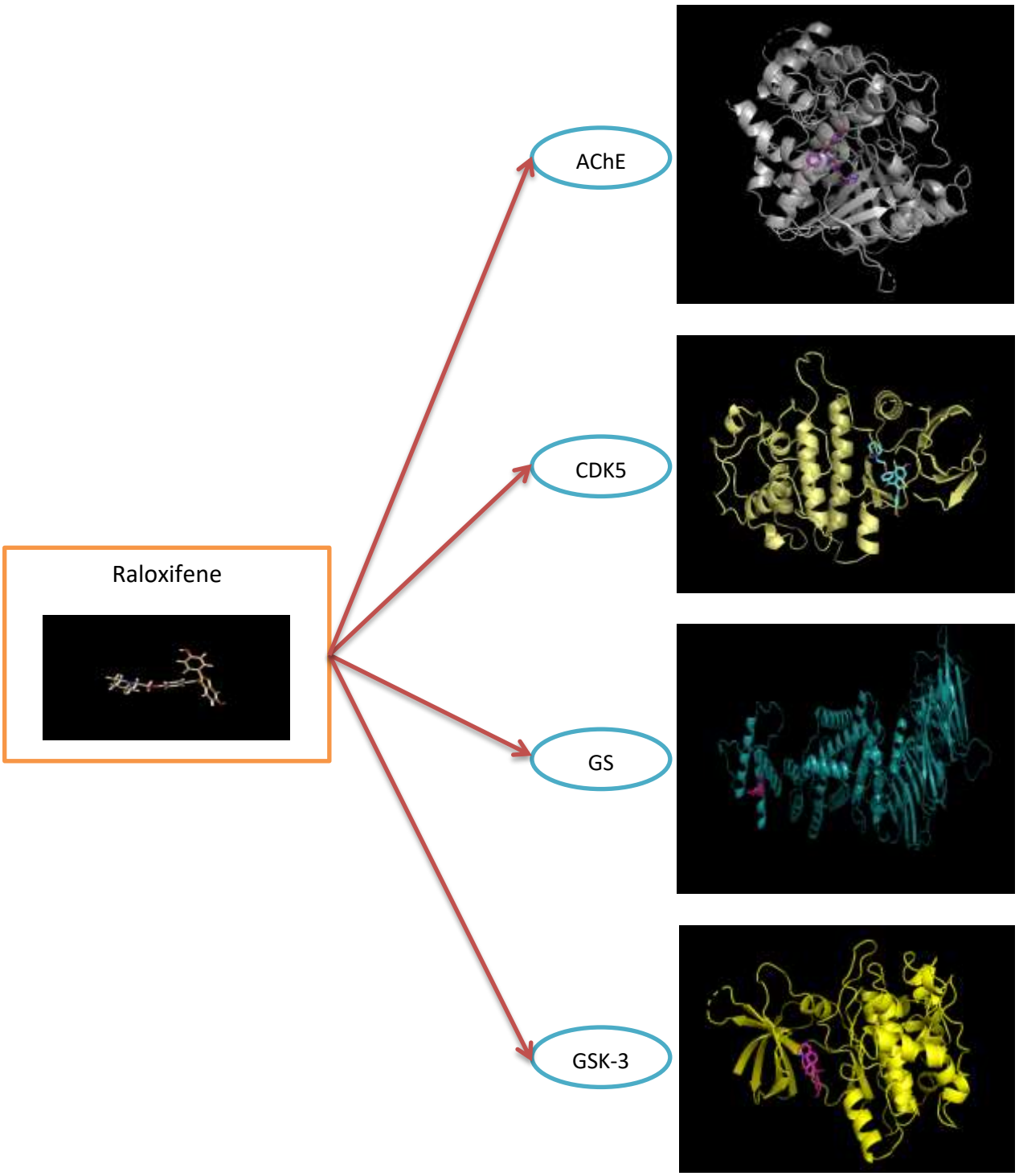


Figure_41: Graphical visualization of Raloxifene with different target proteins.

5.5.2 Using LigPlot Plus

LigPlot Plus an another visualization tool for visualize the molecule for docking.

Figure_42: LigPlot Plus

- Molecular Visualization of the Docked Molecules for Arundic Acid with different Target Proteins.

