AgAnt:A computational tool to assess the mode of interaction between a protein-ligand complex

BY
Bhumika Choudhary

MTech in Computational Biology

Supervisor:
Dr. Arjun Ray
Assistant Professor
IIIT-Delhi

A thesis submitted in partial fulfillment of the requirements for the degree of MTech

Department of Computational Biology
Indraprastha Institute of Information TechnologyDelhi, India

August, 2024

Certificate

This is to certify that the thesis titled "AgAnt:A computational tool to assess the mode of interaction between a protein-ligand complex" being submitted by Bhumika Choudhary 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.

August, 2024

Dr.Arjun Ray
Department of Computational Biology
Indraprastha Institute of Information Technology Delhi
New Delhi 110020

Acknowledgemnt

I extend my heartfelt gratitude to my supervisor Dr. Arjun Ray for his invaluable guidance and support throughout this project. I am deeply thankful to my colleague, Riddhi Sharma, for her assistance and collaboration, which significantly contributed to the project's success. I also wish to express my appreciation to all my lab mates for their support and cooperation, which played a vital role in completing this project.

Abstract

Protein-ligand interactions are fundamental in regulating protein activity within cellular environments, playing a pivotal role in numerous biochemical processes. These interaction studies are important for understanding the mechanisms of biological regulation, and they provide a theoretical basis for the design and discovery of new drug targets. Traditionally, till date, these interactions are uncovered based on simple binding energy calculations. Yet, the lacunae of no functional information arising from these interaction studies remains the biggest challenge. The complexity of these interactions increases when ligands can function as either agonists or antagonists, contingent on specific conditions. Understanding these functional dynamics is crucial for advancing drug discovery and development. This project aims to harness machine learning (ML) techniques to classify agonist and antagonist molecules that relate the biological response of the complexes, based on their mode of action, to their structural features, including hydrophobic, electronic, steric, and topological parameters. The main goal of this study is to make use of limited information to create more general models that cover a wider variety of protein-ligand interactions. This is different from traditional methods that usually focus on specific targets or ligands. In this study, we developed a robust computational tool capable of classifying any protein-ligand complex responses as either agonist or antagonist with a test accuracy of 0.87 and precision of 0.91. We also conducted extensive validation of the model on diverse datasets to ensure its accuracy and reliability in predicting protein-ligand interactions. This novel tool demonstrates the ability to interpret and generalize effectively, ensuring precise classification of the mode of action of any specific ligand-protein pair.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

Emerging multidisciplinary discipline spanning the junction of biology, chemistry, and informatics is chemogenomics. Small compounds interacting with proteins define most of the medications used today. Therefore, drug discovery and design depend critically on an awareness of protein-ligand interaction. [27]. Protein-ligand interactions are crucial for mediating enzymatic catalysis, signal transduction, other protein activities, and hence their disturbance is usually linked to diseases [2].Proteins-ligand-interaction models are induced in the discipline of chemogenomics that includes proteo-chemometrics from data matrices comprising both protein and ligand information together with some empirically determined variable. Unlike conventional methods that concentrate mostly on specific targets or ligands, this quantitative multi-structure property-relationship modelling (QMSPR) approach aims generally in two directions: to take advantage of sparse/incomplete sources of information and to obtain generalised models including larger parts of the protein-ligand space. Usually derived from many sparse/incomplete sources, the data matrices include a series of ligands and proteins coupled with quantified information about how they interact. Ideally, a good model should be simple to understand and able to generalise to novel undiscovered protein-ligand combinations. [27]. In many different processes carried out by living entities, protein-ligand interactions are of basic relevance and initiate a variety of signal transduction mechanisms. [30] Therefore, understanding the biological regulatory processes and for offering a theoretical structure for the identification and development of new therapeutic targets depend on research of protein-ligand interactions. [8]. Through interactions with other molecules including the substrate, a ligand DNA and other domain of proteins, protein fulfils its purpose. The three-dimensional arrangement of protein offers the required shape and physical texture to support these interactions. Structural information about protein surface areas helps to enable thorough research of the link between protein structure and function. [6].

CEEDO, Developed and selected evaluations of the main characteristics of a new and extensive protein–ligand database encompassing inter-atomic connections shown as structural interaction fingerprints. Implementing several interaction kinds in the manner of a bit-vector format known as structural interacting fingerprint (SIFt), the database comprises information about interactions at the inter-atomic levels and represents all potential forms of contact type [2]. The interaction of the protein functioning as a receptor with its corresponding ligand determines its functional effect. An agonist effects within the cell by binding to the receptor. An antagonist blocks the same receptors to a natural agonist, yet binding to it does not cause a response. Unlike a full agonist, a partial agonist can cause an effect throughout a cell that is not optimum then block the receptor [4]. Agonists have efficiency and affinity. Simply said, the capacity of an agonist—or drug—to bind to the receptor is known as affinities and is expressed as the proportion of the rate at which it binds (k+1) and the at which it dissociates (k-1), that is, affinities (kA)= k+1/k-1, and is commonly determined through experimental radioactive binding techniques. On the other hand, effectiveness explains how well the drug once binds to activate the receptor. Whereas partial agonists are not able to provide a full reaction even in their presence of a large quantity of agonists, full agonists are able to produce a maximum response or maximal efficacy [34].

$$A + R \underset{k-1}{\overset{k+1}{\rightleftharpoons}} AR \rightleftharpoons AR^* \underset{\alpha}{\overset{\beta}{\rightarrow}} Effect$$

AR is the bound receptor; AR* is the activated receptor; b and an are the corresponding receptors activation and deactivation constants; A is the agonist, or drug; R is the receptor. Categorised either as reversible or irreversible, competitive receptor antagonists They are incapable to generate a response and hence lack efficacy; they interact to the target and so have affinity. [34]. Moreover, they stop the protagonist from binding so they hinder its capacity to stimulate the receptor:

$$B + R \underset{K-1}{\overset{k+1}{\rightleftharpoons}} BR \rightleftharpoons AR \rightarrow Effect$$

B is the antagonist; R is the receptor; BR is the bound receptor. [34].Opposing effects of agonists are produced by inverse agonists. It is a recently discovered phenomenon only achievable in the lack of an agonist by activated receptors [34]. Drugable partial agonists are those that bind to a receptor to produce a response with a maximum effect, submaximal response attained relative to complete agonist (e.g., buprenorphine). Drugs with agonistic–antagonist action can have agonistic properties at a particular receptor and antagonistic effects at another [28].

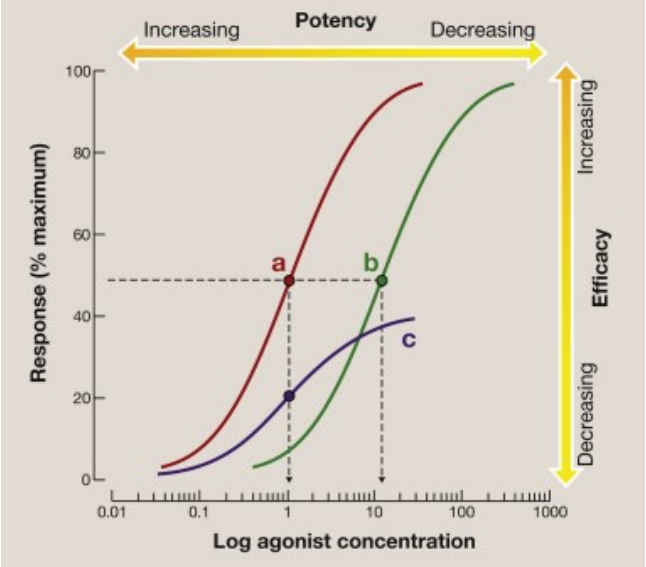

Figure 1.1: Agonists: c has a lesser efficacy (partial ag) than either an or b; an is more potent than b but has equal efficacy; c is equipotent; a and b are full agonists [34]

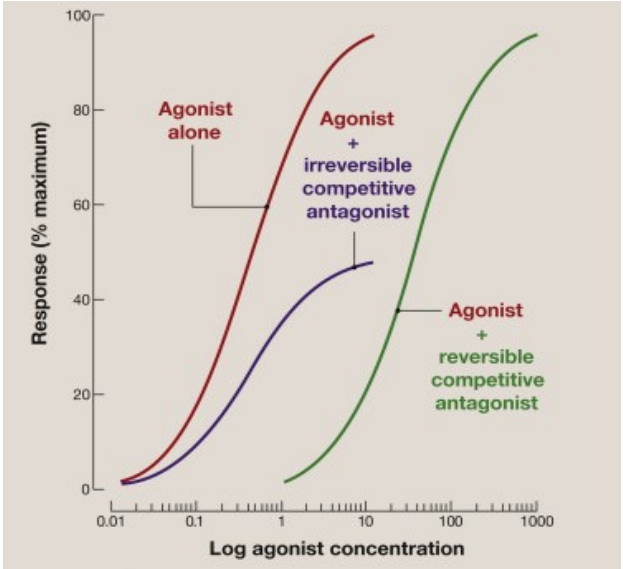

Figure 1.2: While an irreversible (or non-competitive) ant reduces the efficacy of the ag, raising ag concentration (ag curve moves to the right but efficacy remains unchanged) helps one overcome the binding of a reversible competitive ant [34]

For quick early screening of chemicals for putative protein interactions and consequences, computational approaches have certain benefits. Conventional experimental approaches to find molecules interacting with receptors can be time-consuming and costly. Emerging as a cheap and efficient substitute for screening and prioritising possibly protein-active chemicals is computational modelling [11].Particularly in the modelling of ligand-binding protein activating with a growing amount of new chemical compounds synthesised, in silico computational methodologies including machine learning (ML) methods are valuable tools for identifying of agonists and antagonist [21]. From a chemogenomic standpoint, machine learning-based PLI prediction has been investigated combining interaction from both chemical and proteomic domains under a single framework. Among the notable applications are tensor-product-based features mixed with support vector machines (SVM), minimisation of Euclidean distances across common features generated from ligand and protein mappings, and 3D grids with 3D convolutional networks. Approaches also have used deep learning with convolutional neural networks to predict on protein sequences, Bayesian additive regression trees for PLI prediction, and convolutional networks for proteins coupled with graph networks for ligands. Although these approaches give fresh understanding of PLI prediction, they are not a strong basis for direct use in drug development [23], Though the development of Machine Learning (ML), and Deep Learning (DL) techniques has opened a route for solving hitherto difficult unsolvable issues related to biology and chemistry [23]. Any artificial intelligence or statistical method seeks to find trends within training data that might be applied to create predictions outside of the training set. ML models are naturally adaptable and with enough well chosen data as such algorithms have the ability to identify patterns that individuals cannot detect on their own [19] [11]. An algorithm does this by mappings the training set depiction to a space in which inactive and active ligands separate. This conversion can be assisted by physical models but is more typically learnt directly from the data without inclusion of detailed physicochemical knowledge [19]. Because soon the model has been trained the users are just needed to submit normal protein and ligand descriptions as input, end-to- end learning, ML/DL approach, has attracted interest in recent years. The end-to–end learning approach consists in (i) embedding data to lower dimensions, (ii) building different neural networks based on the available data, and (iii) applying backpropagation over the full architecture to minimise loss and update weights [23]. Often used to supplement experimental studies to help with data interpretation and enhance results are computational methods [13].

Current machine learning (ML) and deep learning (DL) methods for predicting protein-ligand interactions (PLIs) face three main challenges, which result in relatively low accuracy. First, the lack of high-resolution

structures of protein-ligand complexes limits the effectiveness of these methods. Second, the 3D grid representation of the target can become an enormous and sparse matrix, which hinders the ability of ML and DL models to learn and predict interactions effectively. Third, the variety of methods used to represent ligand structures complicates the selection of the most effective ligand representation.

To address these issues, researchers have focused on developing techniques to compress the 3D target and ligand structures into 1D representations. This approach creates a denser structure that allows ML and DL models to work with a smaller set of input features. By reducing 3D structural information to 1D, the models can better learn the spatial features of protein secondary structures, thus improving the understanding of ligand interactions [23]. However, predicting interactions remains challenging due to difficulties in identifying suitable chemical descriptors and optimizing machine learning techniques such as deep learning [11].

1.2 Problem Statement

Rational identification of selective modulators would benefit much from the ability to distinguish the agonistic or antagonistic behaviour of substances [39]. Complexity increases in situations when some ligands may function as an agonist in some settings and as a antagonist in others [3]. Although computational techniques for drug development have advanced significantly, existing models for identifying agonists and antagonists may depend on training with particular receptor types, therefore restricting their generalisability across many receptor-ligand couples. This presents a difficulty in creating a universal computational tool based on molecular binding reactions that can precisely characterise agonist and antagonist interactions. Furthermore, it is still difficult to pinpoint important interaction characteristics that set agonist from antagonist behaviour apart over different protein-ligand complexes. Therefore, a thorough computational tool that can not only categorise the mechanism of the relationship between protein-ligand complexes in a generalised manner, irrespective of receptor type, but also extract and analyse the significant features guiding this categorisation.

1.3 Related work

The classification of protein-ligand complexes as either agonists or inhibitors for specific receptors has been widely explored. A notable example involves the binary and multi-class classification of androgen receptor agonists, antagonists, and binders. This study utilized a comprehensive chemical database from the CoMPARA program and applied a multi-class classification approach to differentiate between agonists, antagonists, and inactive compounds. The models achieved 80% accuracy in predicting agonists, with antagonists and binders showing accuracies of 72% and 78%, respectively. Combining agonists, antagonists, and inactive molecules in a multi-class approach yielded 64% accuracy for the evaluation set [9].

DeepSnap-DL represents another DL-based QSAR system that extracts molecular characteristics from 3D chemical structure images [33]. It has demonstrated superior performance compared to traditional machine learning methods such as random forest, XGBoost, LightGBM, and CatBoost [20] [21]. In recent research, prediction models for glucocorticoid receptor antagonists, TGF-beta/Smad antagonists, and thyrotropin-releasing hormone receptor agonists were developed using DeepSnap-DL with DIGITS and an enhanced version, DeepSync-DL, which leverages TensorFlow and Keras. These models were tested on the Tox21 10k library [22]. Additionally, a molecular docking approach was used to predict agonists and antagonists for the estrogen receptor subtype alpha. An in silico method was developed to differentiate between potential binders as agonists or antagonists by using both configurations of the estrogen receptor α [13].

