NATIONAL TECHNICAL UNIVERSITY OF ATHENS
SCHOOL OF ELECTRICAL AND COMPUTER ENGINEERING
SCHOOL OF MECHANICAL ENGINEERING

INTERDISCIPLINARY POSTGRADUATE PROGRAMME
"Translational Engineering in Health and Medicine"

Machine-Learning-based prediction of human SERT protein ligand affinity using molecular docking and interaction analysis

Postgraduate Diploma Thesis

DIMITRIOS K. PAPANAGNOU (03500045)

Supervisor:
PROFESSOR GEORGE MATSOPOULOS,
National Technical University of Athens

Co-Supervisor:
YIANNIS MAKRIS,
Scientific Associate in School of Electrical and Computer Engineering National Technical University of Athens

Athens, June 2025

NATIONAL TECHNICAL UNIVERSITY OF ATHENS
SCHOOL OF ELECTRICAL AND COMPUTER ENGINEERING
SCHOOL OF MECHANICAL ENGINEERING

INTERDISCIPLINARY POSTGRADUATE PROGRAMME
"Translational Engineering in Health and Medicine"

Machine-Learning-based prediction of human SERT protein ligand affinity using molecular docking and interaction analysis

Postgraduate Diploma Thesis

DIMITRIOS K. PAPANAGNOU (03500045)

The postgraduate diploma thesis has been approved by the examination committee on 25 June 2025

1st member	*2nd member*	*3rd member*
Prof. George Matsopoulos *NTUA*	*Prof. P. Tsanakas* *NTUA*	*As. Prof. Ch. Manopoulos* *NTUA*

Athens, June 2025

……

Dimitrios K. Papanagnou
Graduate of the Interdisciplinary Postgraduate Programme,
"Translational Engineering in Health and Medicine",
Master of Science,
School of Electrical and Computer Engineering,
National Technical University of Athens

ABSTRACT

The following thesis presents a multi-disciplinary computational approach for classifying potential inhibitors of the human serotonin transporter (SERT) into three distinct categories: strong binders, moderate binders, and non-binders. SERT is the primary target for many antidepressants. The pipeline integrates several steps including molecular docking, molecular descriptor analysis, residue-level interaction profiling and creation of a supervised machine learning model in order to extract and clarify ligand-SERT interactions. A total of 74 compounds with and without known pharmacological action were studied that belong mainly to wide antidepressant categories, such as SSRIs, SNRIs, TCAs and other unrelated categories which are considered non-binders. The categorization of these ligands into the 3 classes was assigned based on the available inhibition constant values (Ki) with human SERT receptor from authorized pharmacological sources. Initially, molecular docking was employed with the aid of "AutoDock Vina" and "Chimera" software to generate the top ten binding poses for each ligand. The validity of the docking process and protocol was assessed by comparing the predicted binding conformation of the known SSRI drug *"Paroxetine"* with the baseline crystallographic structure from Protein Data Bank (PDB: 5I6X), resulting in a nearly perfect alignment. A custom Python script was applied to select the top five out of ten poses by ranking them based on their binding affinity and root-mean-square deviation values (RMSD). Extensive molecular and residue details were obtained using "BIOVIA Discovery Studio", including *"Surface_Area"*, geometric angles and distance-based features. Statistical analyses were conducted to examine the correlations of features with the target variable, which is the class label and to detect potential multicollinearity among them. Notably, for strong binders, hydrophobic residues, such as *"ALA_173"*, *"ILE_172"*, and *"PHE_341"* were found to be critical. Apart from these, distinct distributions of *"Polar_Surface_Area"* and other angular features like *"ANGLE_HAY"* and *"GAMMA"* were observed. Several machine learning algorithms were trained including Random Forest, XGBoost, LightGBM, Logistic Regression, SVM and Voting Classifier. Nested cross-validation technique was integrated to minimize the risk of overfitting, however performance was moderate, due to the overlapping descriptor distributions between moderate and adjacent classes. Tree-based models outperformed, while at the same time facilitated interpretability of model decisions through SHAP summary and partial dependence plots. These plots highlighted the most predictive and important features across *"STRONG BINDING"* and *"MODERATE BINDING"* classes and confirmed that moderate binders confused the model. Despite the controversial success of the models used, the assumptions and limitations under which the present thesis was conducted, are outlined. Most decisive of them is the limited sample size, the static docking simulations and the custom script that selected the five best poses. Nevertheless, the study suggests for future work to incorporate molecular dynamics simulations from a wider range, include more targeted receptors for docking that are responsible for antidepressant activity, such as NET and DAT and molecular fingerprints that capture atomic level interactions. By this method, the classification accuracy and validity of results will be indisputable. This thesis lays a foundation for an innovative plan for detecting potential antidepressants drugs with the aid of several computational tools, but it requires a lot of optimizations to be considered reliable.

Keywords: Serotonin Transporter (SERT), SERT Inhibitors, Antidepressants, Molecular Docking, Machine Learning, Supervised Classification, Residue-Level Interaction Profiling, Molecular Descriptors, AutoDock Vina, Chimera, BIOVIA Discovery Studio, SHAP Values, Partial Dependence

Plots (PDPs), Nested Cross-Validation, Binding Affinity, Root-Mean-Square Deviation (RMSD), Surface Area, Polar Surface Area, ANGLE HAY, GAMMA, Hydrophobic Interactions, ALA_173, ILE_172, PHE_341, Random Forest, XGBoost, LightGBM, Logistic Regression, Support Vector Machine (SVM), Voting Classifier, Static Docking Limitations, Molecular Dynamics Simulations, Molecular Fingerprints, Multi-target Screening, NET, DAT.

ACKNOWLEDGEMENTS

"It always seems impossible until it's done." – Nelson Mandela

As Nelson Mandela said, this thesis marks the end of a very demanding but deeply meaningful academic journey, initially unreachable but promising for knowledge and personal growth. The path to its completion was long but with daily minor goals, it was achievable.

I would like to express my respected greetings to my Supervisor, Professor George Matsopoulos and Co-Supervisor Yiannis Makris, for their scientific guidance, trust and support in order to complete this thesis successfully.

I am deeply grateful to my family, who has stood by my side over the years and have been a source of unwavering support. Their faith in me and my dreams have been very helpful across this journey.

Lastly, I would like to thank my friend Ioanna, who is always by my side, supporting every new effort and choice I make with patience and care. It would be unfair not to also thank my friends who have been in my life for all over the years and have made positive impact on my personality and moral values.

CONTENTS

Chapter 1: Introduction

1.1 Major depressive disorder

One of the most common psychiatric disorders worldwide, is major depressive disorder (MDD). Based on the World Health Organization (2023) nearly 280 million people suffer from depression including adults, children and particularly women. Persistent sadness, pessimism, feeling of dissatisfaction, permanent fatigue, cognitive dysfunction, disrupted sleep and in severe cases suicidal thoughts are among the list of symptoms of depression. The healthcare professionals mention that the symptoms should persist for more than two weeks to be correlated to depression. Despite technological advancements, the efficacy of antidepressants remains debatable, with more than 75% of patients globally do not have access to treatment, especially in low-income countries. The most recognized theory for the cause of depression is the monoamine hypothesis. This theory claims that major depressive disorder is triggered by a dysregulation of specific neurotransmitters in the human brain, like serotonin (5-HT), norepinephrine (NE) and dopamine (DA) (Delgado, 2000). Based on this theory, scientists developed the selective serotonin reuptake inhibitors (SSRIs) in the mid-1980s with fluoxetine being the first that was launched, serotonin-norepinephrine reuptake inhibitors (SNRIs), tricyclic antidepressants (TCAs) and more recently the serotonin modulators (Hillhouse & Porter, 2015).

1.1.1 Selective Serotonin Reuptake Inhibitors (SSRIs)

The majority of prescribed antidepressant drugs are SSRIs. They operate by blocking the serotonin transporter in the presynaptic nerve, which raises serotonin levels in the synaptic space and therefore allows serotonergic transmission through activation of postsynaptic 5-HT receptors that are responsible for handling mood and anxiety (Chu & Wadhwa, 2023). Common prescribed SSRIs include fluvoxamine, sertraline, citalopram, paroxetine, and fluoxetine (MedlinePlus, 2023). Due to their high selectivity to SERT, they tend to have less side effects than other classes of antidepressants. Based on Mayo Clinic, SSRIs potential side effects include upset stomach, headaches, dry mouth, anxiety, sexual dysfunctions and many more.

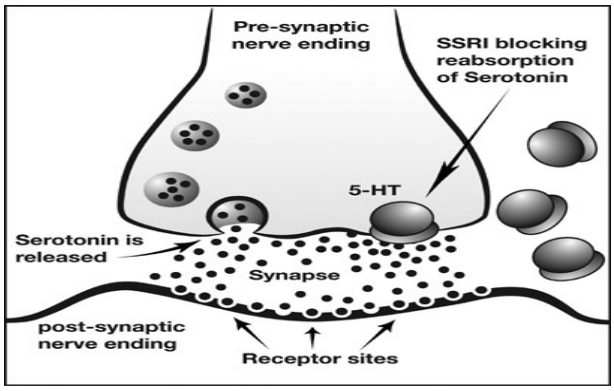


Image 1: Serotonin release from presynaptic nerve to synaptic cleft with SSRIs blockers in SERT transporter (Physiopedia, (n.d.)).

1.1.2 Serotonin Norepinephrine Reuptake Inhibitors (SNRIs)

SNRIs are newer drugs than SSRIs and as their name reveals, they implement a dual mechanism, by which they inhibit both the serotonin and the norepinephrine transporters. This leads to higher concentrations of serotonin and norepinephrine in the synaptic cleft (Randy A Sansone , Lori A Sansone, 2014). Usually, this class of drugs such as Venlafaxine and Milnacipran are prescribed as first class medication like SSRIs, since both have similar side effects, but fewer in number than other categories like TCAs or MAOs based on the Cleveland Clinic.

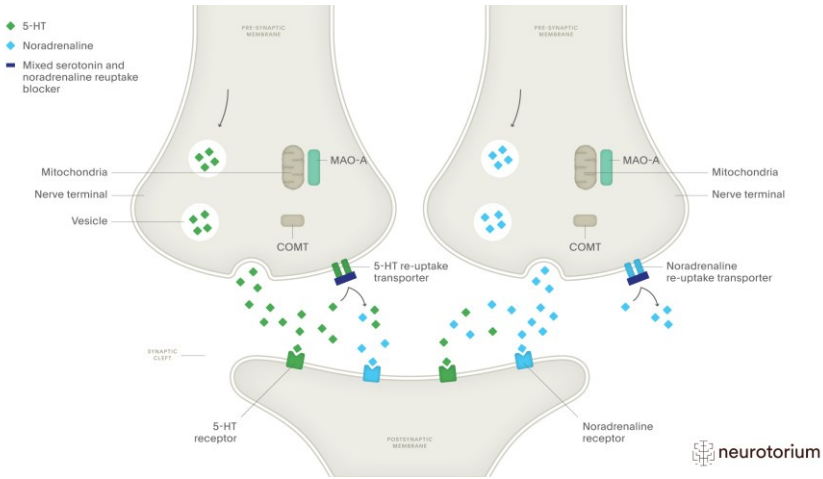


Image 2: Serotonin and norepinephrine release from presynaptic nerve to synaptic cleft with SNRIs blockers in SERT and NET transporters (Neurotorium, (n.d.)).