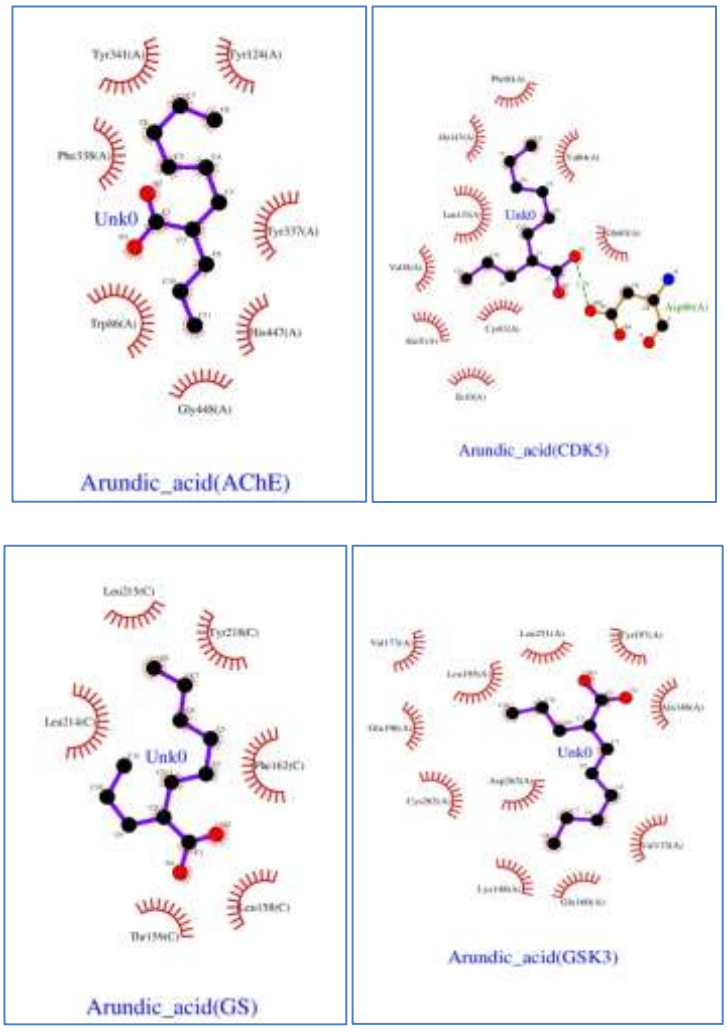


Figure_43: LigPlot Plus visualization of Arundic Acid with different target proteins.

- Molecular Visualization of the Docked Molecules for Celecoxib with different Target Proteins.

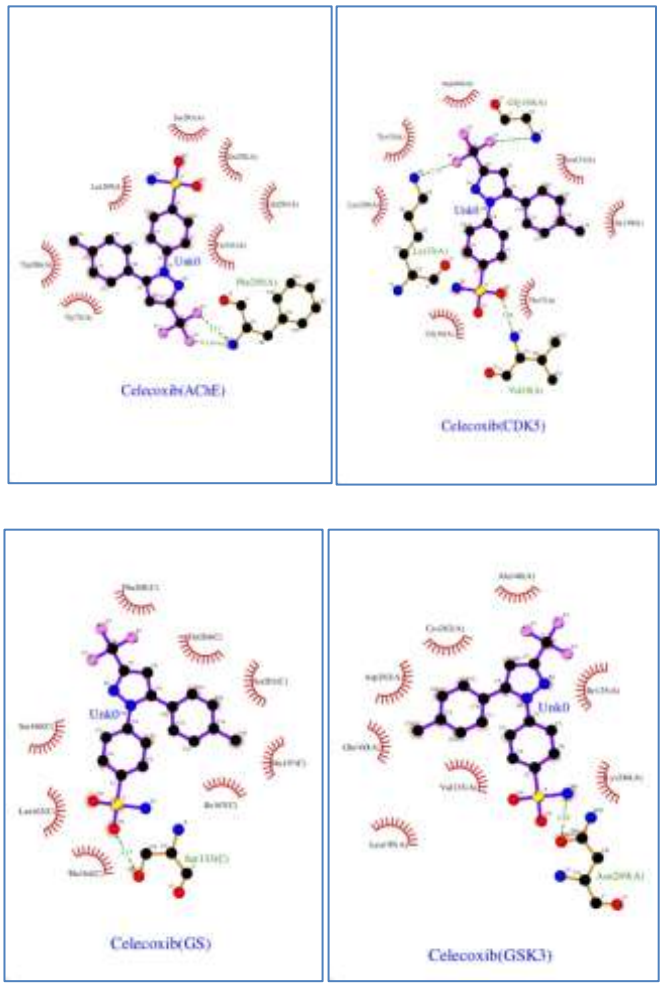
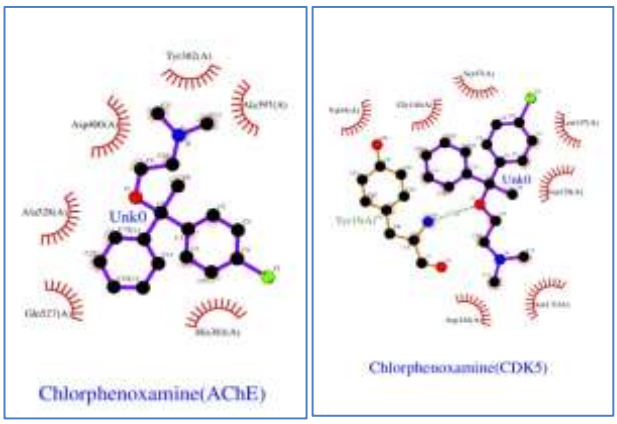