G-protein coupled receptors (GPCRs), which play crucial roles in various physiological processes, are targeted by about 50% of all drugs. GPCR activity is influenced by specific ligands, most of which function as either agonists or antagonists [14]. A ligand-based machine learning approach was designed to identify new human GPCR agonists and antagonists regardless of GPCR type. This two-step algorithm first detects ligands that bind to GPCRs and then classifies them as agonists or antagonists. The combined models exhibited strong performance metrics, with an AUC of 0.795, accuracy of 0.733, sensitivity of 0.716, and specificity of 0.744 [26].

Another approach utilized a Bayesian classification model to predict whether a chemical acts as a GPCR agonist or antagonist. This model employed chemical fingerprints, specifically FCFP-6, from 6,627 experimentally defined compounds classified across all GPCR classes. It identified structural differences between agonists and antagonists: GPCR antagonists often contained planar groups and aromatic amines, whereas GPCR agonists were more flexible and included aliphatic amines. The model achieved test set accuracy of 89.2% and training set accuracy of 90.1%, with a Matthew's Correlation Coefficient (MCC) of 0.806 and an AUC of 0.957. This model is particularly useful in early drug development for categorizing molecules as GPCR agonists or antagonists [14].

DeepREAL is an advanced deep learning system designed for multi-scale modeling of genome-wide ligand-induced receptor activity. It uses pre-trained binary interaction classifications and supervised learning on

millions of protein sequences to predict if a chemical will act as an agonist, antagonist, or not bind at all [29]. In one study involving 258 compounds, thyroid hormone receptor agonists and antagonists were identified using statistical learning techniques. Various models, including C4.5, random forest (RF), and support vector machine (SVM), were evaluated. The C4.5 model showed moderate performance with accuracies of 93.2% for agonists and 57.8% for antagonists, resulting in an average of 83.1% correctly classified compounds in the test set. The RF model achieved average accuracies of 84.0% in cross-validation and 87.1% in external validation. The SVM model performed the best, with accuracies of 96.6% and 97.2% for cross-validation and external validation, respectively [7]. Another study used an SVM model to classify agonists and antagonists of the 5-HT1A receptor, optimizing key parameters with a genetic algorithm (GA) and selecting the most relevant descriptors. Analyzing 284 diverse ligands, the model achieved prediction accuracies of 0.954 for the training set and 0.927 for the test set, with ROC curve AUCs of 0.883 for the training set and 0.906 for the test set [36].

Synthetic truncated nucleoside derivatives were developed and shown to be potent agonists of the human A3 adenosine receptor (hA3AR) with a Ki of 2.40 nM. A minor chemical modification shifted the compounds from antagonist to agonist activity. This change was explained by new hA3AR homology models incorporating ligand pharmacological characteristics. The findings indicated that hydrogen bonding with Thr943.36 likely stabilizes the interaction between the 3'-amino group and TM3 and TM7, with induced-fit effects playing a key role in the agonistic effect. These conclusions were supported by molecular dynamics (MD) simulations and 3D structural network analysis of the receptor-ligand complex [39].

Tool Name	Receptor	Dataset size	Model
Binary-multiclass classification	Androgen receptor	5273 chemicals	Random forest
Deep-Snap DL	Nuclear receptor	7581 compounds	DL
Competitive docking approach	Estrogen receptor *aplpha*	66 ligands, 106 ER Binders, 4018 ER decoys	CDA + SDM
Ligand-Based ML model	GPCR	758 Ag + 950 Ant	Random Forest
Bayesian Model	GPCR	6627 experimentally determined compounds	Bayesian approach
DeepReal	GPCR	GLASS, IUPHAR, Pfam	DL
Thyroid hormone receptor classification	thyroid hormone	258 compunds	C4.5, RF, SVM

Table 1.1: Summary of all the developed ML/DL models for prediction of agonist and antagonists response.

Chapter 2

Materials and Methods

2.1 Data collection and Preproccesing

2.1.1 PDB based Crystal structure dataset

The Protein Data Bank (PDB), the global archive for structural data on biological macromolecules, is the primary data source for this work [5]. By means of experimental techniques including X-ray crystallography, NMR spectroscopy, and, progressively, cryo-electron microscopy, biologists and biochemists worldwide acquire and deposit these structural data. The PDB file format was the first one the PDB adopted. One of several free and open source computer programs allows one to view the structural files: Jmol, Pymol, VMD, Molstar and Rasmol.

Users can access the PDB database through its website or programmatically via its web services or APIs. A keyword search was performed to retrieve the PDBs associated with the terms "Agonist" and "Antagonist". The search query typically retrieves entries (PDB IDs) where these terms are mentioned in the metadata, such as in the title, description, or keyword fields associated with the molecular structure. These terms indicate the biological activity of the molecule in relation to a specific receptor or enzyme. The results of the keyword search provide a list of PDB entries that match the specified criteria. Each entry includes detailed information about the molecular structure, experimental conditions, and any annotations related to its biological function.

2082 PDBs associated with the term "Antagonist" and 2472 PDBs associated with the term "Agonist" were downloaded. The dataset was reduced to 2192 agonists and 2054 antagonists by excluding the PDBs associated with ligands that produced partial agonist, partial antagonist, and inverse agonist responses. The dataset was reduced to 2192 agonists and 2054 antagonists by excluding the PDBs associated with the ligands giving partial agonist, partial antagonist, and inverse agonist responses. cleaned out all the undesirable HETATMs from the PDB including DNA, RNA and ATP as well as crystallizing agents such as NAG, PEG, amino acids associated as ligands and glycerol along with the ions like SO_4^{2-} and Ca^{2+}.

The PDB structure was linked to a number of small molecules, not all of which produced agonistic or antagonistic response. We needed to identify a unique pair of proteins and the ligand associated with the protein, which gave specific responses such as agonist or antagonist. In order to ascertain whether the ligand binding to the protein is eliciting the intended response, we manually curated each PDB using abstracts and literature searches. We discovered not all the ligands bound to protein had the desired result, thus those PDB were eliminated. We now have 787 agonist and 617 antagonist in our final dataset.

Steps performed	**Agonist**	**Antagonist**
Performed Keyword search to fetch PDB ids with word 'Agonist' and 'Antagonist'.	2472	2082
Removed partial Ag and Ant.	2192	2054
There were many ligands in a single structure, hence, we did manual curation to keep only desired ligands.	1123	803
Kept only unique protein-ligand pairs.	856	752
Removed inverse Ag and Ant	787	619

Table 2.1: Workflow for PDB Data Curation

2.1.2 Literature-Based Receptor-Ligand Docking Dataset

We compiled a distinct dataset comprising docked complexes to validate our model's predictive capabilities. To construct this dataset, we conducted an extensive literature review and database mining to collect data on various receptors and their corresponding agonists and antagonists. Utilizing databases such as the IUPHAR/BPS Guide to PHARMACOLOGY and RCSB Protein Data Bank, a dataset of 745 unique receptor-ligand pairs across 24 different receptors was compiled. Bioinformatics tools and molecular docking software, including RCSB PDB (Protein Data Bank) CASTp, UCSF Chimera, OpenBabel, AutoDock Vina, and MGLTools, were employed to prepare and process these receptor-ligand pairs for virtual screening. Finally, virtual screening for 686 unique receptor-ligand pairs was performed. The ultimate goal was to get the docking output with different poses and binding affinities of the ligand with the receptor and create the best ligand-receptor pose output file. This dataset will further be used as a testing dataset for a classification-based Machine Learning model [10].

CastP

We used the Computed Atlas of Surface Topography of proteins (CASTp) program to find and examine the ligand-binding pockets of every receptor in our investigation [6]. For each receptor, we submitted the protein structure to the CASTp server and performed pocket identification. The analysis included the determination of pocket volume and surface area, as well as the identification of key residues lining the pockets. The data obtained from CASTp was used to guide the selection of potential ligand binding sites and to validate the docking results, ensuring that our predictions were consistent with the known structural features of the receptors.

UCSF Chimera

UCSF Chimaera, a highly extendable tool for interactive visualisation and study of molecular structures and associated data, was used to visualise and analyse docked structures [1]. Docked receptor-ligand complexes were visualized to assess the binding interactions and confirm the accuracy of the docking poses. Chimera was used to render the molecular surface and calculate the volume of the binding pockets. This visualization was essential to complement the pocket identification performed using CASTp and to provide a clearer understanding of the spatial arrangement of the binding sites. To compare different receptor conformations and ligand binding modes, Chimera's structure superimposition tools were utilized. This allowed for the alignment of multiple structures to identify common features and differences in binding sites. High-quality images and animations of the receptor-ligand complexes were generated using Chimera for use in presentations and publications. This helped in effectively communicating the structural details and findings of the study.

Open Babel

We used Open Babel, an open-source chemical toolbox made to speak the several languages of chemical data, to prepare the ligands for docking investigations. Together with a programming library, the Open Babel package consists of a series of user apps [25]. The downloaded ligand structures in SDF (Structure-Data File) format were converted to MOL2 and PDBQT formats using Open Babel. The following commands were used to convert the SDF files into a different file formats.

- **SDF to MOL2**: `obabel input.sdf -O output.mol2`

- **SDF to PDBQT**: `obabel input.sdf -O output.pdbqt`

For the filtered molecules, the copy format recommended by "-o copy" is a utility format that directly replicates the precise contents of the input file (for the output without perspective or interpretation) [25]. During the conversion process, Open Babel was also used to optimize the ligand structures. This involved removing any unwanted ions and correcting structural anomalies to ensure that the ligands were in the correct form for docking. The PDBQT files generated by Open Babel included the necessary atomic charges and rotatable bonds, which are essential parameters for docking with AutoDock Vina. By employing Open Babel for these tasks, we ensured that the ligand structures were accurately converted and optimized for subsequent docking studies. This process facilitated the efficient and reliable preparation of ligands, contributing to the overall robustness of our computational analysis.

Autodock Vina

AutoDock Vina, a powerful and efficient molecular docking tool, was employed for both preparing the receptor files and performing the molecular docking. Arguably among the fastest and most extensively used open-source applications for molecular docking is AutoDock Vina [15]. It greatly increases the binding mode forecasts'

average accuracy above AutoDock 4's. Vina views the PDBQT created (interactively or in batch mode) and uses it for input and output using MGLTools. The receptor structures were converted to PDBQT format using AutoDock tools, ensuring the addition of necessary atomic charges and the specification of rotatable bonds. The prepared ligands in PDBQT format were docked into the receptor binding sites using AutoDock Vina [24].

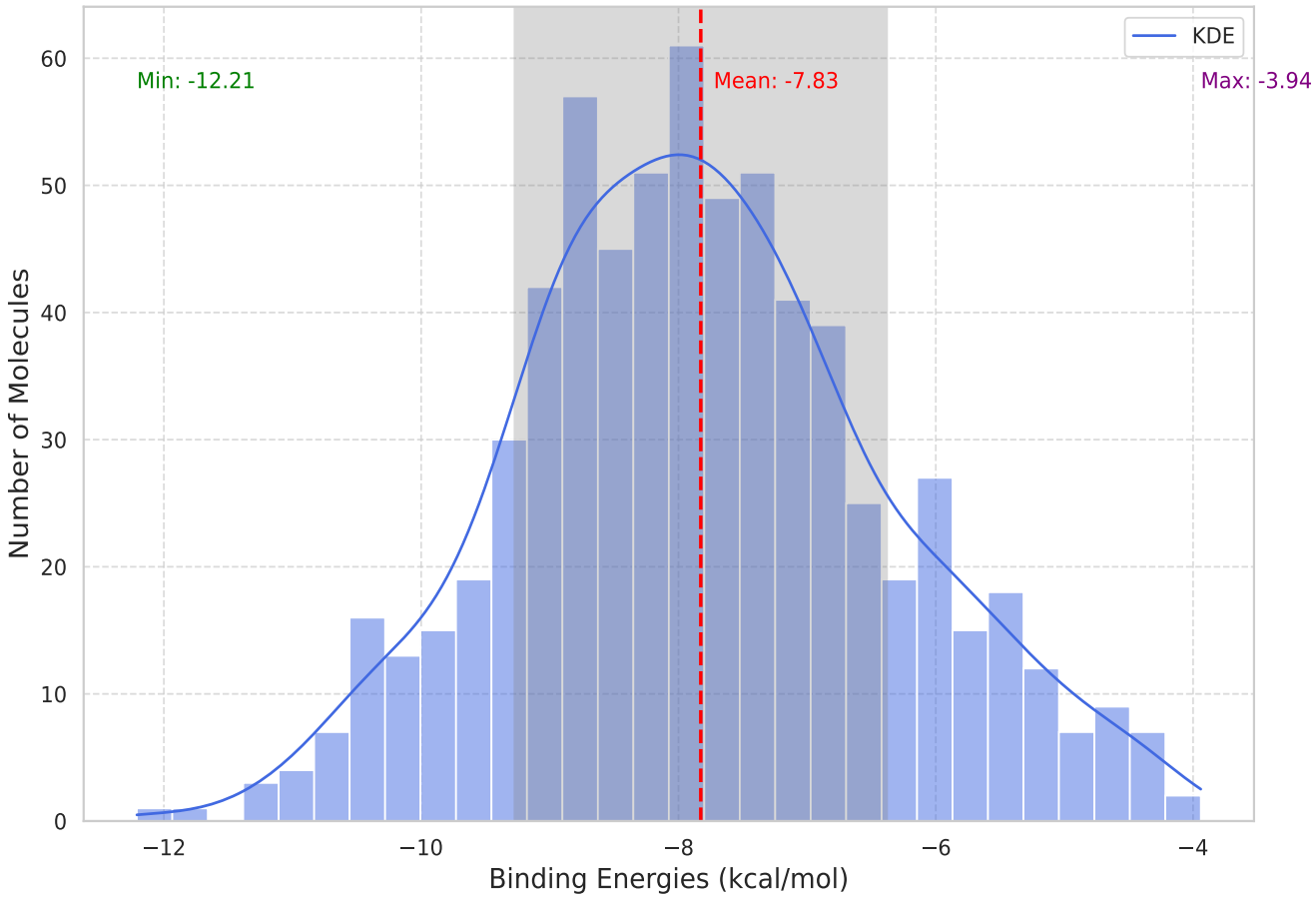

Figure 2.1: A. Types of features extracted using various tools, B. Splitting dataset into Train(70%), Test(15%) and Validation(15%) sets, C. Comparative analysis of top 10 machine learning models

In addition to our initial dataset, we compiled a literature-based docking dataset, which includes 686 docked complexes across 18 different receptors. Figure 2.2 is the binding energy distribution plot which shows a normal distribution with a mean of -7.83 kcal/mol and a standard deviation of approximately 2.5 kcal/mol. The plot indicates that the majority of the binding energies fall within the range of -10 to -6 kcal/mol. The distribution is slightly skewed to the right, suggesting that there are a few complexes with higher binding energies. The minimum binding energy is -12.21 kcal/mol and the maximum is -3.94 kcal/mol. The resulting docked complexes provided a valuable resource for validating our computational models and understanding the molecular basis of receptor-ligand interactions. This dataset not only enhances the robustness of our study by incorporating experimentally validated interactions but also offers insights into the binding mechanisms of diverse receptor-ligand systems.