1.1.3 Tricyclic Antidepressants (TCAs)

Tricyclic antidepressants are the oldest class of antidepressants that were discovered in late 1950s and are still used nowadays as a secondary treatment of depression. Their name arises from the presence of three organic rings in their chemical core structure. Similarly to SNRIs, they inhibit the reuptake of both serotonin and norepinephrine by blocking their respective transporters. However, they are considered as alternative treatment for MDD, because it has been proved that they antagonize various receptors like histamine (H1), muscarinic (M1), and adrenergic (α1) receptors leading to severe side effects, such as sedation, constipation, orthostatic hypertension and cardiovascular conditions like arrhythmias (Moraczewski et.al, 2023). Common TCAs include imipramine and nortriptyline. Additionally, there is another class called tetracyclic antidepressants (TECAs) that operate in a similar manner, but they have 4 aromatic rings in their structure instead of three.

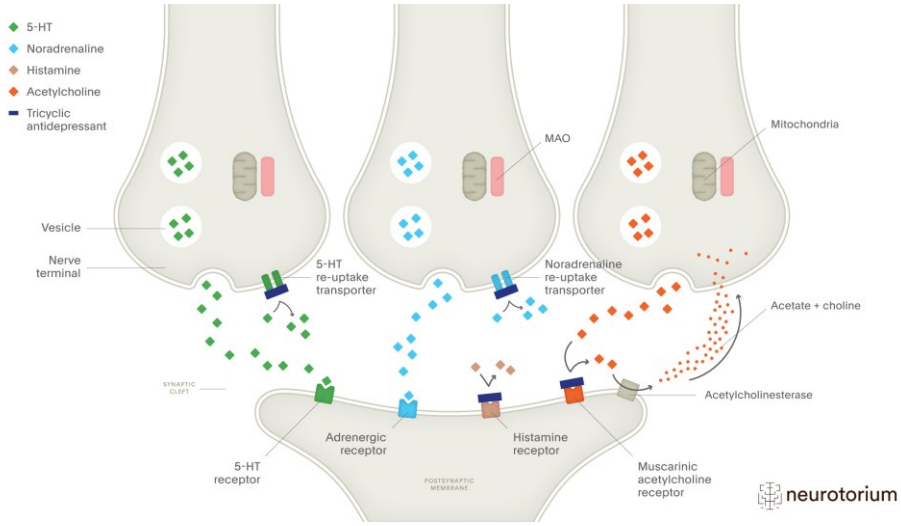


Image 3: TCAs inhibition of SERT and NET and simultaneous antagonizing of adrenergic, muscarinic and histamine receptors (Neurotorium. (n.d.)).

1.1.4 Monoamine Oxidase Inhibitors (MAOIs)

Monoamine Oxidase Inhibitors (MAOIs) were introduced in 1950s for the treatment of depression and are still used today as a third option for handling depressive symptoms. Their mechanism is more complicated than other classes. Inside the mitochondria of the presynaptic neurons, there are two types of enzymes, monoamine oxidase enzymes MAO-A and MAO-B. Their role is the degradation of monoamines like serotonin, norepinephrine, dopamine and tyramine that do not enter the synaptic vesicles in order to reach a balanced concentration of these neurotransmitters in the nerve. MAOIs work by inhibiting these enzymes which therefore increase the levels of neurotransmitters availability for entering the vesicles and this leads to a higher concentration release in the synaptic cleft to regulate

the mood, attention and motivation of a patient. MAOIs can be reversible or irreversible. For instance, phenelzine and tranylcypromine inhibit both MAO-A and MAO-B irreversibly, while moclobemide is a reversible MAO-A inhibitor. Selegiline, for example, inhibits solely MAO-B which is useful for sufferers of Parkinson's disease. As for the side effects, MAOIs have been accused of provoking diarrhea, constipation, drowsiness, insomnia. Howerer, there is a proven danger using these drugs, because in combination with other antidepressants like SSRIs, they are likely to cause serotonin syndrome which is a threatening condition for life. Last but not least, a strict dietary plan should be adjusted to patients when using MAOIs, because these drugs block the breakdown of tyramine from tyramine-rich foods like cheese and beer and as a result blood pressure elevates and this in rare cases leads to cerebral haemorrhage (Laban & Saadabadi, 2023).

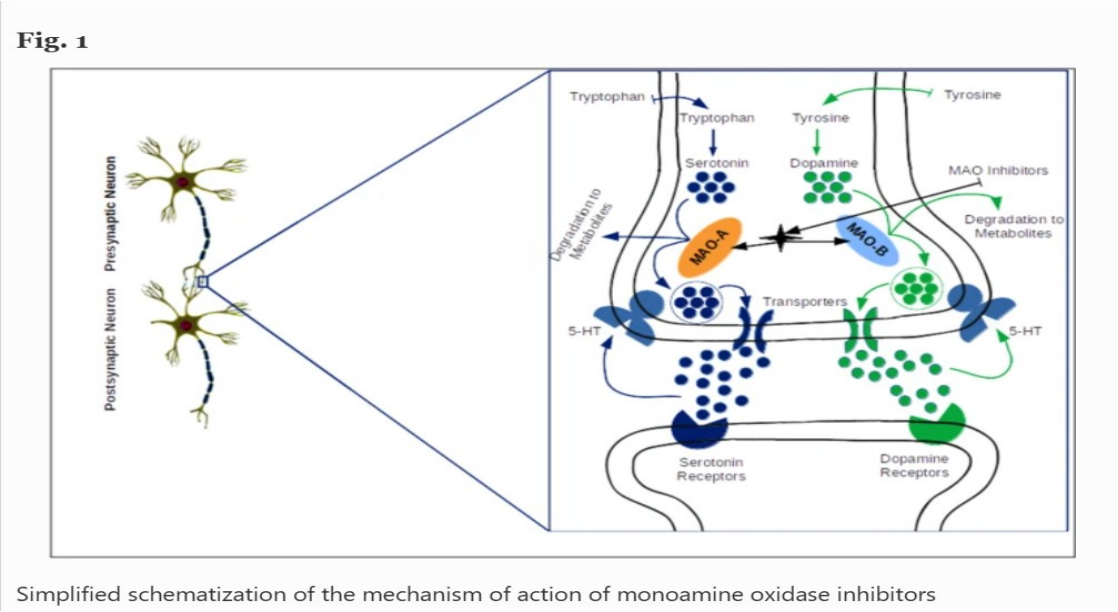


Image 4: Mechanism action of MAOIs inhibition of MAO-A and MAO-B enzymes (A. Reyes-Chaparro et al., 2023).

Drug Class	Mechanism
SSRIS	BLOCK SERT AND HAVE FEW SIDE EFFECTS
SNRIS	BLOCK SERT AND NET TRANSPORTER WITH FEW TO MODERATE SIDE EFFECTS
TCAS	BLOCK SERT AND NET TRANSPORTER PLUS OTHER RECEPTORS WITH SEVERAL SIDE EFFECTS
MAOIS	INHIBIT MAO-A/B WITH SEVERE SIDE

	EFFECTS AND DIETARY RESTRICTIONS

Table 1: Summary of Antidepressants.

1.1.5 Complementary theories of major depressive disorder

Although all the treatment regimens were invented based on the monoamine theory of depression, scientists claim that it is a multifactorial condition for humans and many other theories try to shed light on "solving" the puzzle of what are the factors that are connected to depression. Supporting evidence for this is the delay onset for antidepressants to work, which ranges between 4-8 weeks based on Cleveland Clinic. In addition, only half of the patients taking first and second line medications show reduction of depressive symptoms, while only 30 percent manage complete recovery (Vedrines et al., 2022). Another possible cause of depression is correlated to synaptic dysfunction. Chronic stress and anxiety influence brain-derived neurotrophic factor (BDNF) which is a molecule that regulates synaptic plasticity in regions of the brain responsible for learning and memory processes (Miranda et al., 2019). With this way, there is a loss in dendritic spines that affect the neurons in regions related to mood disorders like prefrontal cortex and hippocampus (Duman , Aghajanian, 2012). An alternative theory that has been proposed as major cause for depression is the glutamatergic system, which is the main excitatory neurotransmitter system. This theory claims that exposure to stressful and anxious environments causes an imbalance in glutamate concentration in prefrontal and limbic regions of the brain leading to synaptic loss and maladaptive dendritic spines. NMDA (N-methyl-D-aspartate) and AMPA (α-amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid) receptors are the drug-targets with antidepressant effect that support this hypothesis (Sanacora et al., 2012). Other group of scientists consider that major depressive disorder is induced by chronic inflammation in the human body. Elevated levels of pro-inflammatory markers, such as Interleukin-6 (IL-6)**,** Tumor Necrosis Factor TNF-α and C-Reactive Protein (CRP) play a substantial role in mood disorders by reducing BDNF and stimulating indoleamine 2,3-dioxygenase (IDO)**,** which is an enzyme that removes tryptophan from the body and produces neurotoxic metabolites like quinolinic acid (Miller, Raison, 2016). Finally, changes in gut microbiota that occurs after chronic stress alter the neurotransmission. As a result, there is a destabilization of mood and overall cognitive behaviour. This neural signaling is motivated by the microbiota–gut–brain axis, which enables bidirectional communication between the central and the enteric nervous system (Hwei-Ee Tan, 2023).

1.2 Central Nervous System and the Structure-Function of the Serotonin Transporter (SERT)

The anatomy of the human nervous system is divided into two main parts, the Central Nervous System (CNS) and the Peripheral Nervous System (PNS). The autonomic nervous system which is a component of PNS seems to be highly correlated to emotional states as it regulates the involuntary physiological processes like emotions and arousal (Hall et al., 2023). Brain and spinal cord form the CNS by receiving and processing the signals from nerves and translate them into cognitive, motor and emotional outputs. The PNS consists of a network of nerves, which branch out of the spinal cord and transport the signals from brain and spinal cord to the organs and entire body (Cleveland Clinic, 2023). Over 100 billion neurons form the nervous system, which are the primary functional unit. Each neuron contains 3 parts: the cell body or soma, the dendrites and the axon. The cell body, inside of which nucleus exists, is vital for neuron's life. Dendrites capture signals from surrounding neurons and pass them through cell body and axon to the adjacent nerve cell. The axon is protected from a fatty molecule called "myelin shealth" which regulates the speed and distance that the signal will travel at the same time (National Institute of Neurological Disorders and Stroke, 2018).