Figure_44: LigPlot Plus visualization of Celecoxib with different target proteins.

- Molecular Visualization of the Docked Molecules for Chlorpheoxiamine with different Target Proteins.



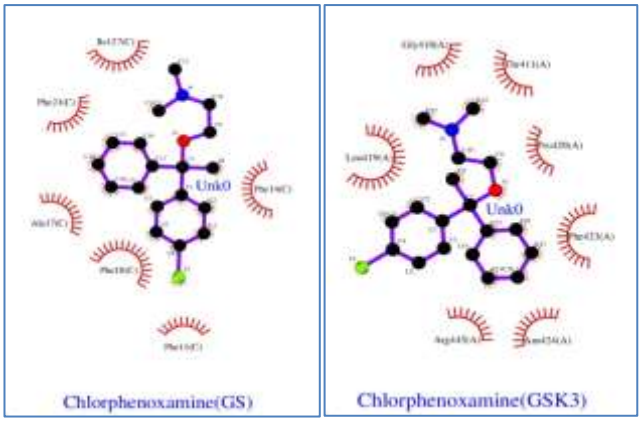


Figure_45: LigPlot Plus visualization of Chlorpheoxiamine with different target proteins.

- Molecular Visualization of the Docked Molecules for Clomocycline with different Target Proteins.

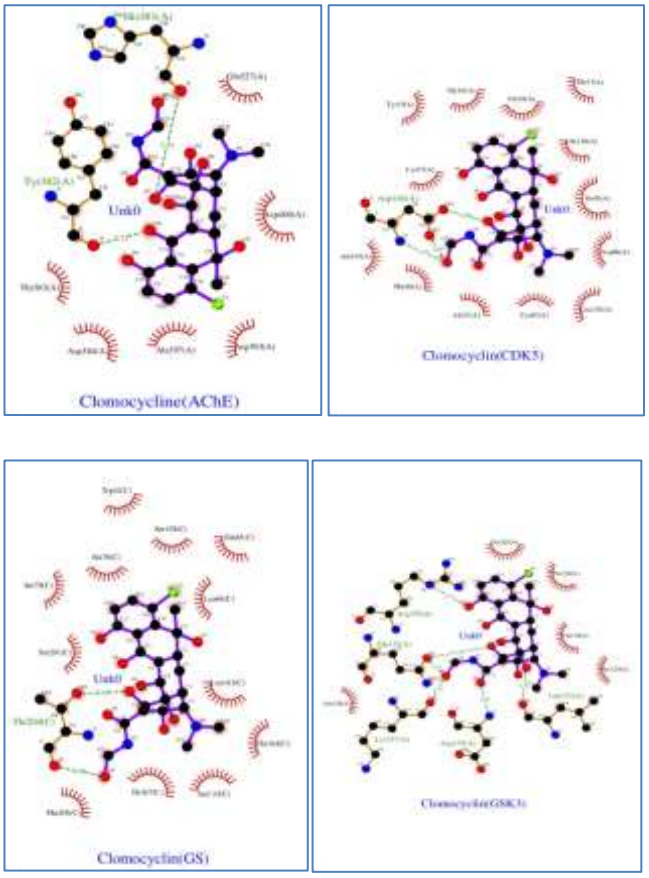


Figure_46: LigPlot Plus visualization of Clomocycline with different target proteins.