2.2 Feature Extraction and Feature Engineering

We calculated a comprehensive set of 1918 features for both proteins and ligands using various computational tools and software. These features were critical for our subsequent feature selection process. To extract the necessary features for analyzing protein-ligand interactions, we employed a range of specialized tools and softwares like PaDEL [37] for calculating molecular descriptors and fingerprints for ligands, dPocket [17] for pocket detection and characterization in protein structures, NACCESS [12] applied to determine the accessible surface

areas of proteins, LigBuilder [38] to obtain the cavity's PDB file, DSSP [35] [16] was employed for secondary structure assignment and calculation of protein structural features and finally biopython was utilized to calculate additional features like average H-bond distance, average B-factor, pi-pi stacking, average aromatic residue distance of whole PDB as well as cavity PDB.

2.2.1 Computation of Ligand Molecular Descriptors with PaDEL

A molecular descriptor is a way of turning information about a molecule's structure into a numerical value or experimental result. This helps in understanding and comparing different molecules [31]. To predict how new chemicals will behave biologically, researchers use molecular descriptors to create QSAR models. These models help quantify and understand the relationship between a chemical's structure and its activity [37]. There are several types of molecular descriptors today. They can be grouped into three main categories: 1D,2D, 3D [32]. PaDEL-Descriptor offers both a graphical user interface (GUI) and command line options. PaDEL-Descriptor is compatible with over 90 different molecular file formats [37].

We had a collection of ligand IDs for which we retrieved the corresponding SMILES (Simplified Molecular Input Line Entry System) data. This SMILES data served as the input for the PaDEL-Descriptor software. PaDEL utilized these SMILES representations to compute a wide array of descriptors that encapsulate the chemical properties of the ligands. These descriptors included both 1D and 2D descriptors, as well as 3D descriptors. Specifically, the software computes 431 three-dimensional (3D) descriptors, which capture spatial and geometric properties of the molecules. Additionally, it generates 1444 one-dimensional (1D) and two-dimensional (2D) descriptors, such as apol, naAromAtom, nAromBond, nAtom, and nHeavyAtom, which encompass various aspects of the chemical structure and properties. In total, PaDEL-Descriptor calculated 1875 chemical descriptors for each ligand, providing a detailed and comprehensive profile of their chemical characteristics.

2.2.2 Ligand Pocket Detection and Featurization Using Dpocket

A major problem of such procedure that has been the focus of growing numbers of studies in the previous decade is the detection and characterisation of pockets and cavities of a protein structure. Though this rule cannot be generalised, the largest pocket often corresponds to the observed ligand binding location [17]. A thorough understanding of pocket features helps one to identify which geometrical (clustering) and chemical (amino acidity content) qualities define a desirable binding site. Furthermore, the approach offers a division of pockets into smaller subpockets [18].

In this work, we featurized protein structures using the well-known computational program Dpocket. Dpocket is meant to arrange descriptor gathering from a set of co-crystallized complexes and find and describe protein binding pockets [17]. Dpocket only needs one single input file. This input file has to be a text file including the following data:

- The PDB file of the protein you want to study.
- The ligand ID would you like to use as a reference to define a specific binding pocket

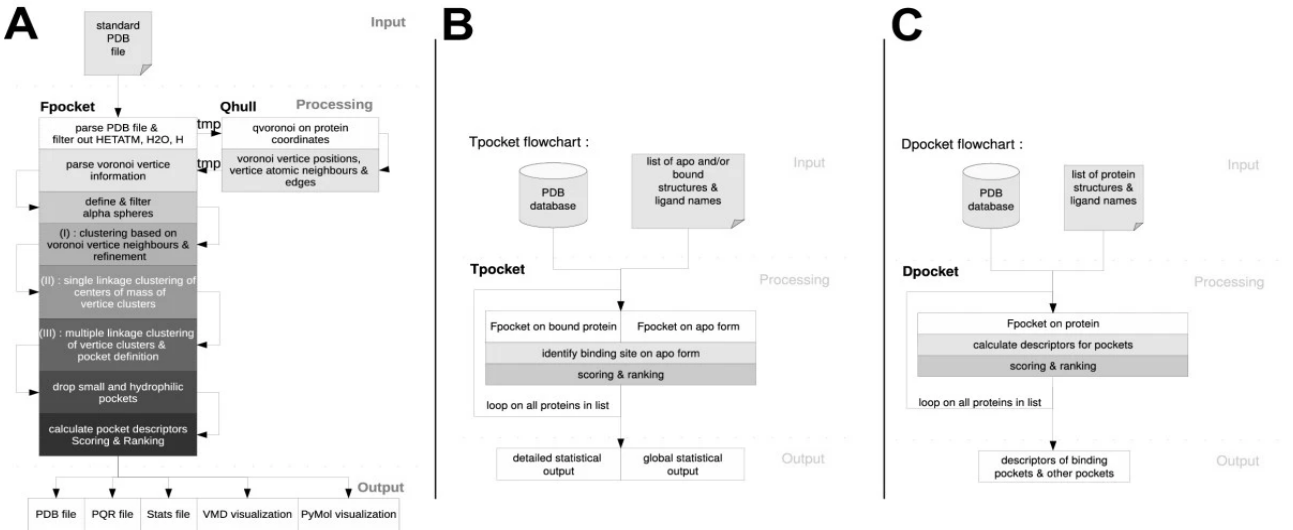

Figure 2.2: flowchart illustrating the workflow of Fpocket, pocket and dpocket [17].

The file used in this (`/10tb-storage/riddhis/DPOCKET.txt`) looks like this:

```
data/3LKB.pdb pc1
data/1ATK.pdb atp
data/7TAU.pdb ABC
```

Dpocket can be launched on this file using the following command:

```
dpocket -f /10tb-storage/riddhis/DPOCKET.txt
```

Dpocket generates three default result files in the current directory:

```
dpout_explicitp1.txt
dpout_fpocketnp1.txt
dpout_fpocketp1.txt
```

Dpocket generates three output files in the current directory by default, each containing pocket descriptors identified by fpocket:

- **fpocketp.txt**:Lists all binding sites discovered by fpocket that fit one of the detection criteria. Stated differently, fpocket discovered numerous pockets in the protein; so, this file will contain descriptors of pockets deemed to be the binding pocket depending on various detection criteria.
- **fpocketnp.txt**: describes, on the other hand, all pockets discovered by fpocket that, following the detection criteria, are not deemed to be the genuine pocket.
- **explicitp.txt**: All pocket descriptors used by fpocket for the explicitly specified binding pocket are contained in this file. Explicitly, in this context, refers to the fact that every PDB file in the input contains a ligand identifier connected with. Fpocket finds the real binding pocket by means of this ligand.

We obtained 55 protein pocket features from Dpocket, including examples such as ligand volume, pocket volume, hydrophobicity score, mean local hydrophobic density, and polarity score.

2.2.3 Utilizing LigBuilder for Detailed Cavity Feature Analysis

Designed for structural-based de novo drug creation and optimisation, Lig Builder V2.0 is a multifunctional tool. One of its modules, cavity plus, searches and examines the ligand bonfng site of the target protein. Calculated as an index of the ligandicity of the binding site is cavity score. CavityPlus identifies potential binding sites on a protein's surface and ranks them based on how likely they are to interact with drugs, using the protein's 3D structure as input. Three submodules— CavPharmer, CorrSite, and CovCys—allow one to examine these possible binding sites more closely. CavPharmer automatically extracts pharmacophore features inside cavities using a receptor-based pharmacophore modelling tool, Pocket. Motion correlation studies between cavities allow CorrSite to detect possible allosteric ligand-binding sites. Specifically helpful for determining new binding sites for covalent allosteric ligands, CovCys automatically finds druggable cysteine residues.refer to [40].

LigBuilder was utilized in whole protein mode to analyze and retrieve the PDB files corresponding to various cavities within the protein structure. This process generated multiple PDB files, each representing a distinct cavity. Among these, the PDB file corresponding to our ligand's cavity exhibited the highest drug score, indicating its superior potential for druggability. Consequently, we selected the surface.pdb file associated with this cavity to extract detailed protein-ligand cavity features for further analysis.

- **Program Dependency:**
  - Ensure that OpenBabel version 2.3.0 or later is installed, and verify that the path to `babel` has been added to the system's executable path. Although OpenBabel is not essential for the design process, its absence will affect the following functions: Automatic mode, 2D-Clustering, Clustering Reporter, and Synthesis
- **Environment Setting:**
  - Add the following line to your `/.bashrc`, `/.cshrc`, or `/.tcshrc` file, depending on your shell:

    ```
    ulimit -s unlimited
    ```

  - Since the program will call `babel` and `obabel` directly, ensure the path to OpenBabel is added to the environment variable if `babel` is not executable in the working directory. This step can be skipped if OpenBabel is installed with root permissions.

- **Installation of LigBuilder V3:**
  - Unpack the package in a Linux shell using the following command:

    ```
    tar -xvf LigBuilderV3.tar.gz
    ```

- **Steps:**
  1. Take the PDB format of the cavity inside the receptor folder, which is located inside the example folder of LigBuilder v2, and run it in PyMOL.
  2. Select for sequence (S) and proceed towards the end of the sequence (UNK), select the UNK (ligand) and save it in .mol2 format in the same folder.
  3. Go to the cavity.input file present in the example folder, change the names of the cavity and ligand to the ones we have taken, and select ligand detection mode (1).
  4. After giving the proper path of the folder, run the command `./cavity cavity.input` in the terminal.
  5. The above step will generate specific files for surfaces, grids, pockets, key sites, and suggest a pharmacophore model inside the receptor folder.
  6. Go to PyMOL and view the .mol2 file along with the surface files. The proper bound surface is selected.

2.2.4 Quantifying Surface and Solvent Accessibilities with NACCESS

The atomic accessible surface area was calculated using Naccess program. This computation rolled a probe over the Van der Waals surface of the cavity PDB file acquired by ligbuilder as well as the complete protein. The path followed by the probe's centre defines the accessible surface [12].

```
REM  Relative accessibilites read from external file "/home/iiitd/softwares/standard.data"
REM  File of summed (Sum) and % (per.) accessibilities for
REM RES _ NUM      All-atoms   Total-Side   Main-Chain    Non-polar    All polar
REM               ABS   REL    ABS   REL    ABS   REL    ABS   REL    ABS   REL
RES LEU A 306   118.99  66.6  77.83  55.1  41.17 109.8  79.52  55.9  39.48 108.7
RES ALA A 307    15.56  14.4  11.00  15.8   4.57  11.8  12.28  17.2   3.28   9.0
RES LEU A 308    70.12  39.3  56.65  40.1  13.47  35.9  57.64  40.5  12.48  34.4
RES SER A 309    91.63  78.7  66.53  85.2  25.10  65.4  51.46 106.0  40.16  59.1
RES LEU A 310    45.77  25.6  42.52  30.1   3.26   8.7  43.59  30.6   2.18   6.0
RES THR A 311    77.25  55.5  76.65  75.4   0.60   1.6  74.05  97.8   3.19   5.0
RES ALA A 312    19.49  18.1  12.44  17.9   7.05  18.3  12.47  17.5   7.02  19.2
RES ASP A 313    84.07  59.9  81.23  79.1   2.84   7.5  39.50  80.2  44.57  48.9
RES GLN A 314    88.34  49.5  86.49  61.3   1.85   4.9  32.38  62.0  55.95  44.3
RES MET A 315     2.12   1.1   2.12   1.4   0.00   0.0   2.12   1.3   0.00   0.0
RES VAL A 316    15.89  10.5  15.55  13.6   0.34   0.9  15.55  13.5   0.34   0.9
RES SER A 317    64.64  55.5  64.17  82.2   0.47   1.2  31.86  65.6  32.78  48.2
RES ALA A 318    33.02  30.6  30.51  43.9   2.51   6.5  31.08  43.5   1.94   5.3
RES LEU A 319     0.74   0.4   0.74   0.5   0.00   0.0   0.74   0.5   0.00   0.0
```

Figure 2.3: An example .rsa file are shown above.

```
ACCALL - Accessibility calculations
MAX RESIDUES     8000
MAX ATOMS/RES     100
PDB FILE INPUT 3RHW.pdb
PROBE SIZE        1.40
Z-SLICE WIDTH   0.050
VDW RADII FILE /home/iiitd/softwares/vdw.radii
EXCL HETATOMS
EXCL HYDROGENS
EXCL WATERS
READVDW  32 residues input
ADDED VDW RADII
CHAINS       15
RESIDUES  3697
ATOMS    28730
```

Figure 2.4: an example .log file

figure 2.5, The log file of Naccess provides a comprehensive summary of the calculation process, including initialization details such as the input PDB file name, probe size, Z-slice width, and van der Waals radii file. It specifies whether HETATM records, hydrogens, and waters are included or excluded. The file also summarizes input data, including the number of residues, atoms, and chains, and notes any non-standard atoms with their guessed van der Waals radii. The calculation summary confirms that atomic accessibilities were computed, lists the relative accessibilities read for amino acids, and provides summed accessibility data over residues.