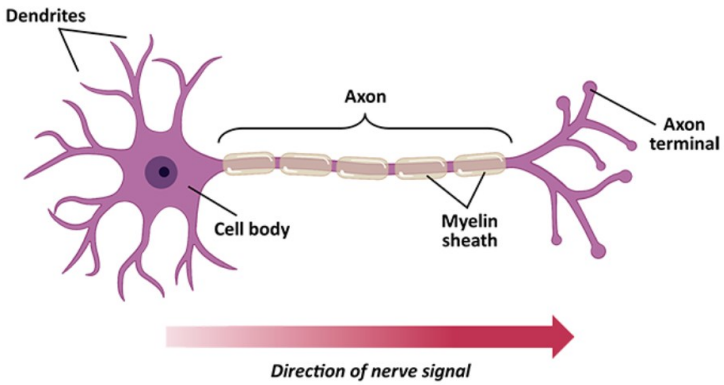


Image 5: Structure of Neuron (National Institute of Neurological Disorders and Stroke, 2023).

Typically, a neuron has a potential of -70mV in resting state, which is called resting membrane potential. This means that the internal part of the cell has a negative electrical charge compared to the external. This electrical polarization is stabilized through controlled flow of Na^+ and K^+ in the membrane cell. When a neuron is stimulated by the adjacent neurons, hypopolarization begins. Above the threshold potential which ranges from -50 to -55 mV, depolarization starts and an action potential occurs that opens voltage-gated sodium channels and allows a large influx of sodium ions to enter the membrane (Vasković, 2023; Society for Neuroscience, 2018).

In a serotonergic neuron this action potential will travel down the axon and as a result the voltage-gated calcium channels will open, allowing Ca^{2+} ions to enter. This influx of Ca^{2+} provokes fusion of presynaptic vesicles, that trigger the release of serotonin in the synaptic cleft and the adjacent neurons through their dendrites will receive this signal message. Depolarization stops when the inside of the cell becomes very positive reaching the known overshoot phase. After this point, sodium influx stops and voltage-gated potassium channels open leading to efflux of potassium. Cell membrane potential becomes again more negative in this repolarization step which purpose is to restore the resting membrane potential (Vasković, 2023; Society for Neuroscience, 2018).

Apart from neurons, proteins also play a substantial role in neural function by mediating signaling and enzymatic activity. Proteins, primarily, consist of huge linear chains, called amino acids that fold into a specific three-dimensional structure that is vital for promoting their relevant biological activity. In addition, they possess a secondary structure, most common of which are the α-helices and the β-sheets. Both are retained through hydrogen bonds. More specifically, in α-helices these bonds are formed every fourth amino acid, while in β-sheets the known "pleats" are formed on the backbone of the polypeptide chain. The unique three-dimensional shape of each protein constitutes the tertiary structure. Finally, there is the quaternary structure which is described as the arrangement of multiple polypeptide subunits in order to interact with each other. All structural levels of a protein are depicted in Image 6 below and they determine how proteins interact with other molecules, like ligands. Every protein has its own unique configuration and amino acids chains and when it is exposed in specific temperatures, pressure, pH or other chemical reactions, it denatures, meaning that its shape alters. This modification may be irreversible and therefore ceases to perform its biological function (Lodish et al., 2016 ; courses.lumenlearning.com, (n.d.)).

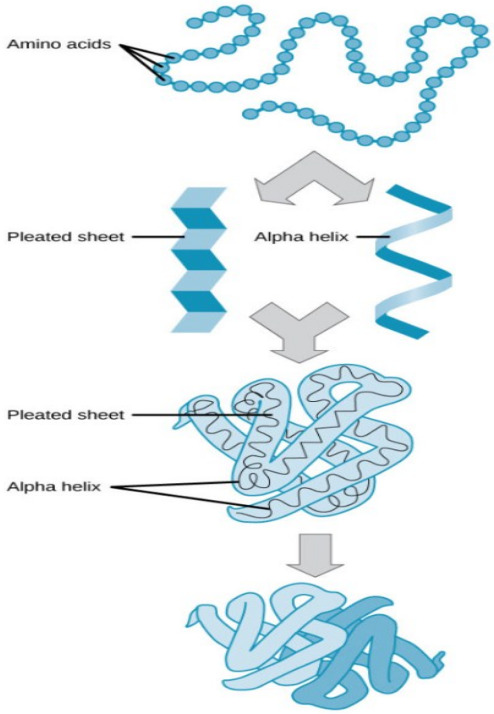


Image 6: Structure of Proteins (courses.lumenlearning.com, (n.d.).

A key pharmacological target for many antidepressants is the serotonin transporter (SERT) which consists of 12 transmembrane a-helices (TM1-TM12). Their role is to separate the intracellular from the extracellular environment by composing a binding cavity. SERT is responsible for the reuptake of serotonin from the synaptic cleft into the presynaptic neuron. Upon ion and substrate binding, some conformational changes occur which enable the transport of ions like Na^+ and CL^- and serotonin across the membrane. Critical regions for the inhibition of SSRIs are TM1, TM3, TM6 and TM8 which compose the central binding site of the protein. However, scientists have discovered that some ligands, such as escitalopram bind also on other secondary binding sites called allosteric sites, which affect the binding dynamics in the central site and stabilize the ligand-receptor complex (Coleman et al., 2016, Plenge et al., 2020).

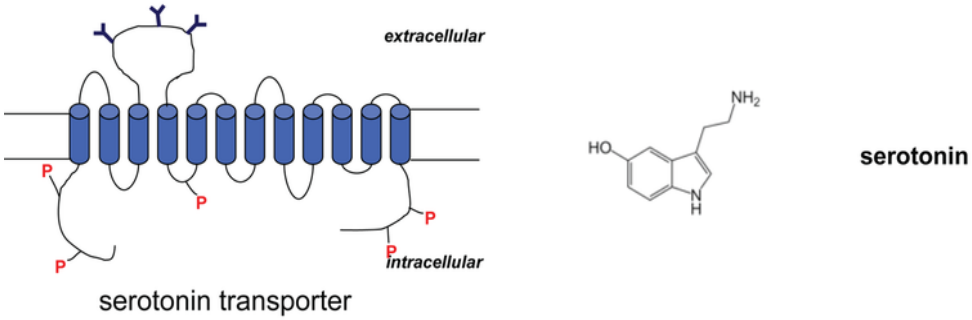


Image 7: Structure of the Human Serotonin Transporter (SERT) (Longone, P., 2011).

1.3 Importance and Tools for Ligand-Protein Binding Affinity Prediction

The treatment of major depressive disorder remains a challenging issue, especially from the moment that a specific cause that triggers it has not been found yet. As usual for several diseases, most experiments rely on the prediction of ligand-protein interactions. The main metric is called inhibitory constant (Ki) or binding affinity. It measures the strength of the binding interaction between a protein like SERT and a potential drug, inhibitor like SSRIs (Malvern Panalytical). The inhibitory constant (K_i) is a type of equilibrium dissociation constant (K_d) and represents the concentration at which the inhibitor-ligand occupies 50% of the receptor sites when no competing ligand is present. The value of Ki is inversely proportional to the binding strength of the ligand relative to its target (Canadian Society of Pharmacology and Therapeutics, 2024). Binding affinity is determined by the intermolecular interactions like hydrogen bonds, hydrophobic bonds, van der Waals bonds and other types of bonds like electrostatic. It reveals different aspects of the structural and functional part of a protein shedding light on the drug discovery processes for the design of new therapies. However, this study does not aim to build a regression model that shows output values of binding affinity, instead it classifies ligands based on the affinity with SERT. Traditionally, binding affinity is measured in a laboratory with several techniques that are mentioned in Chapter 2.1, but it is very time-consuming and has low throughput. As a result, there is an urgent need for the development of novel computational tools that calculate binding affinity, speeding up drug discovery processes at lower costs (Jarmoskaite et al., 2020).

1.4 Thesis Objectives and Structure

The main purpose of this thesis was to develop a supervised machine learning model that classifies ligands based on their binding strength to SERT protein. The categorization labels were strong binders, moderate binders and non-binders and were based on experimentally Ki values from scientific databases such as ChEMBL, BindingDB, DrugBank and PDSP Ki Database. The model integrated molecular features from known antidepressants (such as SSRIs, SNRIs, TCAs, TeCAs and serotonin modulators) alongside non-binding compounds that were extracted from the molecular docking software “AutoDock Vina” in combination with the visualization software “Chimera”. Further

interaction analysis was conducted using the software BIOVIA Discovery Studio to obtain a variety of valuable features. By using known classification machine learning algorithms, the following study strived to identify patterns among molecular interactions that affect binding affinity. Last but not least, this work aimed to determine the most predictive and critical SERT residues that differentiate strong binders from moderate and non-binders using explainable AI tools. The combination of docking features and machine learning methods has become a standard practice in the drug development research due to the mechanistic insights that are provided by docking and generalization by artificial intelligence.

The structure of this thesis is as follows:

Chapter 2: Literature Review

Description of approaches that have been applied for ligand binding classification tasks, including molecular docking and machine learning techniques involving SERT protein.

Chapter 3: Methodology

Configuration and labeling of the dataset, ligands and protein selection, preprocessing steps in docking software and extraction of interaction analysis features. Furthermore, the preprocessing steps of inputs in the model are mentioned, features are correlated, classification algorithms are described and finally model training, evaluation of parameters and explainable AI tools are analyzed.

Chapter 4: Results and Analysis

Perform initially all docking binding affinity scores with RMSD values. Additionally, the evaluation metrics of the algorithms used are performed. Lastly, the most predictive features that drive decision of algorithms towards a specific class are identified, while also common patterns for distinguishing the multi-classification model.

Chapter 5: Conclusion and Discussion

Interpretation of results based on biological evidence from literature review. Enumeration of assumptions and limitations of the dataset, such as sample size or docking assumptions. Key contributions are being pointed out with further suggestions and enhancements for future research in the field of antidepressant classification tasks with the aid of docking and machine learning and potentially facilitating drug discovery process.

1.5 Structure of Dataset

The dataset in this study consists of 74 ligands based on known inhibition constants (Ki) from widely used databases such as DrugBank, ChEMBL, BindingDB and PDSP Ki Database. Many of these

ligands are approved antidepressants, under investigation molecules and several known drugs that do not bind to SERT protein. More specifically the categories are:

I. Selective Serotonin Reuptake Inhibitors (SSRIs): They usually have high binding affinity, therefore low Ki values and are labelled as strong binders
II. Serotonin-Norepinephrine Reuptake Inhibitors (SNRIs): Moderate binding affinity with higher Ki values than strong SERT binders and are labelled as moderate binders
III. Tricyclic and Tetracyclic Antidepressants (TCAs and TeCAs): Different class of antidepressants with moderate Ki values and labelled as moderate binders
IV. Serotonin Modulators and Other Ligands: These drugs do not act the same way as SSRIs, they work as partial agonists or allosteric modulators, and their Ki values vary.
V. Weak binders: Known drugs that have minimal to zero interaction with SERT protein and therefore their Ki values is considered zero.