- Molecular Visualization of the Docked Molecules for Ralxofiene with different Target Proteins.

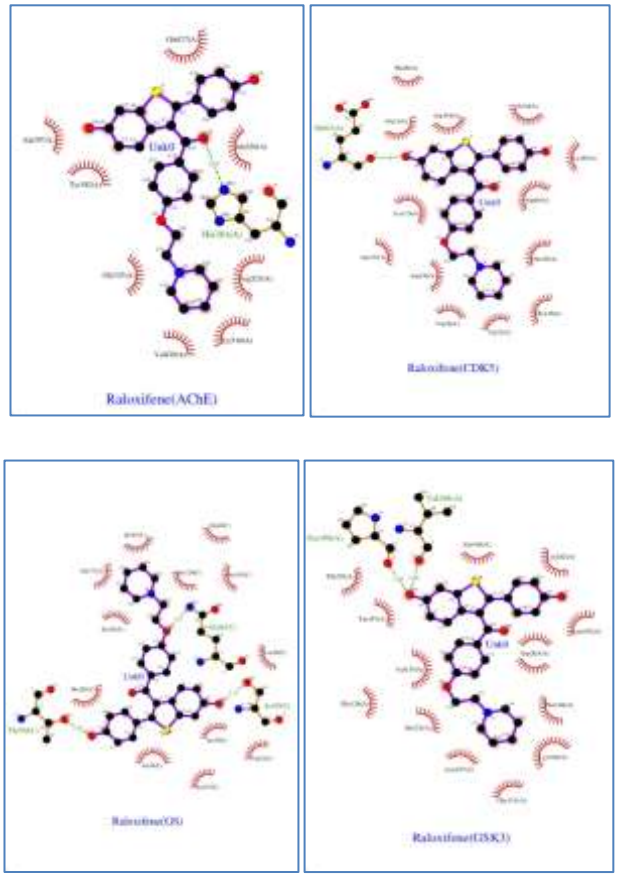


Figure_47: LigPlot Plus visualization of Raloxifene with different target proteins.

Chapter 6

Conclusion and Future aspects

6.1 Conclusion

In this thesis, titled "Machine Learning Based Design Of Multi-targeting Inhibitors for Alzheimer's Disease and Structure-Based Validation," we investigated multi-targeting drugs using regressor prediction models. For the model, the features are extracted by the ChEMBL database in SMILES form and generated descriptors using the RDkit tool, which gives about 1825 types of descriptors or features, including all kinds, i.e., 1D, 2D, and 3D. We use the regression model as our prediction value is continuous, i.e., pIC50, and using different regression models shows that Support Vector and Random Forest are performing well for 2D and 3D descriptors features, as mentioned above in a table. The prediction is done on the DrugBank and Zinc Natural Product, and we get the five drugs from DrugBank, which have the pIC50 value above 7 and show good binding affinity against all our targets after performing the molecular docking.

6.2 Future Scope

- The drugs we get can be the best candidates for the targets (AChE, CDK5, GS, and GSK-3) of Alzheimer's Disease.
- The wet lab validation is needed.
- The model can be used to predict the pIC50 value.