2.2.5 Secondary Structure Analysis Using DSSP Tool

Using the DSSP technique for both the whole protein and the isolated cavity PDB, secondary structural elements were investigated. This study gave thorough understanding of the secondary structure elements—alpha-helices, beta-sheets, loops—present in the entire protein as well as inside the particular cavity region. The DSSP algorithm computes, from a protein's 3D structure, the most likely secondary structure assignment. This is accomplished by reading the atom positions in a protein then computing the H-bond energy between every atom [35] [16].

- **Usage and command line:**
  - `mkdssp -i my-pdb.ent -o my-ss.dssp`

- `-na`: Disables the calculation of accessible surface.
- `-c`: Outputs data in the classic format.
- `-v`: Provides verbose output.
- `--`: Reads input from standard input.
- `-h` or `-?`: Prints help information.
- `-l`: Prints license information.
- `-V`: Prints version information.

```
==== Secondary Structure Definition by the program DSSP, CMBI version 2.0                          ==== DATE=2023-08-07        .
REFERENCE W. KABSCH AND C.SANDER, BIOPOLYMERS 22 (1983) 2577-2637                                                          .
HEADER    TRANSCRIPTION                           04-JUL-15   5AWJ                                                         .
COMPND    MOL_ID: 1; MOLECULE: VITAMIN D3 RECEPTOR,VITAMIN D3 RECEPTOR; CHAIN: A                                           .
SOURCE    MOL_ID: 1; ORGANISM_SCIENTIFIC: RATTUS NORVEGICUS; ORGANISM_COMMON: RA                                           .
AUTHOR    Y.ANAMI,T.ITOH,Y.INABA,M.NAKABAYASHI,T.IKURA,N.ITO,K.YAMAMOTO                                                    .
  251  3  0  0  0 TOTAL NUMBER OF RESIDUES, NUMBER OF CHAINS, NUMBER OF SS-BRIDGES(TOTAL,INTRACHAIN,INTERCHAIN)           .
 12510.5   ACCESSIBLE SURFACE OF PROTEIN (ANGSTROM**2)                                                                    .
  192 76.5   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(J)  , SAME NUMBER PER 100 RESIDUES                         .
    0  0.0   TOTAL NUMBER OF HYDROGEN BONDS IN     PARALLEL BRIDGES, SAME NUMBER PER 100 RESIDUES                         .
    6  2.4   TOTAL NUMBER OF HYDROGEN BONDS IN ANTIPARALLEL BRIDGES, SAME NUMBER PER 100 RESIDUES                         .
    2  0.8   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I-5), SAME NUMBER PER 100 RESIDUES                         .
    0  0.0   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I-4), SAME NUMBER PER 100 RESIDUES                         .
    0  0.0   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I-3), SAME NUMBER PER 100 RESIDUES                         .
    0  0.0   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I-2), SAME NUMBER PER 100 RESIDUES                         .
    0  0.0   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I-1), SAME NUMBER PER 100 RESIDUES                         .
    0  0.0   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I+0), SAME NUMBER PER 100 RESIDUES                         .
    0  0.0   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I+1), SAME NUMBER PER 100 RESIDUES                         .
    8  3.2   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I+2), SAME NUMBER PER 100 RESIDUES                         .
   39 15.5   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I+3), SAME NUMBER PER 100 RESIDUES                         .
  135 53.8   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I+4), SAME NUMBER PER 100 RESIDUES                         .
    3  1.2   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I+5), SAME NUMBER PER 100 RESIDUES                         .
  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30     *** HISTOGRAMS OF ***      .
  0  0  0  1  1  0  2  0  0  0  0  1  0  0  0  1  1  0  0  2  0  1  0  0  0  0  0  1  0  0    RESIDUES PER ALPHA HELIX    .
  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0    PARALLEL BRIDGES PER LADDER .
  0  2  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0    ANTIPARALLEL BRIDGES PER LADDER .
  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0    LADDERS PER SHEET          .
  #  RESIDUE AA STRUCTURE BP1 BP2  ACC     N-H-->O    O-->H-N    N-H-->O    O-->H-N    TCO  KAPPA ALPHA  PHI   PSI    X-CA   Y-CA   Z-CA
    1  123 A K              0   0  174      0, 0.0     2,-2.3     0, 0.0   181,-0.1   0.000 360.0 360.0 360.0 -20.5   48.4  -19.4   -2.3
    2  124 A L        -     0   0   35    180,-0.1     5,-0.2   181,-0.1   177,-0.1  -0.437 360.0-152.1 -67.9  78.2   44.9  -20.8   -2.0
    3  125 A S    >>  -     0   0   58     -2,-2.3     4,-1.3     1,-0.1     3,-0.9  -0.004  18.7-115.6 -50.6 158.9   43.7  -18.3    0.6
    4  126 A E  H 3> S+     0   0  162      1,-0.3     4,-0.9     2,-0.2    -1,-0.1   0.805 114.8  60.3 -72.7 -33.8   40.9  -19.4    3.0
    5  127 A E  H 3> S+     0   0  117      1,-0.2     4,-1.0     2,-0.2    -1,-0.3   0.642 107.6  50.9 -60.6 -18.5   38.4  -16.9    1.8
    6  128 A Q  H <> S+     0   0   24     -3,-0.9     4,-2.2     2,-0.2    -2,-0.2   0.831 103.8  52.3 -90.9 -40.5   38.8  -18.7   -1.6
    7  129 A Q  H  X S+     0   0  101     -4,-1.3     4,-1.4     1,-0.2    -2,-0.2   0.674 110.7  53.0 -71.2 -18.8   38.3  -22.3   -0.5
    8  130 A H  H  X S+     0   0  119     -4,-0.9     4,-3.0     2,-0.2    -1,-0.2   0.899 105.0  51.3 -77.8 -50.0   35.1  -21.0    1.0
    9  131 A I  H  X S+     0   0    5     -4,-1.0     4,-2.4     1,-0.2    -2,-0.2   0.899 112.0  50.0 -47.5 -44.7   34.0  -19.4   -2.3
   10  132 A I  H  X S+     0   0   13     -4,-2.2     4,-3.0     1,-0.2     5,-0.3   0.923 109.6  48.1 -66.0 -48.0   34.7  -22.8   -3.9
   11  133 A A  H  X S+     0   0   51     -4,-1.4     4,-2.6     1,-0.2     5,-0.2   0.920 112.4  50.6 -59.2 -41.8   32.7  -24.7   -1.4
   12  134 A I  H  X S+     0   0   52     -4,-3.0     4,-2.8     1,-0.2    -2,-0.2   0.919 112.6  45.9 -59.2 -46.3   29.8  -22.2   -1.8
   13  135 A L  H  X S+     0   0    0     -4,-2.4     4,-2.1     2,-0.2    -2,-0.2   0.875 113.0  49.2 -70.1 -39.8   29.8  -22.6   -5.5
   14  136 A L  H  X S+     0   0   26     -4,-3.0     4,-2.7     2,-0.2     5,-0.2   0.930 113.0  46.7 -62.7 -45.4   30.0  -26.4   -5.5
   15  137 A D  H  X S+     0   0   72     -4,-2.6     4,-2.8    -5,-0.3    -2,-0.2   0.971 111.4  53.2 -65.6 -45.6   27.2  -26.7   -3.0
   16  138 A A  H  X S+     0   0    0     -4,-2.8     4,-0.6    -5,-0.2    -1,-0.2   0.882 111.1  45.7 -42.7 -51.0   25.2  -24.2   -5.1
   17  139 A H  H >X S+     0   0    3     -4,-2.1     4,-1.3     2,-0.2     3,-1.0   0.907 110.2  53.5 -67.4 -42.3   25.7  -26.3   -8.2
   18  140 A H  H 3< S+     0   0   97     -4,-2.7     3,-0.4     1,-0.3     -2,-0.2   0.911 109.4  49.6 -59.1 -44.1   24.9  -29.5   -6.3
   19  141 A K  H 3< S+     0   0  133     -4,-2.8    -1,-0.3    -5,-0.2    -2,-0.2   0.622 118.8  38.2 -66.9 -14.9   21.6  -28.0   -5.2
   20  142 A T  H << S+     0   0   11     -3,-1.0     2,-0.5    -4,-0.6    -2,-0.2   0.392 107.2  64.0-119.7   3.8   20.7  -26.8   -8.8
```

Figure 2.5: output from DSSP containing secondary structure assignments and other information.

2.2.6 Feature selection

Feature selection means choosing which input variables to keep when creating a predictive model, in order to simplify the model and improve its performance. This step is crucial because it can significantly decrease the computational cost associated with training and deploying the model. Additionally, by eliminating irrelevant or redundant features, feature selection can enhance the model's performance, leading to more accurate and reliable predictions. Simplifying the model in this way not only improves efficiency but also reduces the risk of overfitting, ensuring that the model generalizes better to new, unseen data. Effective feature selection thus plays a vital role in optimizing the balance between model complexity and performance.

Variance Threshold

Since our dataset contained numerous similar input variables, we employed the variance threshold technique to refine our feature set. Removes all features whose variance does not satisfy specified threshold using a basic baseline method for feature selection. It removes all zero-variance features by default—that is, characteristics with the same value in every sample. we had set the variance threshold at 0.8, and were able to filter out these low-variance variables, ensuring that only the most informative features were retained. This step not only streamlined our feature set but also enhanced the model's efficiency and performance by focusing on the variables with greater variability and relevance. Following is the detailed explanation of steps followed for feature selection using variance threshold method:

- **Calculate Variance for Each Feature**
  - For each feature X_j, calculate the variance:

$$\text{Var}(X_j) = \frac{1}{N}\sum_{i=1}^{N}(x_{ij} - \bar{x}_j)^2$$

- **Boolean Features Variance**

$$\text{Var}(X_j) = p_j(1 - p_j)$$

- **Set Variance Threshold**
  - Define a threshold θ. Commonly used thresholds are 0.8 or 1.0, depending on the context and requirements of the model.
- **Feature Selection**
  - Select features with variance greater than the threshold θ:

$$\text{Selected Features} = \{X_j \mid \text{Var}(X_j) > \theta\}$$

ANOVA Method

After employing the variance threshold method to eliminate low-variance features, we conducted additional feature selection using the ANOVA (Analysis of Variance) method. This statistical technique allowed us to further refine our feature set by identifying and retaining the most significant features that contribute to the variance in our data. Consequently, the final number of selected features was reduced to 275, ensuring a robust and efficient feature set. These 275 features were then used across a sample of 1,406 PDB structures, providing a comprehensive and high-quality dataset for subsequent predictive modeling. This meticulous feature selection process not only enhanced the model's performance but also reduced computational complexity, leading to more accurate and efficient predictions.

2.2.7 Exploratory Data Analysis

Exploratory Data Analysis (EDA) stands as a foundational stage in the process of building machine learning models. It involves a systematic approach to scrutinizing and interpreting data to uncover its intrinsic characteristics, identify patterns that might inform subsequent modeling decisions, pinpoint anomalies or outliers that could affect model performance, and elucidate relationships between different variables within the dataset. By employing various statistical and graphical techniques, EDA provides a comprehensive understanding of the data's distributional properties, central tendencies, variability, and potential biases or limitations. This preliminary exploration not only aids in preparing data for modeling but also helps in formulating hypotheses and refining the scope of further analyses. Ultimately, EDA serves as a crucial step in ensuring data quality, understanding its suitability for specific modeling techniques, and guiding the selection of appropriate features and preprocessing steps necessary for constructing robust machine learning models.

Data diversity

The largest category of proteins in the dataset is the "C4 Zinc Finger Nuclear Receptor" and the "G-Protein Coupled Receptor," which together make up approximately 47.7% of the total. The "Winged Helix/Forkhead Transcription Factor" is the third largest category, making up 14.8% of the total. The "Protease Inhibitor" category is relatively small, making up only 1.1% of the total. The "Ligand-Gated Ion Channel" and "Ion Channel" categories are also relatively small, making up 1.7% and 2.4% of the total, respectively. The Others class of protein makes up 2.5% of the total. Given the vast array of protein classes present in our dataset, it would be intriguing to examine the fundamental patterns of protein features that are shared among various classes. This systematic exploration may help us understand the basic concepts surrounding the biological implications of protein features and their interactions. We plotted violin plots to study the diversity of ligands and proteins for both classes. The median of violin plots for protein sequence similarity of agonist is between 0.4 and 0.5 and the median for an antagonist is between 0.4 and 0.3, which implies that the protein collection in our complete dataset is diverse. similarly we calculated tanimoto similarity for ligands to find the diversity of small molecules and then plotted a violin plot. the median of the violin plot for both agonist and antagonist classes of ligands lies in the range of 0.1.

- **Tanimoto similarity**

Tanimoto similarity, also known as the Tanimoto coefficient or Jaccard index, is a metric used to quantify the similarity between two sets. In the context of molecular informatics and cheminformatics, it is commonly used to measure the similarity between chemical compounds based on their structural fingerprints or molecular descriptors.