Chapter 2: Literature Review

2.1 Laboratory and Computational Methods for Binding Affinity Prediction

There are several ways to calculate binding affinity for a specific ligand-protein complex either in a laboratory or in a computer-based method. Enzyme-linked immunosorbent assay (ELISA) is a labelled method where a reagent detects an immobilized ligand bounded to the protein and this facilitates the screening process. On the other hand, there are also some label-free methods, such as isothermal titration calorimetry (ITC) that estimates binding affinity through differences in heat, while surface plasmon resonance (SPR) estimates affinity through changes in refractive index. In addition, biolayer interferometry (BLI) calculates light reflection from a biosensor and grating-coupled interferometry (GCI) which operates with phase-shift signals that emerge during the interaction. All these are some of the most common lab-based techniques for predicting binding affinity (Malvern Panalytical). However, these lab techniques share some disadvantages. For example, ELISA has low specificity for the target molecule, it is highly expensive and time-consuming (Merkel Technologies Ltd). ITC offers unreliable results when the changes in heat are small because the final signal is weak. Also, like ELISA, it is expensive since multiple samples for each ligand should be tested (Palacios-Ortega et al., 2021). The SPR has low throughput as a negative characteristic, while BLI shows poor reproducibility (Nicoya Biotechnology Company). Last but not least, the GCI system requires specialization and right training for handling, it is easily affected by environmental factors like vibrations and it is expensive (Hajnalka Jankovics et al., 2020).

To overcome these limitations, scientists have been focusing on computational methods for predicting binding affinity of ligands recently. A quick and low-cost method is Molecular Docking, in which molecules are tested in various conformations inside the binding pocket of a targeted protein. Another powerful tool is Molecular Dynamics Simulation (MD) that requires more time and high computational units that cannot be supported by conventional computers. MD uses Newtonian physics to simulate how atoms of a specific ligand and protein respond to predefined force fields resulting in their movement and interactions. With this way, researchers study the molecular behaviour under approximate physiological conditions in femtosecond or picosecond scale (Jafar Aghajani et al., 2022). Another widely used tool is the Quantitative Structure–Activity Relationship model which is a mathematical model that correlates ligands binding affinities to chemical descriptors that simulate a realistic environment for the binding pocket of the ligand without prior knowledge of the exact geometry of the receptor, facilitating the whole drug discovery process (Shuxing Zhang et al., 2006). A more straightforward tool, but still very precise, are the free energy calculation formulas, Free Energy Perturbation (FEP) and Thermodynamic Integration (TI). These formulas utilize thermodynamic properties of molecules to measure binding energies. However, they demand high computational costs, since their function is based on comparing two similar ligands through MD simulations by gradually converting one ligand to the other. To overcome this drawback, MM-PBSA (Molecular Mechanics Poisson–Boltzmann Surface Area) has been discovered, which estimates binding energy from snapshots of molecular dynamics simulations without requiring each time to test a pair of ligands (Brandsdal et al., 2003). Recent advances, however, focus more on developing accurate machine-deep learning models for the prediction of ligand binding affinity. Unlike classical methods, these models do not simulate molecular interactions and structure, rather they include different types of machine learning algorithms and basic neural network architectures, such as convolutional neural networks (CNNs) and graph neural networks (GNNs). Models are trained on large datasets like PDBbind or BindingDB to predict by using several key metrics, such as root-mean-square error (RMSE) or area under the curve (AUC), the binding affinity for a specific ligand-protein complex. The problem that emerges, though, is the integrity and validity of each dataset. Since the models rely on large datasets, evaluation and analysis should be carried out by specialists (Huiwen Wang, 2024).

2.2 Molecular Docking and Ligand Preparation

Molecular docking is a computer-based technique to fit a potential ligand into the binding site of a targeted protein. With the aid of integrated searching algorithms, the optimal conformations of a ligand are found and ranked. There are several software programs available for applying molecular

docking, such as AutodockVina, AutoDock, DockThor, GOLD, FlexX and Molegro Virtual Docker. Their differences lie on several aspects as the following table reveals (Fan et al., 2019), such as searching algorithms, scoring functions, flexibility of protein-ligand complex, speed and accuracy.

Table 1 Representative software for molecular docking

Name	Search algorithm	Evaluation method	Speed	Features & Application areas
Flex X [33]	Fragmentation algorithm	Semi-empirical calculation on free energy	Fast	Flexible-rigid docking. It can be used for virtual screening of small molecule databases by using incremental construction strategy
Gold [34]	GA (genetic algorithm)	Semi-empirical calculation on free energy	Fast	Flexible docking. It is a GA-based docking program. The accuracy and reliability of this software have been highly evaluated
Glide [35]	Exhaustive systematic search	Semi-empirical calculation on free energy	Medium	Flexible docking. This software uses domain knowledge to narrow the searching range and has XP(extra precision), SP (standard precision) and high throughout virtual screen modes
AutoDock [36]	GA (genetic algorithm) LGA (lamarckian genetic algorithm)	Semi-empirical calculation on free energy	Medium	Flexible-rigid docking. This software is always used with Autodock-tools and it is free for academic use
ZDOCK [37]	Geometric complement-arity and molecular dynamics	Molecular force field	Medium	Rigid docking. Chen *et al.* [37] propose a new scoring function which combines pairwise shape complementarity(PSC) with desolvation and electrostatic and develop the ZDOCK server [38]
RDOCK [39]	GA(genetic algorithm) MC (monte carlo) MIN (Simplex minimization)	Molecular force field	Medium	Rigid docking. The CHARMm-based procedure for refinement and scoring. Besides predicting the binding mode, it is especially designed for high throughput virtual screening (HTVS) campaigns
LeDOCK [40]	Simulated annealing (SA) Genetic algorithm (GA)	Molecular force field	Fast	Flexible docking. LeDock is a new molecular docking program. From the results of the present study [41], since it is fast and exhibits a high accuracy, it is recommended for the virtual screen task
Dock [42]	Fragmentation algorithm	Molecular force field	Fast	Flexible docking. It is widely applicable and is always used in docking between flexible proteins and flexible ligands
Autodock Vina [6]	GA (genetic algorithm)	Semi-empirical calculation on free energy	Fast	Flexible-rigid docking. AutoDock Vina employs an iterated local search global optimizer and it is faster than the AutoDock 4

Image 8: ***Characteristics of common-used molecular docking software tools (Fan, J., Fu, A., & Zhang, L., 2019).***

The general workflow of molecular docking is illustrated on the following outline (Image 9). Based on this, the files of the targeted protein and ligands should be downloaded from a database, such as Protein Data Bank and PubChem respectively. These files have been obtained from experimental methods and contain the 3D atomic coordinates of a specific molecule providing information about their structure. Then, protonation states, charges and hydrogen atoms should be applied correctly to both structures to reflect a realistic environment. After this, the binding site of the protein is determined either from known biological sources or from specific software programs. Finally, search algorithms and score functions operate to extract the best poses for each ligand conformation (Torres et al., 2019).

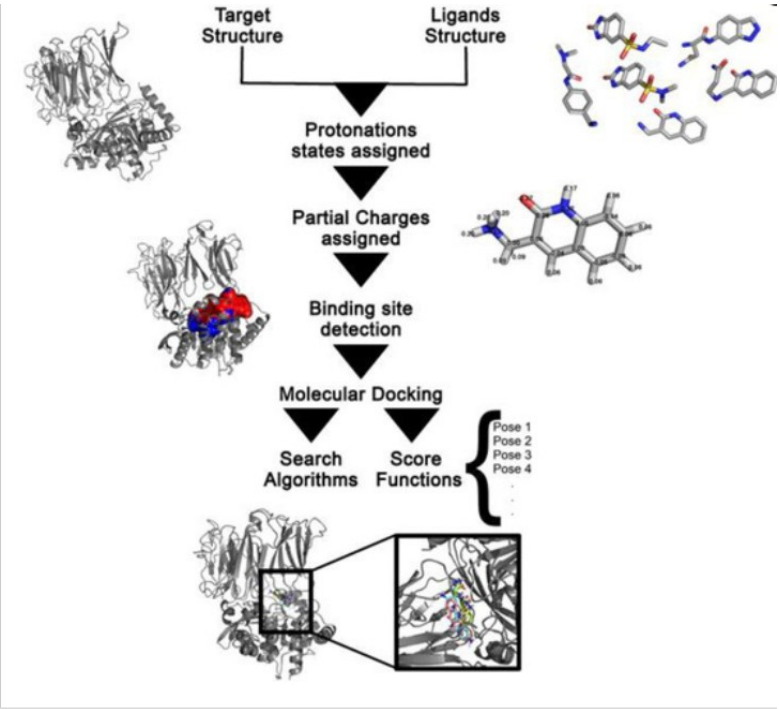


Image 9: Workflow of Molecular Docking Process (Torres, P. H. M. et al., 2019).

Based on the ligand-receptor flexibility, docking is distinctively separated into 3 major categories: Rigid docking, where both receptor and ligand are rigid and the whole simulation time is much faster, but it does not reflect the reality. To balance this drawback, there is the option of partial flexibility of ligand and rigid receptor which is obviously more realistic. On the other hand, exclusively flexible ligand and receptor can be used which requires high computational time cost, although it simulates the ideal condition in a living organism, such as human (Mohanty & Mohanty, 2023).

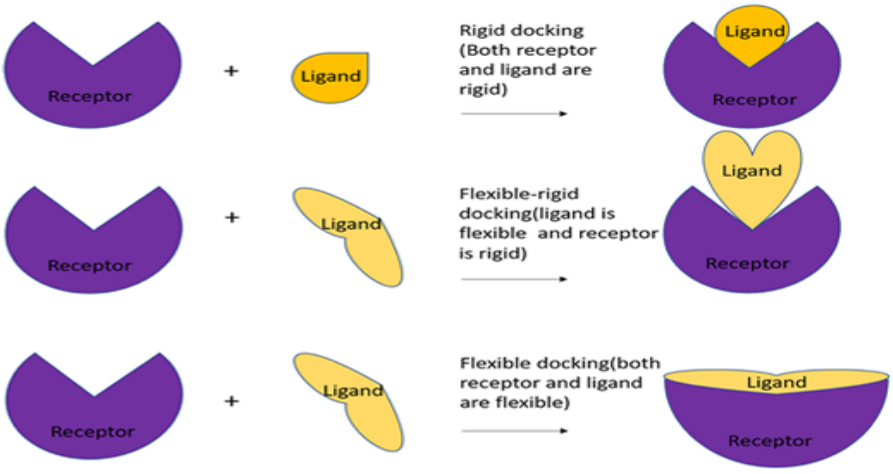


Image 10: Types of Docking Process (Mohanty, M., & Mohanty, P. S., 2023).

There are 2 ways to perform the docking simulation. The first one is called site-specific docking, where the active site of the targeted protein is known from literature or other biological evidence and the user can adjust the grid box accordingly. The other option is called “blind docking” and it is useful when binding site is unknown or if possible allosteric sites exist. It does extensive research in the whole protein to find the most representative binding pocket of a ligand and it requires higher computational cost and time than the site-specific method (Mohanty & Mohanty, 2023).