References

1. V. Andreoli, F. Trecroci, A. La Russa, R. Cittadella, M. Liguori, P. Spadafora, M. Caracciolo, G. Di Palma, C. Colica, A. Gambardella and A. Quattrone, *Alzheimer's & Dementia*, 2011, **7**, 574–578.
2. B. Arosio, L. Mastronardi, C. Vergani and G. Annoni, *International Journal of Alzheimer's Disease*, 2010, **2010**, 1–5.
3. J. Attems, M. Preusser, M. Grosinger-Quass, L. Wagner, F. Lintner and K. Jellinger, *Neuropathology Appl Neurobio*, 2008, **34**, 23–32.
4. J. Avila, F. Wandosell and F. Hernández, *Expert Review of Neurotherapeutics*, 2010, **10**, 703–710.
5. J. Avila, F. Wandosell and F. Hernández, *Expert Review of Neurotherapeutics*, 2010, **10**, 703–710.
6. M. Bajo, B. C. Yoo, N. Cairns, M. Gratzer and G. Lubec, *Amino Acids*, 2001, **21**, 293–301.
7. S. Baressi Šegota, I. Lorencin, Z. Kovač and Z. Car, *Biomedicines*, 2023, **11**, 284.
8. L. Bossaerts, R. Cacace and C. Van Broeckhoven, *Mol Neurodegeneration*, 2022, **17**, 31.
9. D. M. Bowen, S. J. Allen, J. S. Benton, M. J. Goodhardt, E. A. Haan, A. M. Palmer, N. R. Sims, C. C. T. Smith, J. A. Spillane, M. M. Esiri, D. Neary, J. S. Snowdon, G. K. Wilcock and A. N. Davison, *Journal of Neurochemistry*, 1983, **41**, 266–272.
10. A. Brinkmalm, G. Brinkmalm, W. G. Honer, L. Frölich, L. Hausner, L. Minthon, O. Hansson, A. Wallin, H. Zetterberg, K. Blennow and A. Öhrfelt, *Mol Neurodegeneration*, 2014, **9**, 53.
11. F. Checler, C. Alves Da Costa, K. Ancolio, N. Chevallier, E. Lopez-Perez and P. Marambaud, *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease*, 2000, **1502**, 133–138.
12. X. Chen, H. Li, L. Tian, Q. Li, J. Luo and Y. Zhang, *Journal of Computational Biology*, 2020, **27**, 1397–1406.
13. L. N. Clark, L. A. Kartsaklis, R. Wolf Gilbert, B. Dorado, B. M. Ross, S. Kisselev, M. Verbitsky, H. Mejia-Santana, L. J. Cote, H. Andrews, J.-P. Vonsattel, S. Fahn, R. Mayeux, L. S. Honig and K. Marder, *Arch Neurol*, , DOI:10.1001/archneurol.2009.54.
14. V. Consonni and R. Todeschini, in *Recent Advances in QSAR Studies*, eds. T. Puzyn, J. Leszczynski and M. T. Cronin, Springer Netherlands, Dordrecht, 2010, vol. 8, pp. 29–102.
15. D. Culpan, S. H. MacGowan, J. M. Ford, J. A. R. Nicoll, W. S. Griffin, D. Dewar, N. J. Cairns, A. Hughes, P. G. Kehoe and G. K. Wilcock, *Neuroscience Letters*, 2003, **350**, 61–65.
16. B. DaRocha-Souto, M. Coma, B. G. Pérez-Nievas, T. C. Scotton, M. Siao, P. Sánchez-Ferrer, T. Hashimoto, Z. Fan, E. Hudry, I. Barroeta, L. Serenó, M. Rodríguez, M. B.