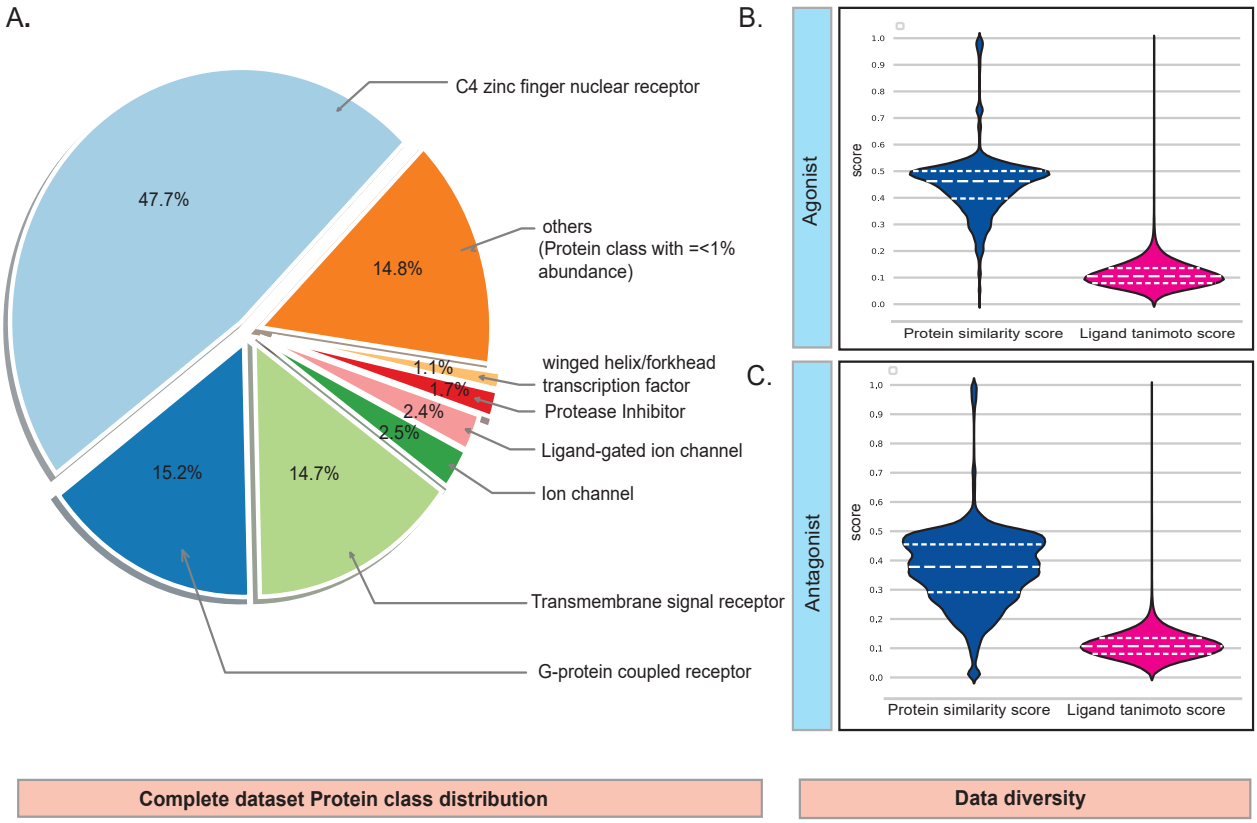

Figure 2.6: A. Pie chart showing the distribution of protein classes in complete dataset. B. Violin plot representing the protein sequence similarity (blue) and the ligand tanimoto score (pink) for agonist and C. Violin plot representing the protein sequence similarity (blue) and the ligand tanimoto score (pink) for antagonist

Visualization of Data Distribution Using t-Distributed Stochastic Neighbor Embedding (t-SNE)

T-Distributed Stochastic Neighbour Embedding (t-SNE) is a dimensionality reduction method based on a randomised approach meant to nonlinearly lower a dataset's dimensionality. It's a means of visualising high-dimensional data that underlines the preservation of the local data structure in the lower-dimensional space. This approach enables exploration of high-dimensional data by mapping it into lower dimensions while preserving local structures. Visualizing these mappings in 2D or even 3D plots allows us to gain insights into the data's structure.

We plotted t-SNE (t-distributed Stochastic Neighbor Embedding) to visualize and reveal our dataset's hidden patterns and complexity. The data points in our dataset for both classes overlap, and no distinct clusters are obtained, which implies that the data points from both classes are not well separated and are not easily distinguishable based on the features. the silhouette score obtained for the clustering was 0.017, which is very low.

Dimensionality Reduction and Visualization with UMAP

Like t-SNE, Uniform Manifold Approximation and Projection is again a dimension reduction method available for general non-linear dimension reduction as well as visualisation. It aims to understand the multifarious structure of your data and identify a low dimensional embedding maintaining the fundamental topological structure of that manifold. UMAP's n_neighbors value controls the initial high-dimensional graph's construction's number of nearest neighbours. It affects UMAP's balance between local and global structure: lower values give local details top priority while larger values stress global structure but may sacrifice fine details. The minimum distance between points in the low-dimensional embedding is set by the min_dist variable. Lower values lead

to tighter clustering of points, while larger values result in more dispersed embeddings, focusing on preserving broader topological features.

Despite extensive experimentation with different hyperparameters in UMAP for dimensionality reduction and data visualization, we encountered challenges in obtaining clear and distinct clusters in our data distribution. Despite adjusting parameters aimed at influencing the balance between local and global structure, such as n_neighbors and min_dist, the resulting embeddings did not effectively separate the data into cohesive clusters that would facilitate straightforward visual interpretation or analysis

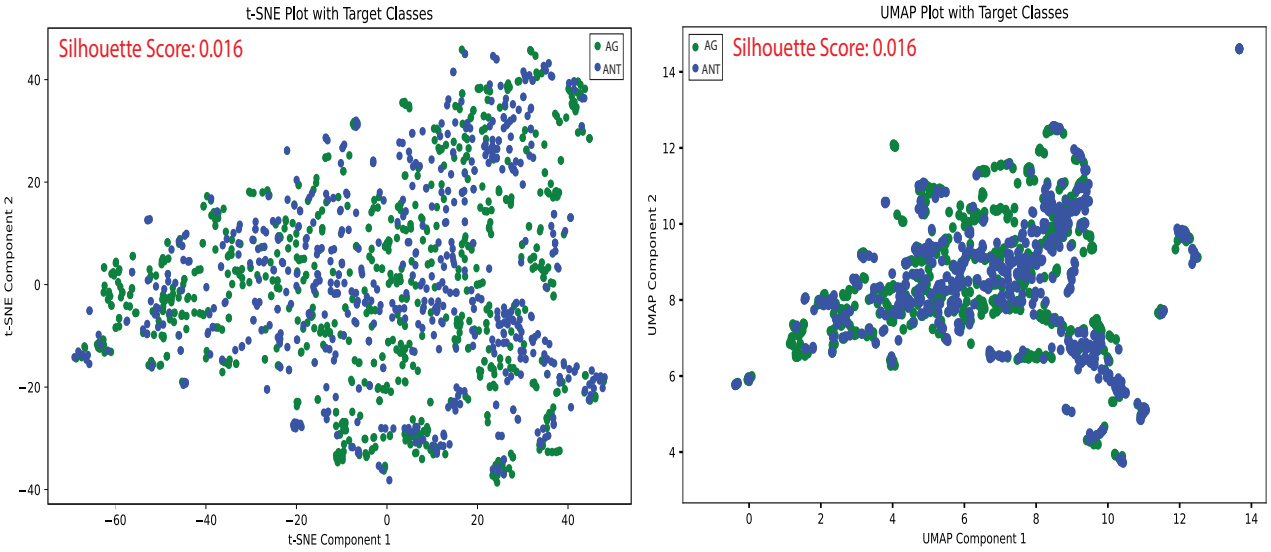

Figure 2.7: Two 2D representations of the entire dataset are available: A. tsne and B. UMAP. In addition to the silhouette score being extremely low, no distinguishable clusters are obtained.

2.3 Machine learning model

according to our study it was required for us to develop a binary class classification model that was capable of precisely classifying the positive class as agonist and and negative class as antagonist.

2.3.1 Model selection

We tested several AutoML pipelines, including FLAML, AutoGluon, AutoKeras, H2O, and LazyPredict, to identify the most effective model among the available options. The evaluations revealed that the best-performing models recommended by all pipelines were gradient boosting models, followed by Random Forest. Subsequently, we compared the evaluation metrics of various machine learning models such as Logistic Regression (LR), Random Forest (RF), Gradient Boosting, Naive Bayes, K-Nearest Neighbors, Support Vector Machines (SVC), Decision Trees, XGBoost (XGB), LightGBM, and CatBoost. Each model was tested with some hyperparameter tuning to optimize performance. It was found that XGBoost consistently provided the best results on our dataset.

In addition, we evaluated these machine learning models on our dataset using different feature selection methods, including Boruta, Variance Threshold, Forward Selection, L1 Regularization, and Backward Selection. Among these combinations, the XGBoost model applied to the dataset with features selected by the Variance Threshold method yielded the best performance. This result highlights the effectiveness of the Variance Threshold method in selecting relevant features that enhance the predictive accuracy of the XGBoost model. Overall, our comprehensive evaluation underscores the superior performance of XGBoost, particularly when paired with the Variance Threshold feature selection method.

After applying the Variance Threshold method to initially reduce the feature set, we further refined the selected features using the ANOVA feature selection technique. By applying the ANOVA method, we aimed to identify the most statistically significant features from the already reduced set. We then employed the XGBoost machine learning model on these ANOVA-selected feature subsets to evaluate their performance. Our goal was to determine which specific combination of features, as filtered by ANOVA, resulted in the highest accuracy for the XGBoost model. This iterative selection process allowed us to pinpoint the optimal set of features that maximize the predictive power of the model.

Feature Selection Method	LR	RF	SVC	XGB
Boruta method	0.64	0.67	0.58	0.66
Variance threshold	0.67	0.70	0.60	0.71
Forward selection	0.66	0.69	0.59	0.64
L1 Regularization	0.62	0.68	0.48	0.71
Backward selection	0.59	0.64	0.48	0.63

Table 2.2: Performance of Different Feature Selection Methods

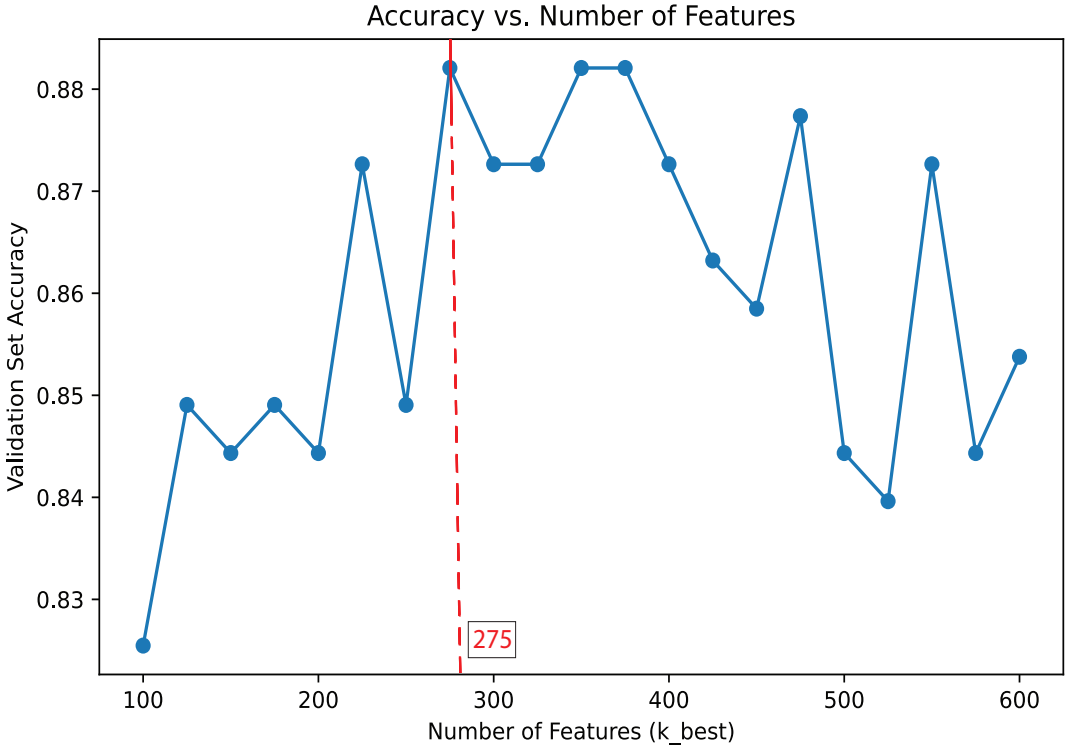

Figure 2.8: Two 2D representations of the entire dataset are available: A. tsne and B. UMAP. In addition to the silhouette score being extremely low, no distinguishable clusters are obtained.

2.3.2 XgBoost Model

Under ensemble learning, XGBoost is a machine learning algorithm. For supervised learning chores like classification and regression, it's hip. XGBoost creates a predictive model iteratively by aggregating, frequently decision tree, predictions of several independent models. The method incrementally adds weak learners to the ensemble, with each new learner focusing on correcting the errors made by the previous ones. It reduces a predetermined loss function during training by use of a gradient descent optimisation method.

XGBoost Algorithm has important characteristics like regularisation methods to prevent overfitting and incorporation of parallel processing for effective computing; it can also manage complicated relationships in data. By use of L1/L2 penalties on the weights and biases of every tree, XGBoost provides regularisation, hence enabling control overfitting. and the weighted quantile sketch approach helps one to manage sparse data sets. While keeping the same computational complexity as previous methods such as stochastic gradient descent, this methodology lets us handle non-zero items in the feature matrix.

- **Hyperparameter tuning**:

The hyperparameters chosen for the XGBoost classifier were tailored to optimize its performance across the dataset. The 'objective' parameter was set to 'binary ' to indicate binary classification tasks, while 'max_depth' constrained each decision tree to a depth of 6 levels, thereby controlling model complexity. A low 'learning_rate' of 0.01 was employed to gradually shrink the contribution of each tree during boosting, enhancing overall model stability. With 'n_estimators' set to 1000, the model iteratively built 1000 trees, refining its predictions with each successive tree. 'Subsample' and 'colsample_bytree', both set to 1, ensured that each tree sampled all

training instances and features, respectively. Regularization was enforced with 'gamma' set to 0.4, 'reg_alpha' and 'reg_lambda' both set to 0.3, and 'min_child_weight' set to 1, influencing tree structure to prevent overfitting. Additionally, 'scale_pos_weight' of 0.46 was used to balance class weights for imbalanced datasets. The model was evaluated using 'logloss' as the evaluation metric, and 'gbtree' was chosen as the boosting type, optimizing the classifier's ability to generalize and make accurate predictions.