As for the docking model, there are 3 different cases which are shown in the figure below (Image 11). First, there is the lock-and-key model where both ligand and receptor are rigid and have complementary shapes for interaction. Secondly, in the induced-fit model, a conformational change in shape undergoes in the receptor after ligand binding, while in the last model there are more variations in protein-shape and the ligand binds selectively to one of these (Mohanty & Mohanty, 2023).

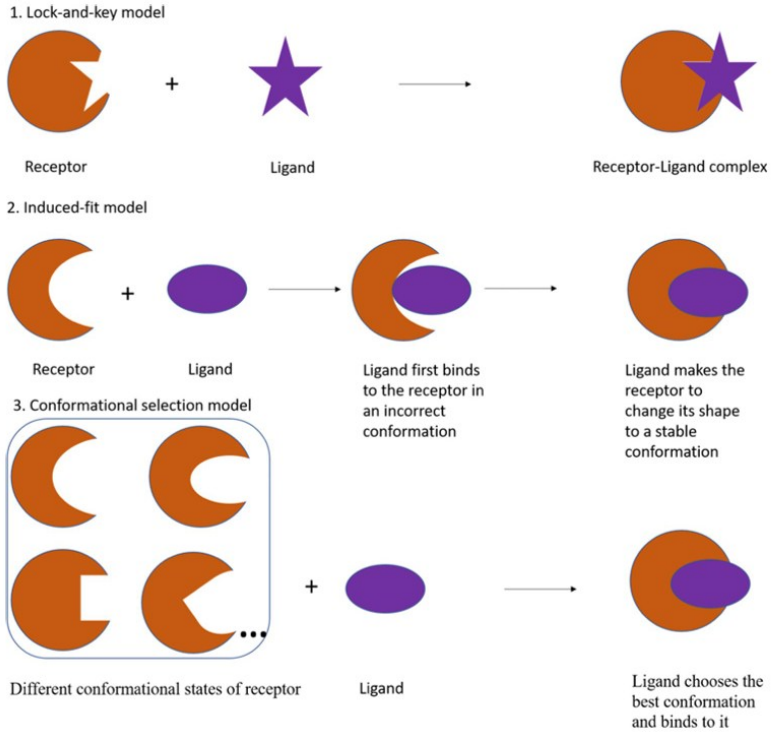


Image 11: Docking Models (Mohanty, M., & Mohanty, P. S., 2023).

2.2.1 AutoDock Vina

A widely used and compatible tool for molecular docking is the AutoDock Vina which was developed at The Scripps Research Institute in San Diego California, and it belongs to the AutoDock suite

platform. This tool is applied to the present study. It surpasses in speed and ease upon handling in comparison to other software programs by utilizing multiple central processing units. At the same time it achieves higher accuracy than the previous AutoDock4 program (Eberhardt et al., 2021). Since Autodock Vina is a newer and more refined version than other docking tools, it does not require for the user to regulate many parameters before implementing the docking simulation and it also enables multiple docking simultaneously. Vina employs a Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm for local optimization combined with a stochastic global search using Monte Carlo sampling to create an iterated local search global optimizer. This iteration is achieved through random perturbations and local energy minimization steps, causing extensive search in the conformational space to find the optimal poses of a tested ligand. A quasi-Newton optimization algorithm, the BFGS method quickly converges on a local energy minimum by approximating the second derivative of the scoring function, therefore improving docking accuracy (Trott & Olson, 2010).

The output that emerges from molecular docking software programs is the binding affinity of a ligand and some RMSD values. These outputs arise from the prediction of some scoring functions which can be classified broadly into four categories, force field-based, empirical, knowledge-based, and machine learning-based scoring functions. Force field-based functions apply classical laws of physics based on energy terms from protein–ligand complex non-bonded interactions and internal ligand energy, taking into account the solvent environment of the docking. The solvent environment is regulated through continuum models like Poisson–Boltzmann (PB) or Generalized Born (GB). Examples of softwares that integrate force field-based functions are DOCK and DockThor. Empirical functions work through regression analysis and match energy terms with experimental known data. Knowledge-based functions conduct statistical analysis of the interacting atoms of the ligand-protein complex and convert it into the preferred geometries (Guedes et al., 2018). Machine learning-based scoring functions use known algorithms or neural networks that are trained on large datasets with known ligand affinities for a specific protein. They do not focus on physics, but they try to “learn” some common patterns from the dataset in order generalize satisfactorily in new unseen samples (Ballester & Mitchell, 2010).

AutoDock Vina utilizes an empirical formula that is the following:

$$\Delta G = -0.0356 * Gauss1 + (-0.00516) * Gauss2 + 0.840 * Repulsion + (-0.0351) * Hydrophobic + (-0.587) * Hydrogen\ bonding + 0.0585 * N\_rot,$$

Where:

- **Gauss1 and Gauss2**: Gaussian steric interaction terms: Both terms reflect the electrostatic Van der Waals forces between a ligand and a targeted protein (Trott & Olson, 2010).

- **Repulsion**: Steric clashes are penalized by this term, meaning that atoms that are close to others do not promote the occurrence of the interaction (Eberhardt et al., 2021).
- **Hydrophobic**: Non-polar regions of a ligand and protein contribute also to the binding process (Eberhardt et al., 2021).
- **Hydrogen bonding**: Hydrogen bonding between ligand and protein primarily affects the affinity, with electronegative atoms like oxygen or nitrogen have a higher probability of accepting a hydrogen atom (Trott & Olson, 2010).
- **N_rot**: Number of rotatable bonds. This term penalizes the system because the partial ligand flexibility decreases the entropy upon binding (Eberhardt et al., 2021).

This formula was derived from the PDBbind dataset, and it shows the measure of strength in kcal/mol of a ligand binding to a protein and which type of bonds contribute more to the binding process with more negative values indicating stronger affinity (Trott & Olson, 2010). On the other hand, RMSD (Root Mean Square Deviation) values indicate the conformational variations of poses that occur because of the movement of heavy atoms only, based on the best-ranked pose with the lowest binding affinity. Two values are obtained, the upper bound RMSD (rmsd/ub) and the lower bound RMSD (rmsd/lb). The former compares each atom of one pose to the exact same on another but ignores the symmetry factor in the molecule. The latter term aims to minimize deviation across poses by comparing each atom of one pose to the closest element type atom from another, leading to a more realistic result (Trott & Olson, 2010; AutoDock Vina Manual). Both of them give valuable insights into the resulting docking simulation which drive drug discovery processes in many cases.

2.2.2 Chimera software

Among the well-established molecular modelling and visualization software programs, is UCSF Chimera. It was developed by the University of California, San Francisco's Biocomputing, Visualization, and Informatics team. Due to its known advantages, such as the flexible settings and easy handling, it is widely used in the field of structural and computational biology, especially for the drug discovery process. Chimera supports a variety of formats like PDB, MOL2, SDF, and PDBQT files. Users can explore in detail the three-dimensional structure of a protein, ligand or a complex and recognize all types of interactions that occur during docking or just inspect the structures alone prior to any simulation. It is very useful for ligand and protein preparation, since specific tasks can be performed, for example addition of hydrogen atoms, assignment of charges, removal of unwanted atoms or chains and adjustment of force field parameters (Pettersen et al., 2004).

Chimera offers the ability to measure distances, torsions, angles, surface area and binding pockets of a receptor. It can be connected to docking engines, such as AutoDock or AutoDock Vina and visualize the docking results there. In addition, other software platforms like BIOVIA Discovery Studio support output files that emerge from AutoDock Vina in combination to Chimera, to extract two-dimensional maps that depict as graphs the molecular interactions and residues included in the docking simulation (Butt et al., 2020). A more advanced and newer version of USCF Chimera is ChimeraX which prevails in functional and modelling settings of the classic version, though USCF Chimera is used in this study.

2.3 Machine Learning Applications in Ligand-Protein Affinity

In recent years, the drug discovery community has focused on inventing machine learning techniques that overcome the difficulties one faces during traditional molecular docking and molecular simulation software platforms in order to predict binding affinity of a ligand. These computational tools require several assumptions during their application, they are highly time-consuming and computationally expensive. For this reason, machine learning models act as promising solution for approaching ligand binding either numerically or categorically. Machine learning models aim to "learn" specific patterns from a variety of molecular or interaction features in order to increase the generalization to real world scenarios with high validity, speed and less costly (Ballester & Mitchell, 2010).

Machine learning methods can be separated into 2 types, traditional machine learning methods and deep learning techniques. In the former case, traditional algorithms, such as Random Forest, Logistic Regression, k-Nearest Neighbors and XGBoost are employed in a model. As for features, they usually consist of molecular descriptors, like molecular weight and polar surface area, molecular interaction details, such as hydrogen, hydrophobic bonds, angles etc. and some specific residue analysis of the binding site of the protein. Models are trained on approved and large datasets, for example BindingDB and ChEMBL and they can serve different purposes. Regression tasks involve the determination of a continuous numerical value which in such cases is the binding affinity of a ligand to a specific protein or receptor. On the other hand, there are the classification models, either binary or multiclassification, in order to categorize a number of ligands based on one or more features. Several evaluation metrics are used to evaluate results, such as root mean squared error (RMSE) and mean absolute error (MAE) for regression analysis, F1-score, accuracy, precision and recall for classification problems (Wang et al., 2024).

A more sophisticated approach than the traditional machine learning models, are the deep learning algorithms. The advancement is due to the spatial component that they can integrate in training process. This means, that deep learning algorithms learn from raw molecular formats instead of

receiving inputs of molecular features extracted from other sources, such as Docking and Visualization Softwares. Convolutional neural networks (CNNs) can interpret and understand the three-dimensional coordinates of atoms included in the protein-ligand complex, while Graph Neural Networks (GNNs) utilize molecular graphs where atoms and bonds are represented as bonds and edges respectively (Torres et al., 2019).

In the context of the following thesis, a supervised machine learning framework will be constructed to classify ligands based on serotonin transporter (SERT) inhibitors as strong, moderate and non-binders. Intrinsic molecular characteristics, interaction information and residue level contact analysis derived from the docking and visualization programs will serve as input to the machine learning model.

2.4 Related Work on Ligand Binding Prediction and Machine Learning in SERT and Docking Studies

Predicting ligand–protein binding interactions computationally has been a vital component of drug discovery in recent years. In the vast majority of studies, the serotonin transporter (SERT) has been the main focus of ligand screening in depression research and has been used for potential inhibitor or modulator identification. Machine learning (ML) integrated with molecular docking features show increased reliability and cost-effectiveness of experiments, decreased screening time and decode structure–activity relationships. In this section, relevant studies have informed the present thesis, in the context of docking, dataset structure and ML classification for predicting ligands binding to SERT protein transporter.