Sánchez, B. T. Hyman and T. Gómez-Isla, *Neurobiology of Disease*, 2012, **45**, 425–437.
17. N. Embi, D. B. Rylatt and P. Cohen, *European Journal of Biochemistry*, 1980, **107**, 519–527.
18. W. J. Geldenhuys, *CNS Drugs*.
19. J. Gobom, A. Brinkmalm, G. Brinkmalm, K. Blennow and H. Zetterberg, *Molecular & Cellular Proteomics*, 2024, 100721.
20. R. Holma, P. Salmenperä, I. Virtanen, H. Vapaatalo and R. Korpela, .
21. J.-Y. Hur, *Exp Mol Med*, 2022, **54**, 433–446.
22. S. Jayadev, J. B. Leverenz, E. Steinbart, J. Stahl, W. Klunk, C.-E. Yu and T. D. Bird, *Brain*, 2010, **133**, 1143–1154.
23. S. Jiang, Y. Li, C. Zhang, Y. Zhao, G. Bu, H. Xu and Y.-W. Zhang, *Neurosci. Bull.*, 2014, **30**, 295–307.
24. T. Kanekiyo and G. Bu, *Front. Aging Neurosci.*, , DOI:10.3389/fnagi.2014.00093.
25. G. Kaur and E. Levy, *Front. Mol. Neurosci.*, , DOI:10.3389/fnmol.2012.00079.
26. B. Kim, J. Park, K.-T. Chang and D.-S. Lee, *Free Radical Biology and Medicine*, 2016, **90**, 184–194.
27. J. Kim, J. M. Basak and D. M. Holtzman, *Neuron*, 2009, **63**, 287–303.
28. D. S. Knopman, H. Amieva, R. C. Petersen, G. Chételat, D. M. Holtzman, B. T. Hyman, R. A. Nixon and D. T. Jones, *Nat Rev Dis Primers*, 2021, **7**, 33.
29. I. V. Kurochkin, E. Guarnera and I. N. Berezovsky, *Trends in Pharmacological Sciences*, 2018, **39**, 49–58.
30. R. M. Lane, S. G. Potkin and A. Enz, *Int. J. Neuropsychopharm.*, 2005, **9**, 101.
31. S.-L. Liu, C. Wang, T. Jiang, L. Tan, A. Xing and J.-T. Yu, *Mol Neurobiol*, 2016, **53**, 4328–4342.
32. Y. Liu, Y. Wang and J. Zhang, in *Information Computing and Applications*, eds. B. Liu, M. Ma and J. Chang, Springer Berlin Heidelberg, Berlin, Heidelberg, 2012, vol. 7473, pp. 246–252.
33. K.-G. Ma and Y.-H. Qian, *Neuropeptides*, 2019, **73**, 96–106.
34. R. MacLeod, E.-K. Hillert, R. T. Cameron and G. S. Baillie, *Future Science OA*, 2015, **1**, fso.15.9.
35. R. Malinow, *Current Opinion in Neurobiology*, 2012, **22**, 559–563.
36. E. Masliah, L. Hansen, M. Alford, R. Deteresa and M. Mallory, *Annals of Neurology*, 1996, **40**, 759–766.
37. M. C. McCulley, I. N. M. Day and C. Holmes, *American J of Med Genetics Pt B*, 2004, **124B**, 50–53.
38. E. J. Mufson, S. E. Counts and S. D. Ginsberg, .
39. J. L. Muir, .
40. T. Nagai, P. L. McGEER, J. H. Peng, E. G. McGEERI and C. E. Dolman, .
41. R. J. O'Brien and P. C. Wong, *Annu. Rev. Neurosci.*, 2011, **34**, 185–204.