Hyperparameter	**Value**
Objective	`binary:logistic`
Max Depth	`6`
Learning Rate	`0.01`
Number of Estimators	`1000`
Subsample	`1`
Column Subsample (By Tree)	`1`
Gamma	`0.4`
Regularization Alpha	`0.3`
Regularization Lambda	`0.3`
Minimum Child Weight	`1`
Scale Positive Weight	`0.46`
Evaluation Metric	`logloss`
Booster	`gbtree`

Table 2.3: XGBoost Hyperparameters

- **Train-test split**:

The dataset was divided into three distinct subsets: 70% for training, 15% for testing, and 15% for validation. This method of splitting the data helps ensure that the performance metrics of the model, such as accuracy, precision, and recall, are both reliable and reflective of its effectiveness in real-world scenarios.

- **Model Evaluation**:

The evaluation of the XGBoost model included rigorous assessment on various datasets, including a blind dataset and one prepared after molecular docking. Performance metrics such as F1 score, recall, Matthews correlation coefficient (MCC), and precision were meticulously computed to assess the model's predictive capabilities comprehensively. Furthermore, the performance of the XGBoost model was compared with that of other models to provide context and insights into its relative effectiveness and strengths. Detailed findings from this comparative analysis are presented in the subsequent results section, offering a comprehensive evaluation of the model's generalization and predictive accuracy across different datasets and relative to alternative modeling approaches.

2.4 Summary

This study employed a comprehensive suite of computational tools and software packages to analyze protein structures and ligand interactions. Dpocket was utilized to extract 55 protein pocket features, including ligand volume, pocket volume, hydrophobicity scores, and mean local hydrophobic density, crucial for understanding binding site characteristics. PaDEL generated 1875 ligand features from SMILES strings, encompassing 1D and 2D descriptors like ALogP and apol, as well as 3D descriptors, and provided 12 types of fingerprints to characterize molecular properties empirically. Naccess calculated atomic accessible surface areas, yielding 5 features such as total all-atoms RSA and main chain RSA, aiding in understanding molecular accessibility and surface interactions. Additionally, Naccess output included cavity-specific features such as DrugScore and Predicted Average pKd from Mol2 files, facilitating the selection of optimal cavity structures. DSSP predicted secondary structure elements and hydrogen bonds, offering insights into protein folding with features such as percentage helix and hydrogen bonding profiles. Using the Biopython package, additional structural features were computed for both the entire PDB structure and the isolated cavity PDB. These features encompassed a wide range of metrics including average hydrogen bond distances, average B-factors, pi-pi stacking interactions, and average distances between aromatic residues. In total, 42 distinct features were derived for each PDB structure and its corresponding cavity. These comprehensive analyses provided a detailed characterization of the structural properties, facilitating a thorough comparative study between the entire protein and its specific cavity environment. Ultimately, a dataset comprising 1407 PDB samples and 275 features was assembled. This dataset was partitioned into a training set (70%) for machine learning model training, and separate test (15%) and validation (15%) sets were reserved. These datasets were further utilized to assess and validate the model's performance against various other datasets.

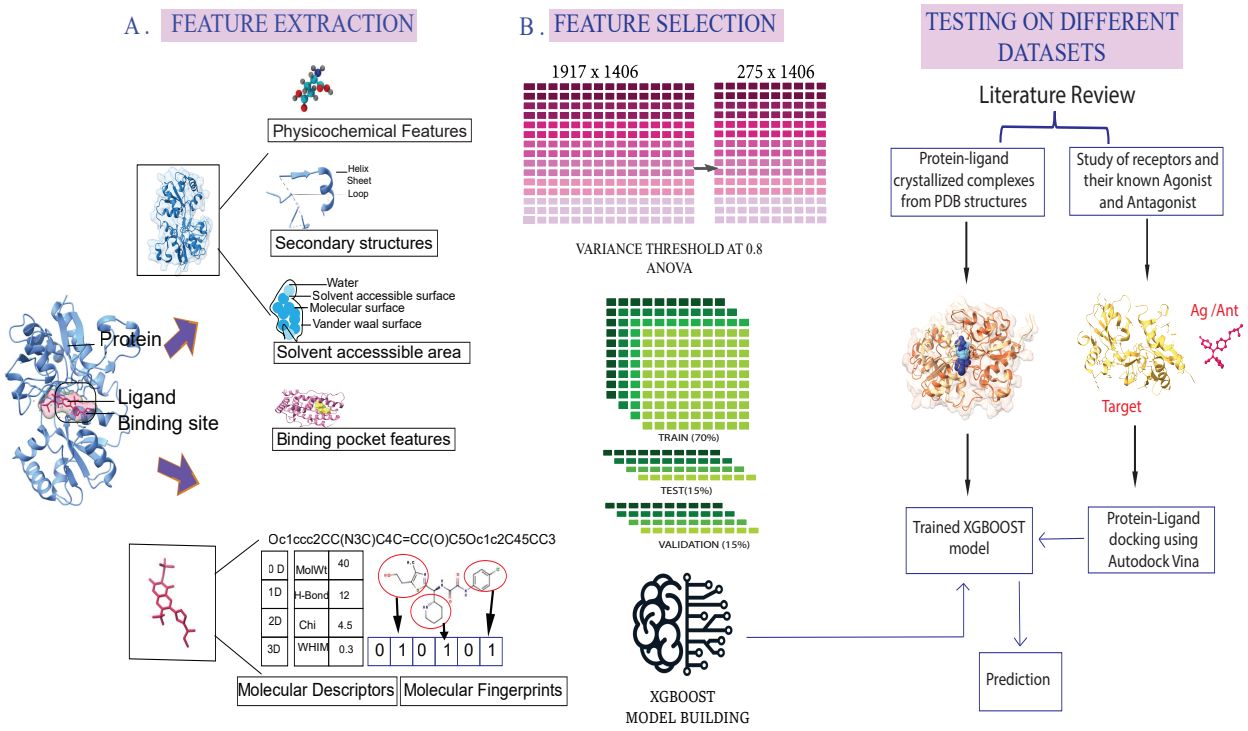

Figure 2.9: A. Types of features extracted using various tools, B. Splitting dataset into Train(70%), Test(15%) and Validation(15%) sets, C. Comparative analysis of top 10 machine learning models

Chapter 3

Results and Discussion

3.1 ML performance on PDB-based Crystal structure dataset

Several measures were used to judge the model's performance, such as F1-score, Youden's Index, Matthews Correlation Coefficient (MCC), and the area under the curve (AUC). To make sure the review was strong, the dataset was split into training, testing, and validation sets. A 10-fold cross-validation method was used, in which the data was randomly split into ten equal parts. In each round, the model was trained on nine of these parts and then checked against the last one. Ten times, this process was done, and each part was used as a confirmation set once. We got a good idea of how well the model worked by averaging the performance metrics over these ten rounds.

The best hyperparameters were chosen based on how well the end model did on the validation set (see Table 2.7). The model was right 86% of the time on the test set. With an F1 score of 0.87, the model does a good job of balancing recall and accuracy, which means it handles false positives and false negatives well.

ROC-AUC scores, which show how well the model can tell the difference between classes, were used to test the model's ability to discriminate. Our model got a great ROC-AUC score of 0.94, which shows that it is very good at classifying things.

The Matthews Correlation Coefficient (MCC) and Youden's Index were also used to evaluate the model. The MCC looks at both true and false positives and negatives. It can be used to judge success on data that isn't balanced. Youden's Index (J) shows how well the test can avoid both fake positives and false negatives. Both tests gave a result of 0.72, which shows that the model is very good at classifying things. Based on the evaluation measures as a whole, our model gets very accurate and reliable classification results.

Confusion Matrix

True positives and true negatives, which appear along the diagonal of the matrix, indicate correctly predicted instances of the positive and negative classes, respectively. False negatives occur when the model incorrectly predicts the negative class, while false positives occur when the model incorrectly predicts the positive class.

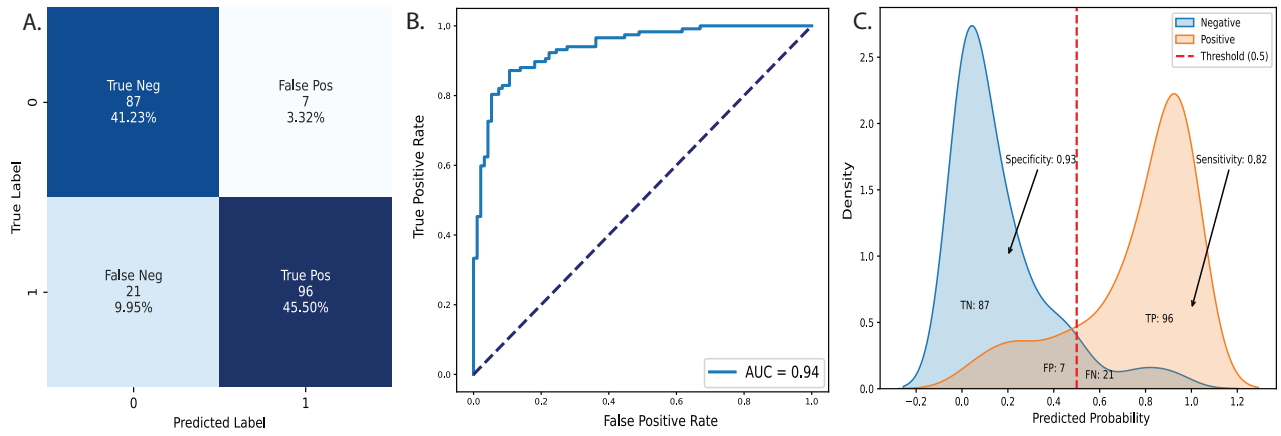

Figure 3.1: A. Confusion Matrix of the XGBoost Model on the Test Set. B.ROC Curve for the XGBoost Model on the Test Set, with an AUC score of 0.94 indicating excellent performance. C.Predicted Probability Density Curve Showing Sensitivity and Specificity for the XGBoost Model on the Test Set.

3.2 ML model evaluation on independent dataset

We thoroughly evaluated our model using a Blind dataset that includes G protein-coupled receptors (GPCRs) and nuclear receptors, both of which are crucial in biological systems and pharmacology. GPCRs are involved in various physiological functions, while nuclear receptors regulate gene expression, making them important targets for therapeutic drugs. Testing our model on these datasets allowed us to gauge its effectiveness in classifying compounds as agonists or antagonists, which is vital for drug discovery and pharmacological research.

The evaluation metrics—Accuracy (83%), F1-score (0.88), AUC-ROC (0.88), Matthews Correlation Coefficient (MCC) (0.71), and Youden's Index (0.71)—demonstrate strong performance across these receptor types. These metrics provide a thorough assessment of the model's ability to differentiate between various pharmacological activities, highlighting its robustness in complex biological scenarios. This validation confirms the model's potential to advance drug development by improving the identification and characterization of bioactive compounds interacting with GPCRs and nuclear receptors.

The confusion matrix indicates good performance, with a high true positive rate (58.05%) and a low false negative rate (8.05%). The ROC curve illustrates the model's ability to accurately classify positive cases, and the predicted probability density graph demonstrates the model's precision in estimating the likelihood of positive cases. These results suggest that the model is well-calibrated and capable of making accurate predictions. Overall, the findings show that the model fits the data well and can be relied upon for making dependable predictions.

The results of the model evaluation are provided below

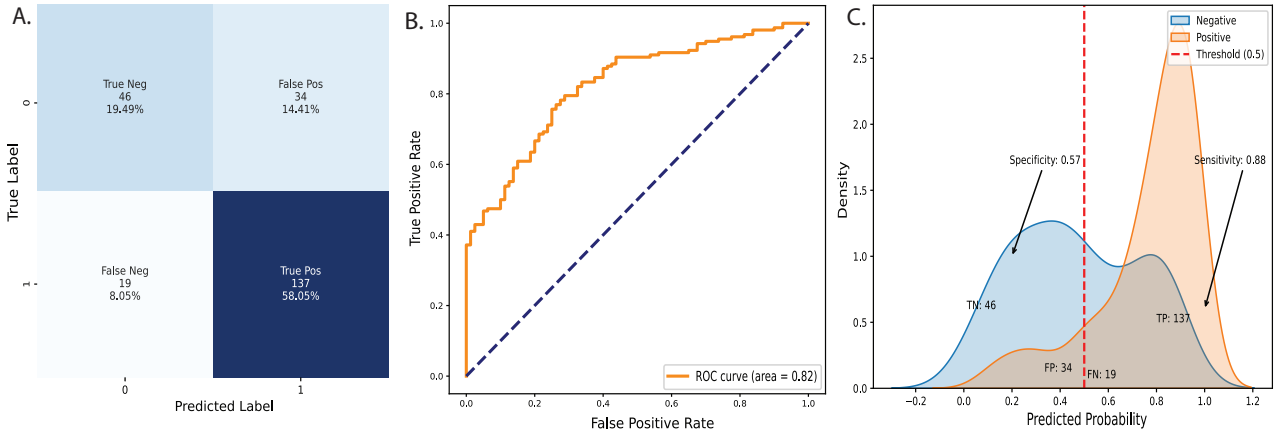

Figure 3.2: A. Confusion Matrix of the XGBoost Model on the Test Set. B.ROC Curve for the XGBoost Model on the Test Set, with an AUC score of 0.94 indicating excellent performance. C.Predicted Probability Density Curve Showing Sensitivity and Specificity for the XGBoost Model on the Test Set.

3.3 ML model evaluation on Literature-based receptor docking dataset

To assess the performance of our model on docked complexes, we analyzed a subset of these complexes. Our initial dataset included 846 docked complexes, featuring a range of receptors such as insulin receptors, Toll-like receptors, muscarinic receptors, and G-protein coupled receptors (GPCRs). Out of this total, predictions were successfully made for 175 complexes. Among these, 84 were labelled as agonists and 91 as antagonists. The accuracy of these predictions was 68.5The predictions for the remaining complexes are still pending. Once these additional predictions are completed, we expect the overall accuracy of the dataset to improve.