In the study titled "Prediction of 5-hydroxytryptamine Transporter Inhibitor Based on Machine Learning", researchers managed to apply a binary classification task related to SERT inhibitors and non-inhibitors. Samples were collected from ChEMBL and DrugBank databases and were separated based on the IC50 variable. IC_{50} is the concentration of the competitive antagonist required to reduce the activity/binding of an agonist to a specific enzyme, receptor or transporter by 50% (Canadian Society of Pharmacology and Therapeutics, 2024). Specifically, compounds with $IC_{50} < 500$ nM were labelled as inhibitors, while those with $IC_{50} > 1000$ nM were considered non-inhibitors. The dataset consists of 812 inhibitors and 400 non-inhibitors. For this reason SMOTE (Synthetic Minority Over-sampling Technique) technique was applied to reduce the severe class imbalance. Final number of samples reached 1920 and a variety of features including circular, topological and physicochemical descriptors from RDKit tool were extracted. These features capture not only the atomic correlations but also the global structure of the complex. Conventional machine learning algorithms were applied such as Random Forest (RF), k-Nearest Neighbor (KNN), Support Vector Machines (SVM), Logistic regression (LR) and Voting ensemble Classifier (VOL_CLF). The results of this model prove that on

internal tests precision in inhibitors of SERT ranged from 90.7% in Random Forest, to 93.3% in the VOL_CLF algorithm. Highest recall was achieved also with RF. In non-inhibitors highest precision was achieved with RF classifier and highest recall with VOL_CLF. On the external test set, RF again showed better performance in non-inhibitors with precision 95.7% and recall 73.3% respectively, while in SERT-inhibitors the results were not satisfactorily (Kong et al., 2019).

Another research approach conducted by Sharma and Dang in 2022, was to compare the binding affinity and interactions of phytochemical compounds with SERT protein to the standard SSRIs, for possible antidepressant activity. For docking, AutoDock 4.2 software and the Lamarckian Genetic Algorithm (LGA) were used, while for visualization Discovery Studio Visualizer 2020 and PLIP. Ten natural compounds were downloaded from PubChem database and tested in parallel to the standard drug Paroxetine and the results are in the following table (Image 12):

S.No	Name of Natural Compound/Standard Drug	Binding Energy(kcal/mol)	Amino acid residues Showing Interaction
1	Withaferin A	-8.75	ASP401,LEU25,LEU29,PHE320,ALA319,VAL33,ILE111
2	Piperine	-6.80	ARG30,LEU25,LEU29,ALA319,VAL33,ILE111
3	Hyperforin	-6.80	ARG30,ALA319,VAL33,ASP404,PHE405,ILE475,TYR471,TRP467
4	Naringenin	-6.62	ALA319,PHE320,LLEU162,VAL393,LEU400,SER399
5	Resveratol	-6.60	ARG30,LEU25, LEU29, ALA319,VAL33,ILE111,GLY26PHE253
6	Hypericin	-6.53	ARG30,ALA319,VAL33,GLU37,PHE405,TRP467,THR409,ILE249
7	Paroxetine (Standard Drug)	-8.05	ARG30,LEU25,LEU29,ALA319,PHE320,VAL33,GLN34, PHE253

Image 12: Binding Energy and Amino acids residues involved in interactions with the Ligands and standard drug Paroxetine (Sharma, S., & Dang, S., 2022).

As the above table depicts, although it is known that paroxetine strongly binds to SERT with binding affinity here -8.05 kcal/mol, Withaferin A outperformed in binding energy and shared 5 common amino acids residues in binding site interaction with paroxetine, indicating that it could be helpful in treating depressive symptoms. However, the rest molecules did not show any promising antidepressant effect based solely on binding energy (Sharma S. & Dang S., 2022).

As for the study “Synthesis, molecular docking and binding studies of selective serotonin transporter inhibitors” another chemical class was tested, called arylpiperidine oxime ethers for possible SERT inhibition. Reference drugs were paroxetine and fluoxetine. Ten compounds were tested in different stereoisomeric forms, which emerge from the different spatial conformation their atoms have around the double bond. A combination of experiments was conducted to investigate structure–activity relationships (SAR). Radioligand binding assays in a laboratory environment with the radio-labelled [^{3}H]-paroxetine and docking simulations in the computational tool GOLD were applied. The inhibition constant (Ki) for the ethers ranged from 10.28 nM to 396.5 nM for SERT, while it was 0.31 and 5.80 for paroxetine and fluoxetine respectively. Through docking simulations, researchers came up to the conclusion that the primary binding site for known SSRIS is between transmembrane regions TM1,

TM3, TM6, and TM8 of SERT protein. Key residues in these regions are aspartame 98, which forms electrostatic bonds with the amine group of paroxetine, threonine 439 that forms hydrogen bonds and finally tyrosine 95, Phenylalanine 341, and Asparagine 177 that enable hydrophobic interactions. All these residues are in agreement with what literature mentions about critical residues for SERT inhibition (Nencetti et al., 2011).

In their general review, Crampon et al. (2022) try to overcome the limitations of traditional molecular docking techniques with integrated machine learning and deep learning pipelines. These disadvantages lie on the fact that docking requires high number of computational units while at the same time scoring functions and accuracy remain a controversial issue. A machine learning model approach is proposed called "SIEVE-Score" which encodes specific residues that enable protein-ligand binding by using variables related to individual energy terms, such as Van der Waals interactions, electrostatic and hydrogen bonding. The machine learning model is then trained on these input features in order to define the structure of the complex and the pattern of the interactions that occur locally. In addition, it is underlined, that for better and more valid results, explainable artificial intelligence algorithms can be implemented, like Random Forests or Decision Trees feature importance tools. These tools allow researchers to link specific residues interactions with known physical quantities, for example molecular weight or polar surface area. In conclusion, this combination of structurally and biologically meaningful features increase the generalizability of the models with more reliable and quicker results than conventional docking, making it a useful tool for the drug discovery community (Crampon et al., 2022).

An informative research paper that analyze thoroughly the key residues that are involved in SERT ligand binding has been conducted by Andersen et al in 2009. Research team tested the known SSRI drug "escitalopram" with the current brand names "Cipralex" and "Lexapro". It is the S-enantiomer of racemic citalopram which is more potent and selective than R-citalopram. To explore the binding site of this known SSRI, scientists combined mutational mapping and modelling through radioligand binding assays and docking simulations. Sixty four mutations were applied to SERT in order to define the most critical residues that are highly connected with inhibition action. The critical region was between transmembrane regions TM1, TM3, TM6 and TM8 as the following image shows.

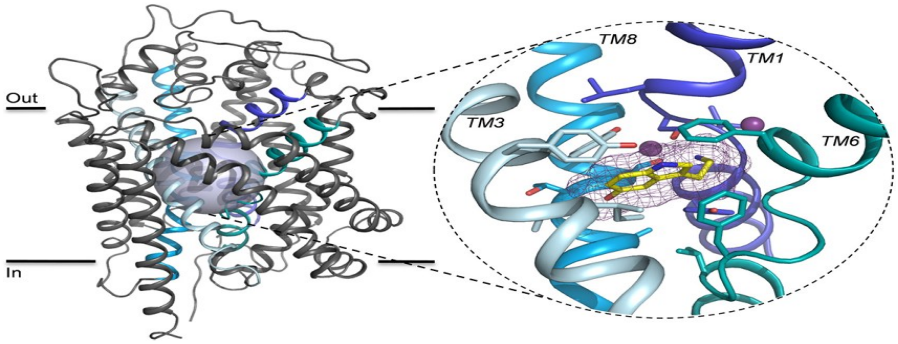


Image 13: Critical Binding sites of SERT inhibition (Andersen, J. et al, 2010).

As several research papers reveal, there are many residues that affect SERT binding. However in this study, the most critical seemed to be TYR95, ILE172**,** PHE341**,** SER438, ASP98 and ASN177 which altered the inhibition constant 10 to 400-fold meaning much weaker inhibition. First 3 residues were involved in hydrophobic bonds, SER438 in polar, ASP98 in electrostatic interactions and ASN177 in hydrogen bonding (Andersen et al., 2010).

One of the most recently published research papers is titled "Predicting Selective Serotonin Re-Uptake Inhibitors Potency: Machine Learning and Molecular Docking Approach" and was conducted by Adejoro and Adewara in 2025. The goal was to predict the potency of 2616 ligands for SERT protein. They used a cheminformatic tool named "PaDEL-Descriptor" in order to extract 12 molecular fingerprints, each of which gives a different structural and molecular information. For example "MACCS" is a binary vector that shows the presence or absence of some specific chemical groups, such as aromatic rings. In addition, "Estate" descriptors inform researchers about the electronic status and the polarity of the protein-ligand complex, while "Klekotha–Roth" fingerprint is a high-dimensional vector that acts as "MACCS", but it integrates a large chemical library with a huge number of subgroups to search for. Ten ensemble machine learning algorithms were trained and the best results were obtained with Extra Trees Regressor in combination with the Klekota–Roth fingerprint. Coefficient of determination (R^2) was calculated 0.92 and root mean square error (RMSE) 0.01. The former proves that the model can effectively explain the variance of inhibition constant results, while the latter shows that predicted values of Ki do not differ significantly from real values. Both metrics increase the generalizability of the model and research in general. The fifty molecules with the lowest Ki value were then selected for docking studies and interaction paths with SERT protein by using AutoDock Vina and PyRx software platforms. Several compounds performed better than known SSRIs in the binding affinity term. TYR95, ILE172, and SER438 were the critical residues that facilitate SERT binding as it was indicated by the interaction analysis. These residues are located in the S1 domain of the binding pocket and the first 2 participate in hydrophobic interactions, while SER438 in hydrogen bonding (Adejoro & Adewara, 2025).

Chapter 3: Methodology

3.1 Overview of the Workflow

The present study proposes a hybrid computational approach that integrates molecular docking simulations, deep interaction analysis and machine learning model in order to classify ligands based on the affinity they have with SERT protein. First step was to obtain the structure of SERT from RCSB protein Data Bank (PDB ID: 5I6X) and prepare it effectively for docking in UCSF Chimera software. Similar procedure was applied to the 74 ligands that comprised the dataset and then they were docked to the protein with the aid of AutoDock Vina. Chimera enabled visualization of the top 10 ranked poses, while through a custom script in Python the best five were selected taking into account both binding affinity and RMSD values. Structural analysis was then conducted through BIOVIA Discovery Studio to extract a variety of meaningful features related to interactions, residues and geometrical details. A comprehensive dataset was constructed consisting of hundreds of features, while for samples the top 5 poses for each ligand were aggregated by the median value into one single vector, resulting in 74 ligands. Categorization was made into strong, moderate and non-binders with SERT based on experimentally validated Ki values from established databases such as ChEMBL, PDSP Ki Database, DrugBank and BindingDB. A supervised machine learning model pipeline was built and classical algorithms were trained and evaluated on these features. Finally, explainable AI techniques were applied to capture the most informative and predictive features for the decision of models. With all this pipeline, virtual screening process for the discovery of antidepressant drugs can be facilitated. However enhancements and additions should be implemented in order to be considered a scientifically reliable approach.

3.2 Ligand and Protein Selection and Preparation

3.2.1 Protein Preparation

The target protein used in this study was retrieved from RCSB Protein Data Bank and the identification number is 5I6X.