42. S. E. O'Bryant, S. C. Waring, V. Hobson, J. R. Hall, C. B. Moore, T. Bottiglieri, P. Massman and R. Diaz-Arrastia, *J Geriatr Psychiatry Neurol*, 2010, **23**, 49–53.
43. S. Oh, M. S. Kim, S. Park, H. Son, S. Kim, M. Kim, D. Jo, E. Tak and J. Lee, *Brain Pathology*, 2019, **29**, 217–231.
44. J.-J. Pei, I. Grundke-Iqbal, K. Iqbal, N. Bogdanovic, B. Winblad and R. F. Cowburn, .
45. G. Perry, R. Friedman, G. SHAWt and V. Chauo, .
46. V. Plevris, G. Solorzano, N. Bakas and M. Ben Seghier, in *8th European Congress on Computational Methods in Applied Sciences and Engineering*, CIMNE, 2022.
47. E. Rogaeva, Y. Meng, J. H. Lee, Y. Gu, T. Kawarai, F. Zou, T. Katayama, C. T. Baldwin, R. Cheng, H. Hasegawa, F. Chen, N. Shibata, K. L. Lunetta, R. Pardossi-Piquard, C. Bohm, Y. Wakutani, L. A. Cupples, K. T. Cuenco, R. C. Green, L. Pinessi, I. Rainero, S. Sorbi, A. Bruni, R. Duara, R. P. Friedland, R. Inzelberg, W. Hampe, H. Bujo, Y.-Q. Song, O. M. Andersen, T. E. Willnow, N. Graff-Radford, R. C. Petersen, D. Dickson, S. D. Der, P. E. Fraser, G. Schmitt-Ulms, S. Younkin, R. Mayeux, L. A. Farrer and P. St George-Hyslop, *Nat Genet*, 2007, **39**, 168–177.
48. N. Sasaki, S. Toki, H. Chowei, T. Saito, N. Nakano, Y. Hayashi, M. Takeuchi and Z. Makita, *Brain Research*, 2001, **888**, 256–262.
49. A. Savioz, G. Leuba and P. G. Vallet, *Ageing Research Reviews*, 2014, **18**, 86–94.
50. L. S. Schneider, J. A. Severson, H. C. Chui, V. E. Pollock, R. B. Sloane and E. R. Fredrickson, .
51. R. C. Sequeira and A. Godad, *Mol Neurobiol*, , DOI:10.1007/s12035-023-03839-1.
52. S. R. Shajahan, S. Kumar and M. D. C. Ramli, *Front. Aging Neurosci.*, 2024, **15**, 1274452.
53. M. Singh, M. Kaur, H. Kukreja, R. Chugh, O. Silakari and D. Singh, *European Journal of Medicinal Chemistry*, 2013, **70**, 165–188.
54. A. K. Somavarapu and K. P. Kepp, *Journal of Neurochemistry*, 2016, **137**, 101–111.
55. S. H. Son, N.-R. Lee, M. S. Gee, C. W. Song, S. J. Lee, S.-K. Lee, Y. Lee, H. J. Kim, J. K. Lee, K.-S. Inn and N.-J. Kim, *ACS Cent. Sci.*, 2023, **9**, 417–426.
56. K. M. Stouffer, X. Grande, E. Duezel, M. Johansson, B. Creese, M. P. Witter, M. I. Miller, L. E. M. Wisse and D. Berron, *Brain*, 2023, awad411.
57. D. C. Swinney, in *Annual Reports in Medicinal Chemistry*, Elsevier, 2011, vol. 46, pp. 301–317.
58. K. Takahashi, Q. Kong, Y. Lin, N. Stouffer, D. A. Schulte, L. Lai, Q. Liu, L.-C. Chang, S. Dominguez, X. Xing, G. D. Cuny, K. J. Hodgetts, M. A. Glicksman and C.-L. G. Lin, *Journal of Experimental Medicine*, 2015, **212**, 319–332.
59. Z. Tian, B. Feng, X.-Q. Wang and J. Tian, *Front. Mol. Neurosci.*, 2022, **15**, 1030639.
60. L.-H. Tsai, M.-S. Lee and J. Cruz, *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics*, 2004, **1697**, 137–142.
61. D. Twohig and H. M. Nielsen, *Mol Neurodegeneration*, 2019, **14**, 23.
62. G. K. Uyanık and N. Güler, *Procedia - Social and Behavioral Sciences*, 2013, **106**, 234–240.

63. H. Wang, M.-S. Tan, R.-C. Lu, J.-T. Yu and L. Tan, *BioMed Research International*, 2014, **2014**, 1–9.
64. M. Yaar, S. Zhai, P. F. Pilch, S. M. Doyle, P. B. Eisenhauer, R. E. Fine and B. A. Gilchrest, .
65. N. Zaręba and M. Kepinska, *IJMS*, 2020, **21**, 9003.
66. B. Zdrazil, E. Felix, F. Hunter, E. J. Manners, J. Blackshaw, S. Corbett, M. de Veij, H. Ioannidis, D. M. Lopez, J. F. Mosquera, M. P. Magarinos, N. Bosc, R. Arcila, T. Kizilören, A. Gaulton, A. P. Bento, M. F. Adasme, P. Monecke, G. A. Landrum and A. R. Leach, *Nucleic Acids Research*, 2024, **52**, D1180–D1192.
67. F. P. Zemlan, O. J. Thienhaus and H. B. Bosmann, *Brain Research*, 1989, **476**, 160–162.
68. M. Zetterberg, A. Sjölander, M. Von Otter, M. Palmér, S. Landgren, L. Minthon, A. Wallin, N. Andreasen, K. Blennow and H. Zetterberg, *Mol Neurodegeneration*, 2010, **5**, 11.
69. X. Zhan, G. C. Jickling, B. P. Ander, B. Stamova, D. Liu, P. F. Kao, M. A. Zelin, L.-W. Jin, C. DeCarli and F. R. Sharp, *JAD*, 2015, **44**, 1213–1229.
70. L. Zhong, Q. Dong-hai, L. Hong-ying and L. Qing-feng, *Eur J of Neuroscience*, 2009, **30**, 1831–1836.