3.4 Protein and ligand features found to be important

To identify the crucial features that distinguish between agonist and antagonist complexes, we performed an extensive SHAP (SHapley Additive exPlanations) analysis. By calculating the SHAP values for each feature, we pinpointed those with the highest values, indicating their significant contribution to the model's output. We then selected the top 20 features for further examination. Using these top 20 features, we created a SHAP dot plot to visualize the contributions. This plot serves a dual purpose: it highlights the global feature importance and illustrates the local feature contribution for each instance within the dataset. The scattering in the beeswarm plot provides a comprehensive view of how individual features impact specific predictions, allowing us to see

Metrices	Value
Total docked complexes	175
Total agonist	84
Total antagonist	91
Total predictions	120
Prediction accuracy	0.68
True positive	39
True negative	81
False positive	9
False negative	46
total wrong prediction	55

Table 3.1: docked complexes prediction results

both the overarching trends and the nuances in feature contribution. This combined analysis offers a robust understanding of the factors driving the classification of complexes as either agonists or antagonists.

3.4.1 SHAP summary plot

We calculated the SHAP absolute values for all the features and selected only top 20 features on the basis of highest absolute SHAP values to plot a beeswarm plot.

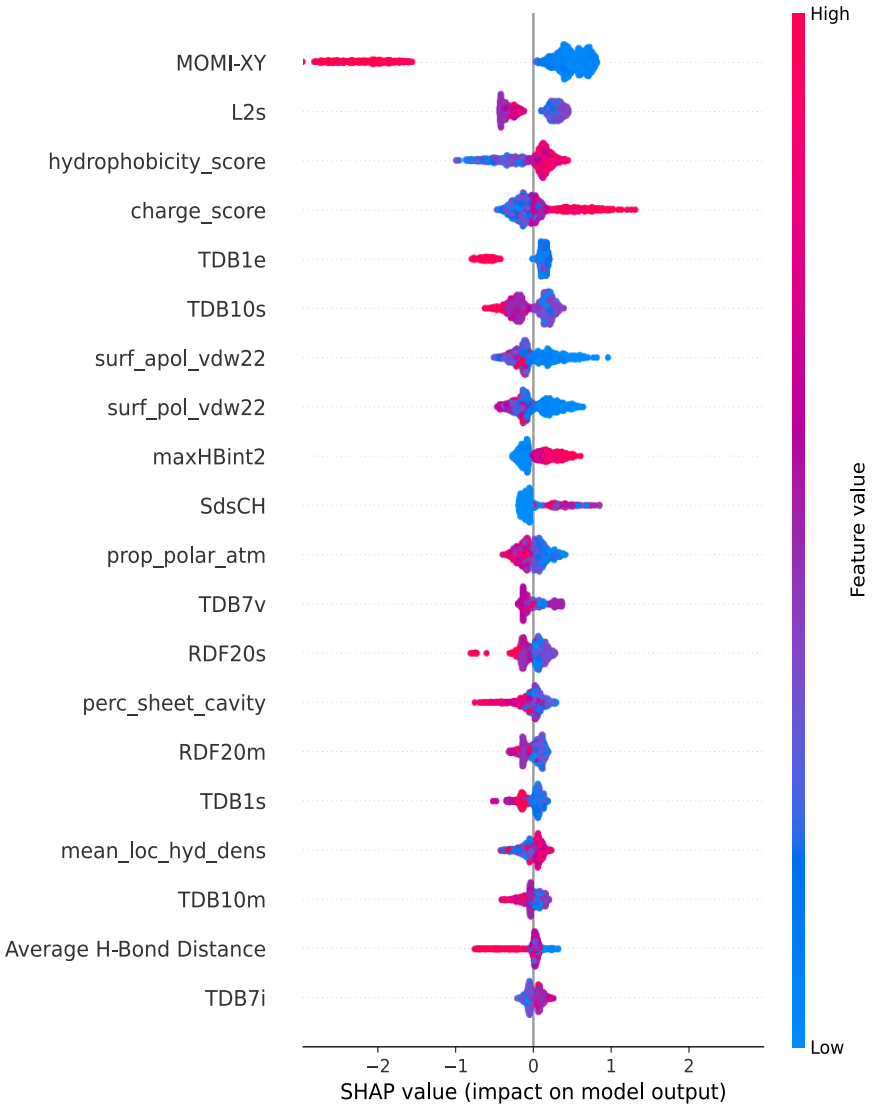

Figure 3.3: Shapely dot plot showing the magnitude and directionality of each contributing features (top 20 according to their shap values) in the model output

The plots demonstrate that both protein and ligand features significantly contribute to determining the response of a protein-ligand complex. Specifically, protein features such as hydrophobicity score, charge score, average hydrogen bond distance, and secondary structure characteristics of the cavity, when considered alongside ligand features, play a crucial role in modulating whether the complex behaves as an agonist or antagonist. The hydrophobicity score of the protein is indicative of the protein's tendency to interact with hydrophobic (water-repelling) regions of the ligand, which can influence the binding affinity and stability of the complex. Similarly, the charge score reflects the electrostatic interactions between the protein and ligand, which are essential for the specificity and strength of the binding. The average hydrogen bond distance provides insights into the quality and strength of hydrogen bonds formed between the protein and ligand. Shorter hydrogen bond distances typically indicate stronger interactions, which can be critical for the functional response of the complex. Moreover, secondary structure features of the protein cavity, such as the presence of alpha helices or beta sheets, contribute to the structural compatibility and proper orientation of the ligand within the binding site. These structural elements can significantly influence the efficacy of ligand binding and, consequently, the agonistic or antagonistic response.When these protein features are considered in conjunction with ligand characteristics, they collectively determine the functional outcome of the interaction. Ligand features, such as molecular size, shape, functional groups, and flexibility, interact synergistically with the protein's attributes to activate (agonist) or inactivate (antagonist) the complex.

These findings establish the hypothesis that both protein and ligand features are integral to the activation or inactivation of a protein-ligand complex. Understanding the interplay between these features can lead to better predictions of complex behavior and inform the design of more effective drugs and therapeutic agents. The plots underscore the necessity of a holistic approach in analyzing protein-ligand interactions, considering the multifaceted contributions of both components to the overall response.

3.4.2 tsne-jointplot

We generated a joint clustered t-SNE plot using the top 20 features identified through SHAP analysis. This visualization technique effectively reveals distinct groupings or clusters within the data. These clusters signify that the selected features collectively contribute to segregating data points based on their inherent characteristics. Such a detailed analysis highlights the pivotal role these features play in delineating nuanced patterns and relationships in protein-ligand interactions. By providing a spatial representation of feature similarities across the dataset, the plot enhances our understanding of how specific protein and ligand attributes influence complex behaviors, thereby supporting more informed biomedical and pharmacological insights.

The plot begins with the application of t-SNE, which projects high-dimensional feature data onto a 2D plane while preserving local similarities between data points. This transformation allows for the visualization of complex relationships and patterns that may not be discernible in the original feature space. Data points are displayed as dots in a scatter plot, where each dot represents an instance from the dataset. Points that are closer together in the plot are more similar in the original high-dimensional space, indicating clusters or groups of data points with similar characteristics. Alongside the scatter plot, density plots are overlaid along the axes of t-SNE component 1 and component 2. These density plots provide a visual representation of the distribution of data points along each t-SNE dimension. They reveal where data points are concentrated and help identify regions of high density, which can correspond to significant clusters or patterns in the data. The t-SNE joint plot with density and clustering provides valuable insights into the structure and relationships within the dataset. It helped us identify clusters of protein-ligand interactions that share similar features, elucidating underlying patterns and potentially revealing new insights into the mechanisms governing complex interactions.

The analysis yielded distinct clusters, underscoring the significance of the selected features in providing valuable information about the target variable. A mutual information score of 0.68, employed to assess the quality of these clusters, reinforces this finding by indicating a substantial amount of shared information between the feature space and the target variable. This score suggests that the features selected are highly relevant and contribute significantly to understanding the characteristics or outcomes related to the target. In the t-SNE plot, specifically along t-SNE component 1, the density distributions exhibit overlapping regions with distinct peaks. This pattern signifies that while there is some overlap in the feature distributions, there are also discernible separations or concentrations of data points. These distinct peaks indicate clusters or groups of data instances that share similar characteristics or belong to distinct categories, contributing to the overall structure revealed by t-SNE.

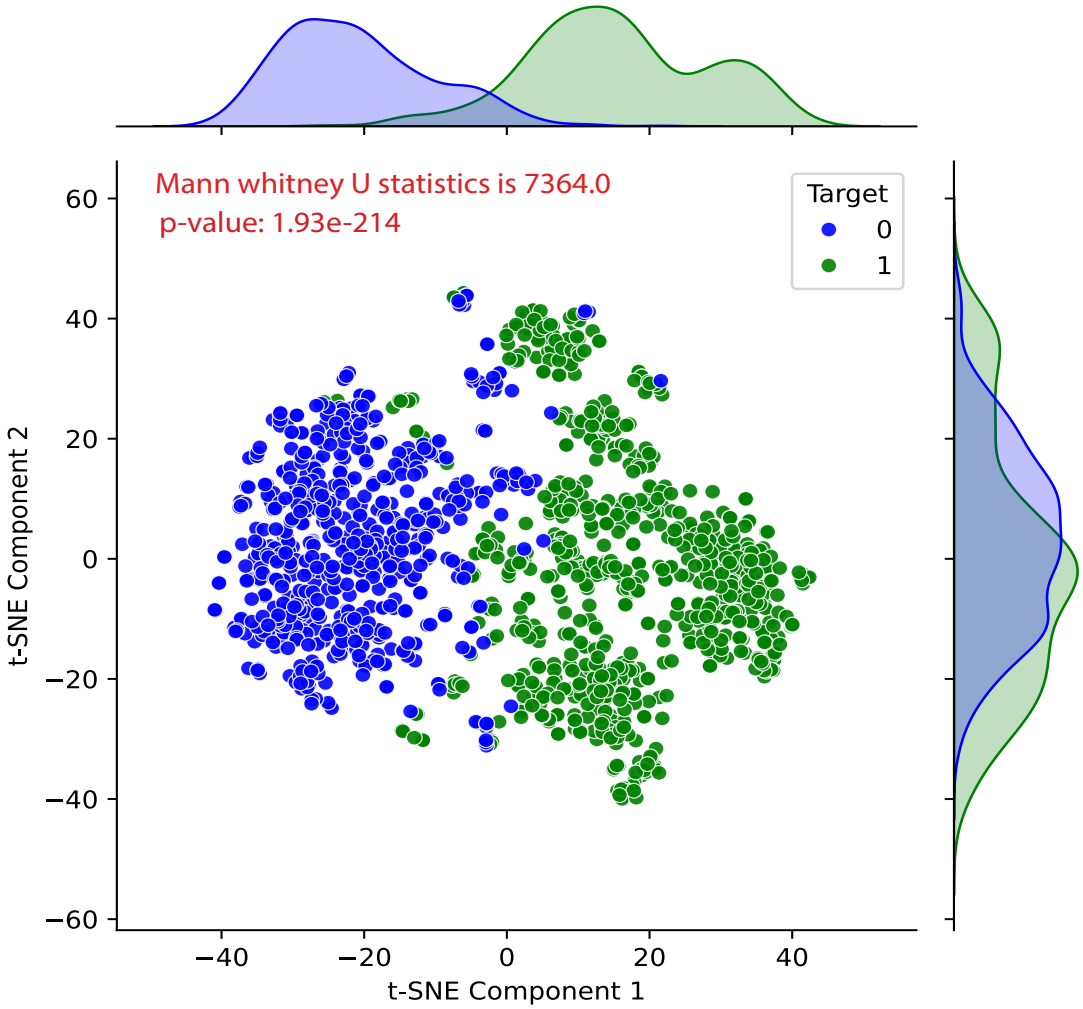

Figure 3.4: Shapely dot plot showing the magnitude and directionality of each contributing features (top 20 according to their shap values) in the model output

The figure presented in the analysis illustrates the non-parametric distributions of both classes along both t-SNE components. t-SNE component 1, the densities of the two classes show some overlap but also exhibit distinct peaks. These distinct peaks indicate potential clusters or subgroups within the dataset that correspond to different classes or categories defined by the target variable. To further evaluate the significance of the separation observed in t-SNE component 1, Mann-Whitney U statistics were performed. The Mann-Whitney U statistics value obtained was 7364, which reflects the rank sum of data points between the two classes in t-SNE component 1. More importantly, the p-value associated with the Mann-Whitney U test was calculated to be 1.93e-214.

Mann-Whitney U Statistics

Let $X = \{X_1, X_2, \ldots, X_m\}$ and $Y = \{Y_1, Y_2, \ldots, Y_n\}$ be two independent samples from populations with unknown distributions.

1. **Ranking the Data:**
Combine the data and rank them in ascending order, assigning ranks R_i to each observation.
2. **Calculating Mann-Whitney U:**
Compute the Mann-Whitney U statistic as follows:

$$U = \sum_{i=1}^{m} R_i - \frac{m(m+1)}{2}$$

$$U = \min(U_X, U_Y)$$

where U_X and U_Y are the sums of ranks for samples X and Y, respectively.

3. **Computing the p-value:**

The p-value is determined based on the distribution of U. For large sample sizes (typically m and n ¿ 20), the p-value can be approximated using the normal distribution:

$$Z = \frac{U - \mu_U}{\sigma_U}$$

where μ_U and σ_U are the mean and standard deviation of U under the null hypothesis.

Alternatively, exact p-values can be calculated using tables or computational methods specific to the Mann-Whitney U distribution.

This extremely small p-value indicates a highly significant difference between the distributions of the two classes along t-SNE component 1. In statistical terms, this means that the observed separation in the t-SNE plot is unlikely to have occurred by chance. Instead, it provides strong evidence that the differences observed in the distribution of data points along t-SNE component 1 are statistically significant and meaningful.