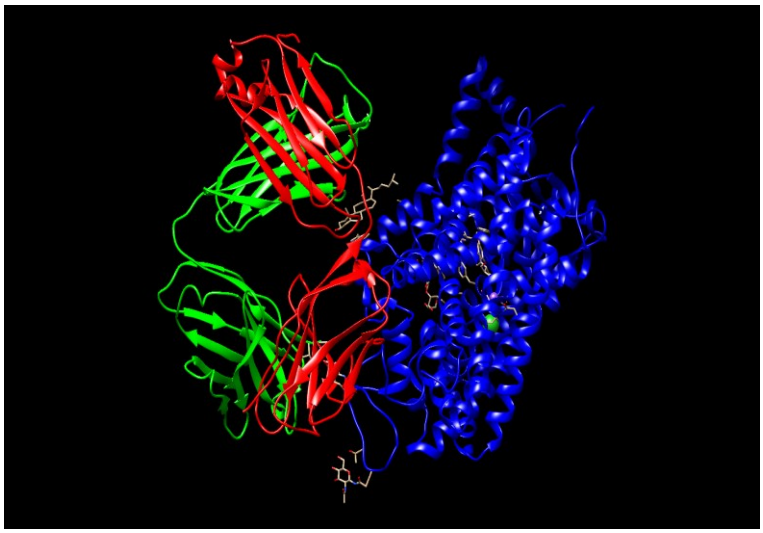


Image 14: 5I6X structure in Chimera.

The image above, which shows paroxetine bound to the central site of the protein, was obtained through X-ray crystallography and has 3.14 Å resolution. 5I6X structure includes apart from the main chain A (blue colour in image 14), which is the human serotonin receptor protein, chain B (green colour) and chain C (red colour) which are heavy and light chains respectively of the 8B6 monoclonal antibody derived from "*Mus musculus"*. They were placed only for crystallization purpose and that was to stabilize SERT protein. In addition, chain A is surrounded by several ligands. Sodium (Na^+), chloride ions (Cl^-) and water molecules (H_2O) enable the proper functioning of the protein. Cholesterol (CLR) on the other side might be useful for stabilizing the membrane protein environment, while Dodecyl-beta-D-maltoside (LMT) is a detergent molecule. Both seem not to contribute to the binding process of a ligand in SERT. 2-acetamido-2-deoxy-beta-D-glucopyranose (NAG) is observed only in glycosylation sites far from the binding pocket. Paroxetine is mentioned as 8PR and it is the reference ligand in the whole structure. The actual protein sequence of human SERT is depicted in image 15 which was obtained from UniProt database (ID: P31645). It consists of 630 amino acids and its molecular weight is 70.325 Daltons. The protein sequence that was used in the present study is in image 16.

```
                10        20        30        40        50        60        70        80        90
METTPLNSQK QLSACEDGED CQENGVLQKV VPTPGDKVES GQISNGYSAV PSPGAGDDTR HSIPATTTTL VAELHQGERE TWGKKVDFLL

               100       110       120       130       140       150       160       170       180
SVIGYAVDLG NVWRFPYICY QNGGGAFLLP YTIMAIFGGI PLFYMELALG QYHRNGCISI WRKICPIFKG IGYAICIIAF YIASYYNTIM

               190       200       210       220       230       240       250       260       270
AWALYYLISS FTDQLPWTSC KNSWNTGNCT NYFSEDNITW TLHSTSPAEE FYTRHVLQIH RSKGLQDLGG ISWQLALCIM LIFTVIYFSI

               280       290       300       310       320       330       340       350       360
WKGVKTSGKV VWVTATFPYI ILSVLLVRGA TLPGAWRGVL FYLKPNWQKL LETGVWIDAA AQIFFSLGPG FGVLLAFASY NKFNNNCYQD

               370       380       390       400       410       420       430       440       450
ALVTSVVNCM TSFVSGFVIF TVLGYMAEMR NEDVSEVAKD AGPSLLFITY AEAIANMPAS TFFAIIFFLM LITLGLDSTF AGLEGVITAV

               460       470       480       490       500       510       520       530       540
LDEFPHVWAK RRERFVLAVV ITCFFGSLVT LTFGGAYVVK LLEEYATGPA VLTVALIEAV AVSWFYGITQ FCRDVKEMLG FSPGWFWRIC

               550       560       570       580       590       600       610       620       630
WVAISPLFLL FIICSFLMSP PQLRLFQYNY PYWSIILGYC IGTSSFICIP TYIAYRLIIT PGTFKERIIK SITPETPTEI PCGDIRLNAV
```

Image 15: Protein Sequence of human SERT (UniProt database, ID: P31645).

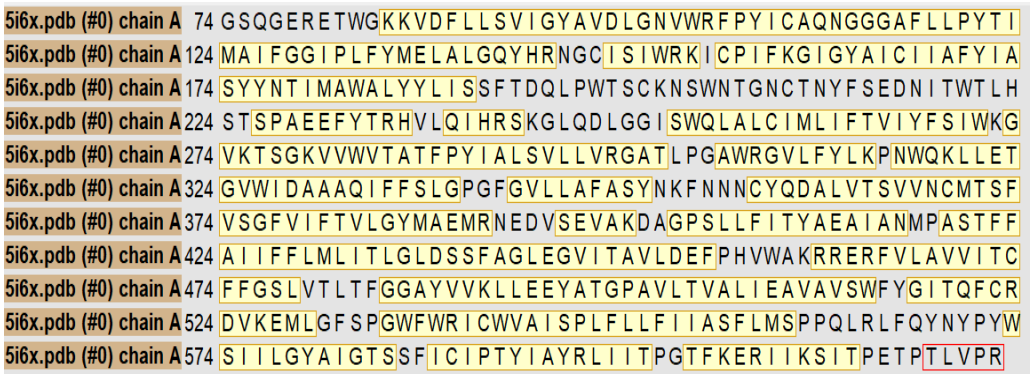


Image 16: Protein sequence of human SERT structure 5I6X from RCSB.

First of all, SERT protein transporter (5I6X) was imported in Swiss-Pdb Viewer (DeepView) software for a deepened inspection of the structure. It is a widely used computational tool for identifying and repairing missing atoms and residues of a given protein. Mutations can also be recognized and engineered for further simulation purposes. As the following table reveals (Table 2), there were several missing atoms in different residues of SERT structure. As a result, Swiss-Pdb Viewer was used in order to ensure the accurate replacement of missing parts and artifacts of the protein, leading to a more realistic environment in the docking process.

ID	RESIDUE	CHAIN	RESIDUE NUMBER	MISSING ATOMS
1	GLN	A	76	CG, CD, OE1, NE2
2	ARG	A	79	CG, CD, NE, CZ, NH1, NH2
3	TRP	A	82	CG, CD1, CD2, NE1, CE2, CE3, CZ2
4	TRP	A	82	CZ3, CH2
5	ASN	A	145	CG, OD1, ND2
6	LYS	A	201	CG, CD, CE, NZ
7	GLU	A	215	CG, CD, OE1, OE2
8	LYS	A	275	CG, CD, CE, NZ
9	GLU	A	463	CG, CD, OE1, OE2
10	ARG	A	464	CG, CD, NE, CZ, NH1, NH2
11	GLU	A	494	CG, CD, OE1, OE2
12	GLN	A	562	CG, CD, OE1, NE2
13	LEU	A	597	CG, CD1, CD2

14	ILE	A	598	CG1, CG2, CD1
15	ILE	A	599	CG1, CG2, CD1
16	THR	A	600	OG1, CG2
17	THR	A	603	OG1, CG2
18	PHE	A	604	CG, CD1, CD2, CE1, CE2, CZ
19	LYS	A	605	CG, CD, CE, NZ
20	GLU	A	606	CG, CD, OE1, OE2
21	ARG	A	607	CG, CD, NE, CZ, NH1, NH2
22	ILE	A	608	CG1, CG2, CD1
23	ILE	A	609	CG1, CG2, CD1
24	LYS	A	610	CG, CD, CE, NZ
25	SER	A	611	OG
26	ILE	A	612	CG1, CG2, CD1
27	THR	A	613	OG1, CG2
28	THR	A	616	OG1, CG2
29	GLU	B	20	CG, CD, OE1, OE2
30	ASN	C	232	CG, OD1, ND2

Table 2: Missing atoms of residues in 5I6X structure from RCSB.

After these modifications, the structure was uploaded into UCSF Chimera for further preparation. Chains B and C were deleted, as they did not participate in ligand binding. As for the co-crystallized ligands, they were evaluated in order to be retained or removed based on biological relevance. Dodecyl-β-D-maltoside (LMT) and N-Acetyl-D-glucosamine (NAG) were deleted since they did not fall into the binding pocket of SERT protein. Water molecules, sodium and chloride ions were maintained not only because they promote the function of the transporter, but also are located near the binding site and are the target molecules for many ligands. Cholesterol (CLR) was also kept since it is a valuable molecule for membrane stabilization. Finally, paroxetine was not removed at this stage.

Following this cleanup, next crucial step was to replace the mutations that existed in structure. From the validation report of 5I6X in RCSB Protein Data Bank it was obvious that 4 mutations have been engineered in chain A (Image 17).

Chain	Residue	Modelled	Actual	Comment	Reference
A	74	GLY	-	cloning artifact	UNP P31645
A	75	SER	-	cloning artifact	UNP P31645
A	291	ALA	ILE	engineered mutation	UNP P31645
A	439	SER	THR	engineered mutation	UNP P31645
A	554	ALA	CYS	engineered mutation	UNP P31645
A	580	ALA	CYS	engineered mutation	UNP P31645
A	619	LEU	-	cloning artifact	UNP P31645
A	620	VAL	-	cloning artifact	UNP P31645
A	621	PRO	-	cloning artifact	UNP P31645
A	622	ARG	-	cloning artifact	UNP P31645

Image 17: Discrepancies of modelled and reference sequence (validation report of 5I6X, RCSB Protein Data Bank).

By the option represented in image 18 in Chimera, alanine 291 was selected and replaced with the actual amino acid isoleucine. This was achieved through the rotamer selection library and more specifically the Dunbrack 2010 library which provides statistical data for possible side chains conformations based on probability scores. For example in case of replacement of alanine 291 with isoleucine (Image 18), the first conformation was applied since it had approximately 77% probability of occurrence. Chi 1 and Chi 2 columns were the torsion angles of side chains around their bonds, reflecting the final structure of the replaced region. Same procedure was followed for the rest 3 mutations. Serine 439 was replaced by threonine, while alanine 554 and 580 by 2 cysteine amino acids.