3.4.3 Correlation heatmap

A correlation heatmap was generated for the top 20 features to examine their interrelationships within each class. This visualization allows for a comprehensive exploration of how these features correlate with each other, providing insights into their mutual dependencies and potential multicollinearity. By analyzing correlations separately for each class, the heatmap reveals patterns of association that may vary across different groups or categories defined by the target variable. This analysis is crucial for understanding the underlying structure of the data and for informing subsequent modeling and interpretation in the study of protein-ligand interactions or similar complex systems.

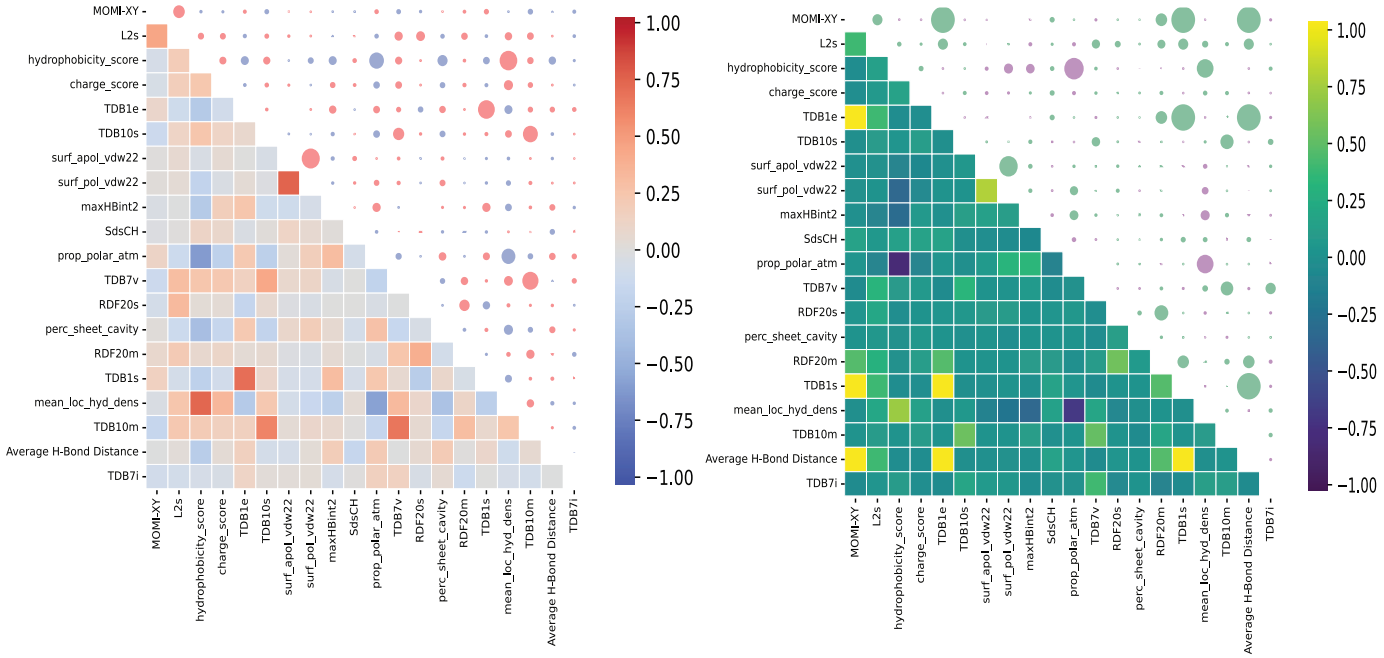

Figure 3.5: Correlation heatmap among features for agonist and D. correlation heatmap among features for antagonist

In the analysis of agonists and antagonists, observed through a heatmap, features such as hydrophobicity score exhibit diverse correlations with other characteristics. In agonists, this score demonstrates both positive associations with attributes like cavity sheet percentage, where increases in one tend to coincide with increases in the other, and negative relationships with mean local hydrophobic density, suggesting that as one feature rises, the other declines. Conversely, in antagonists, various features display mixed regulatory patterns: some exhibit positive correlations, indicating simultaneous increases, while others show negative correlations, where an increase in one feature corresponds to a decrease in another. These intricate relationships, visualized through heatmaps, provide critical insights into how agonists and antagonists interact with biological systems, guiding further investigation into their pharmacological and biological implications.

Chapter 4

Conclusion

The successful development of the "AgAnt" database and web server tool represents a significant advancement in the field of computational biology and drug discovery. This innovative tool is designed to identify and classify agonist and antagonist reactions within protein-ligand complexes, which are pivotal in understanding the functional mechanisms of various biological systems and the design of therapeutic agents. By classifying these interactions as either agonistic or antagonistic, the tool provides valuable insights into how ligands modulate the activity of proteins. This capability is crucial for researchers aiming to decipher complex biochemical processes and for pharmaceutical scientists engaged in the development of drugs that specifically target these interactions. In addition to its core functionality, the "AgAnt" web server is designed with user accessibility in mind, allowing researchers to easily upload and analyze their PDB files through an intuitive interface. The tool's robust analytical algorithms and comprehensive database ensure high accuracy and reliability in its predictions, making it a valuable resource for both academic research and practical applications in drug discovery.

Our research has successfully validated the hypothesis that both protein and ligand features are critical for classifying protein-ligand complex responses as agonists or antagonists. We demonstrated that protein features such as binding site topology and electrostatic properties, combined with ligand attributes like molecular size, shape, and functional groups, significantly influence the interaction outcomes. This dual dependence highlights the importance of considering both sets of features to accurately predict the behavior of protein-ligand complexes. By integrating these aspects, our study provides a more comprehensive framework for understanding molecular interactions and enhances the precision of classifications in drug discovery and development.

When it comes to protein-ligand interactions, the model shows amazing predictive power. Not only does it accurately classify crystallized structures, but it also predicts responses for docked complexes that were not encountered during training with a high degree of accuracy. The model's ability to handle both well-characterized and new protein-ligand interactions is demonstrated by its dual functionality. The model is a useful tool for comprehending and forecasting the behaviour of protein-ligand systems outside the range of the training data since it can generalise to previously encountered docked complexes, indicating that it captures essential aspects of the interaction dynamics.

Bibliography

[1] Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE. Ucsf chimera–a visualization system for exploratory research and analysis., 2023.

[2] Adrian Schreyer, Tom Blundell. Credo: A protein–ligand interaction database for drug discovery. *Chemical Biology and Drug Design*, 2009.

[3] Armstrong, N., Mayer, M., Gouaux, E. Tuning activation of the ampa-sensitive glur2 ion channel by genetic adjustment of agonist-induced conformational changes. *Proceedings of the National Academy of Sciences of the United States of America*, 2003.

[4] Barbara J Pleuvry. Receptors, agonists and antagonists. *Anaesthesia Intensive Care Medicine*, 2004.

[5] Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., Bourne, P. E. . The protein data bank. *Nucleic acids research*, 2000.

[6] Binkowski, T. A., Naghibzadeh, S., Liang, J. Castp: Computed atlas of surface topography of proteins. *Nucleic acids research*, 2003.

[7] Fangfang Wang, Jinyi Xing . Classification of thyroid hormone receptor agonists and antagonists using statistical learning approaches. *Mol Divers*, 2019.

[8] Fu, Y., Zhao, J., Chen, Z. Insights into the molecular mechanisms of protein-ligand interactions by molecular docking and molecular dynamics simulation: A case of oligopeptide binding protein. *Computational and mathematical methods in medicine*, 2018.

[9] Geven Piir, Sulev Sild, Uko Maran. Binary and multi-class classification for androgen receptor agonists, antagonists and binders. *chemosphere*, 2021.

[10] Armstrong J. F. Faccenda E. Southan C. Alexander S. P. H. Davenport A. P. Spedding M. Davies J. A. Harding, S. D. The iuphar/bps guide to pharmacology in 2024. *Nucleic acids research*, 2024.

[11] Heather L. Ciallella, Daniel P. Russo, Lauren M. Aleksunes, Fabian A. Grimm, Hao Zhu,. Predictive modeling of estrogen receptor agonism, antagonism, and binding activities using machine- and deep-learning approaches. *Laboratory Investigation*, 2021.

[12] Hubbard, S. J. and Thornton, J. M. Naccess, 1993. Computer Program, Department of Biochemistry and Molecular Biology, University College London.

[13] Hui Wen Ng1, Wenqian Zhang1, Mao Shu1, Heng Luo2, Weigong Ge1, Roger Perkins1, Weida Tong1,Huixiao Hong. Competitive molecular docking approach for predicting estrogen receptor subtype a agonists and antagonists. *BMC Bioinformatics*, https://doi.org/10.1186/1471-2105-15-S11-S4.

[14] Inhee Choi and Hanjo Kim and Jihoon Jung and Nam, Ky Youb and Yoo, Sung Eun and Kang, Nam Sook and No, Kyoung Tai. Bayesian model for the classification of gpcr agonists and antagonists. *Bulletin of the Korean Chemical Society*, 2010.

[15] Jerome Eberhardt, Diogo Santos-Martins, Andreas F. Tillack, and Stefano Forli. Autodock vina 1.2.0: New docking methods, expanded force field, and python bindings. *J. Chem. Inf. Model.*, 2021.

[16] Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 1983.

[17] Le Guilloux, V., Schmidtke, P. Tuffery, P. Fpocket: An open source platform for ligand pocket detection. *BMC Bioinformatics*, 2009.

[18] Luca Gagliardi, Walter Rocchia. Siteferret: Beyond simple pocket identification in proteins. *J. Chem. Theory Comput.*, 2023.

[19] Lucy J Colwell. Statistical and machine learning approaches to predicting protein–ligand interactions. *Current Opinion in Structural Biology*, 2018.

[20] Matsuzaka Y., Uesawa Y. Prediction model with high-performance constitutive androstane receptor (car) using deepsnap-deep learning approach from the tox21 10k compound library. *Int. J. Mol. Sci.*, 2019.

[21] Matsuzaka Y., Uesawa Y. Deepsnap-deep learning approach predicts progesterone receptor antagonist activity with high performance. *Frontiers in bioengineering and biotechnology*, 2020.

[22] Matsuzaka, Yasunari, and Yoshihiro Uesawa. A deep learning-based quantitative structure–activity relationship system construct prediction model of agonist and antagonist with high performance. *International Journal of Molecular Sciences*, 2020.

[23] Niraj Verma,Xingming Qu,Francesco Trozzi,Mohamed Elsaied,Nischal Karki,Yunwen Tao,Brian Zoltowski,Eric C. Larson, Elfi Kraka. Ssnet: A deep learning approach for protein-ligand interaction prediction. *International Journal of Molecular Sciences*, 2021.

[24] O. Trott, A. J. Olson. Autodock vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization and multithreading. *Journal of Computational Chemistry.*, 2010.

[25] O'Boyle, N.M., Banck, M., James, C.A. . Open babel: An open chemical toolbox. *J Cheminform*, 2011.

[26] Oh, J., Ceong, Ht., Na, D. . A machine learning model for classifying g-protein-coupled receptors as agonists or antagonists. *BMC Bioinformatics*, 2022.

[27] R. Andersson, Claes; G. Gustafsson, Mats; Strombergsson, Helena. Quantitative chemogenomics: Machine-learning models of protein-ligand interaction. *Current Topics in Medicinal Chemistry*, 2011.

[28] S. Bristow, V. Singh, J. Ballantyne. Opioids. *Encyclopedia of the Neurological Sciences*, 2014.

[29] Tian Cai, Kyra Alyssa Abbu2, Yang Liu, Lei Xie1. Deepreal: a deep learning powered multi-scale modeling framework for predicting out-of-distribution ligand-induced gpcr activity. *Bioinformatics,*, 2022.

[30] Tingting Sun, Yuting Chen, Yuhao Wen, Zefeng Zhu, Minghui Li. Prempli: a machine learning model for predicting the effects of missense mutations on protein-ligand interactions. *Communications Biology*, 2021.

[31] Roberto Todeschini and Viviana Consonni. *Handbook of molecular descriptors.* John Wiley & Sons, 2008.

[32] Roberto Todeschini and Viviana Consonni. *Molecular descriptors for chemoinformatics: volume I: alphabetical listing/volume II: appendices, references.* John Wiley & Sons, 2009.

[33] Uesawa Y. Quantitative structure-activity relationship analysis using deep learning based on a novel molecular image input technique. *Bioorg Med Chem Lett*, 2018.

[34] Weir, C. J. Ion channels, receptors, agonists and antagonists. *Anaesthesia Intensive Care Medicine*, 2013.

[35] Wouter G Touw, Coos Baakman, Jon Black, Tim AH te Beek, E Krieger, Robbie P Joosten, Gert Vriend. A series of pdb related databases for everyday needs. *Nucleic Acids Research*, 2015.

[36] Xue-lian Zhu, Hai-yan Cai, Zhi-jian Xu, Yong Wang, He-yao Wang, Ao Zhang, Wei-liang Zhu. Classification of 5-ht1a receptor agonists and antagonists using ga-svm method. *Acta Pharmacol Sin*, 2011.

[37] CHUN WEI YAP. Padel-descriptor: An open source software to calculate molecular descriptors and fingerprints. *Journal of Computational Chemistry*, 2010.

[38] Yaxia Yuan, Jianfeng Pei, and Luhua Lai. Ligbuilder 2: A practical de novo drug design approach. *J. Chem. Inf. Model.*, 2011.

[39] Yoonji Lee, Xiyan Hou, Jin Hee Lee, Akshata Nayak, Varughese Alexander, Pankaz K. Sharma, Hyerim Chang, Khai Phan, Zhan-Guo Gao, Kenneth A. Jacobson, Sun Choi Lak Shin Jeong. Subtle chemical changes cross the boundary between agonist and antagonist: New a3 adenosine receptor homology models and structural network analysis can predict this boundary. *J. Med. Chem.*, 2021.

[40] Youjun Xu, Shiwei Wang, Qiwan Hu, Shuaishi Gao, Xiaomin Ma, Weilin Zhang, Yihang Shen, Fangjin Chen, Luhua Lai, Jianfeng Pei. Cavityplus: a web server for protein cavity detection with pharmacophore modelling, allosteric site identification and covalent ligand binding ability prediction. *Nucleic acids research*, 2018.