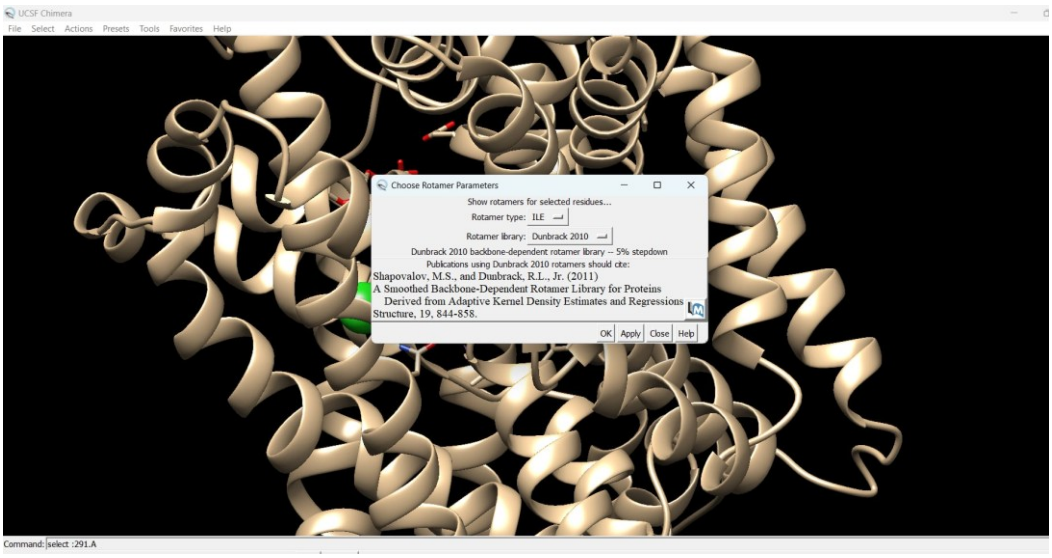


Image 18: Rotamer library for handling mutations.

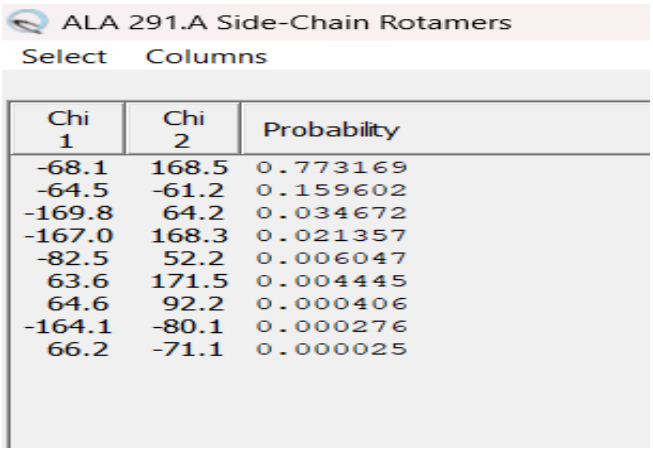


Image 19: Side-chain rotamers probabilities.

After rotamer replacement, geometric clash analysis was performed. By the *Find Clashes/Contacts* option in Chimera, it was manageable to identify potential steric hindrances to the whole structure and minimize these steric effects for a more realistic protein environment. Initially, unprocessed 5I6X complex had 12.24 clashscore and was calculated from MolProbity web-based tool. This metric reveals the number of clashes per 1000 atoms. Clashes were derived from overlaps or unrealistic van der Waals interactions and were visualized in Chimera for further optimization.

The minimization of structure was achieved through two major steps. Firstly, local minimization was performed with 200 steepest descent steps to the four mutated amino acids from rotamer library. In addition, several steric clashes, especially in critical regions of the protein that are shown from PDB files or MolProbity were alleviated. Raw crystal structure, as previously mentioned, had 12.24 clashscore, while after the usage of Swiss-Pdb Viewer the number increased to 20.52 possibly due to the addition of missing atoms and reconstructed chains. To compensate this high value, a global minimization step to the whole protein was applied. Conjugate gradient steps were set at 10 in order for the system to converge faster to a local minimum. With this way, high strains and clashes in structure were reduced. For this reason, the clashscore of the final structure after local and global minimization steps was calculated 3.3 in MolProbity, increasing the validity of results and the docking simulations.

Next step crucial for docking was the *Dock Prep* tool from structure editing option in Chimera (Image 20). Solvent and non-complexed ions options were deactivated, because water and ions of sodium (Na^+) and chloride (Cl^-) were present in the structure and critical for ligand binding and functionality of protein. Incomplete side chains were replaced from Dunbrack rotamer library. Another important step for docking and scoring functions was the addition of hydrogen atoms at physiological pH which were not represented in X-ray PDB structure. Last but not least, partial charges were assigned by using the Gasteiger–Marsili method. This is a fast, heuristic and iterative algorithm that calculates charges based on the electronegativity and polarizability of the atoms participating in bonds, while it also

accounts for the role and influence of the adjacent atoms whether they are donors or acceptors of electrons (Gasteiger & Marsili, 1980).

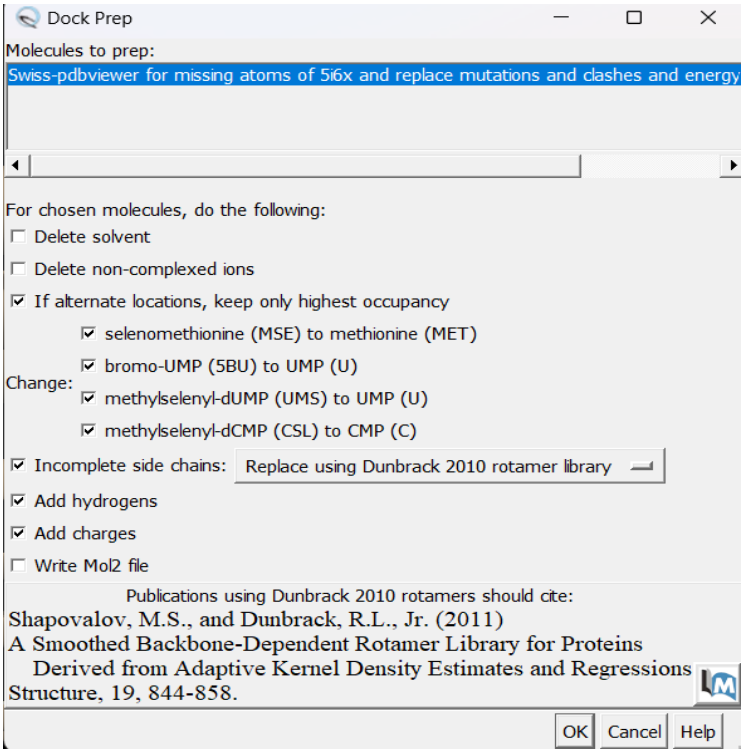


Image 20: Dock Preparation in Chimera.

The resulting structure that included the above steps, namely deletion of unwanted chains and ligands, handling of clashes, minimization of protein and docking preparation, was the following (Image 21). Blue colour is Na^+, green is Cl^- and yellow is the bound paroxetine.

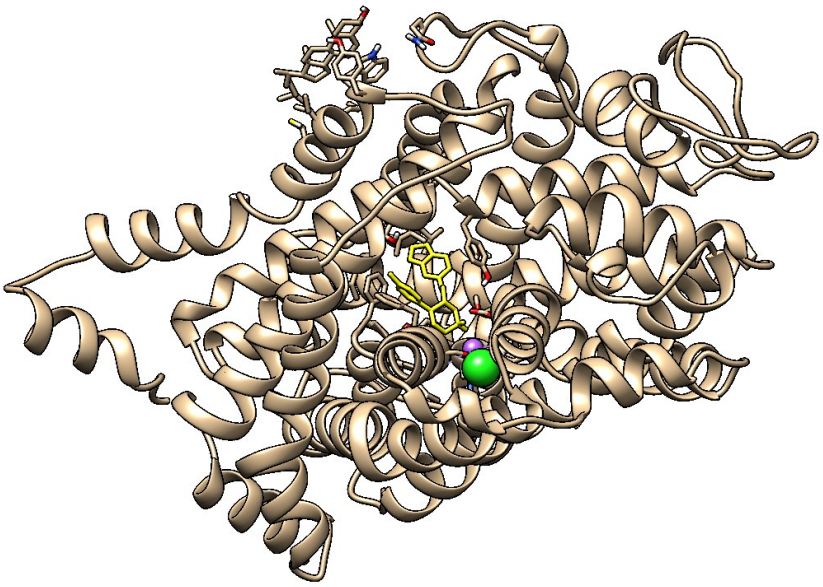


Image 21: Final SERT structure for docking.

3.2.2 Ligand Selection

The selection of the 46 ligands for docking followed the specific criteria below:

1. Similar assay type: Since the categorization of the dataset was based on inhibition constants, the assays from which Ki values were obtained must be comparable. For this reason, only radioligand binding assays were included that reflect directly ligand binding instead of uptake inhibitions that show the functionality of the protein.
2. Validated radioligands: Radioligand that were approved for use were known SERT selective compounds, such as [^{3}H] Paroxetine, [^{3}H] Citalopram and [^{3}H] Imipramine.
3. The experiments must involve only human SERT protein, so animal based experiments were excluded.
4. Ki values were obtained from validated databases, for example ChEMBL, PubChem, PDSP Ki Database from NIMH Psychoactive Drug Screening Program and DrugBank.
5. The present study focused solely on Ki values. This means that studies which calculated other similar variables like half-maximal inhibitory concentration (IC50) or Kd were excluded.

The dataset was divided into 2 parts. First part includes 21 compounds without any brand or generic name in literature. They may be under investigation for possible antidepressant activity or they may be tested in structure-activity relationship (SAR) studies, where different derivatives of the same initial

ligand are examined. The second part includes 28 ligands with known antidepressant action in human SERT protein, such as SSRI's, SNRI's, TCA's, TeCA's and other serotonin modulators.

As for the 21 compounds, they were retrieved from 2 literature sources. In first study several indole cyclopropylmethylamines analogues were examined as potential selective serotonin reuptake inhibitors. Although the authors mention that they applied radioligand displacement assays, the exact radioligand is not stated. However, it is mentioned that binding affinities were determined based on literature methods from other research papers, where [^{3}H]-citalopram was used (Taber et al., 2005), (Schmitz et al., 2005). The experiments were conducted in vitro using human SERT protein (Mattson et al., 2005).

On the other hand, the other study analyzed new compounds of piperazine and diazepane amides with simultaneous serotonin reuptake inhibition and histamine H3 receptor antagonism. As described in reference 13 of the study, researchers used radioligand binding assays with [^{3}H]citalopram to determine binding affinities of tested ligands (Barbier et al., 2007). Similar to previous study, experiments were conducted in vitro by using hSERT (Ly et al., 2008).

In conclusion both studies, although they did not clearly mention the exact steps for their experiments, they met the above criteria to a great extent. For this reason, their findings and calculations were integrated into the present thesis by forming a big part of the dataset. Below is a table with 20 compounds from the 2 aforementioned studies and their experimentally measured inhibition constants (Ki) against human SERT (Table 3). CID_44351345 is a tested ligand from another research paper, where radioligand binding assay was implemented with [^{3}H]paroxetine in humans (Takeuchi et al., 2006).

ID	Compound	Ki value	SERT Inhibition
1	CID_11310988	0.56	STRONGLY BINDING
2	CID_11447499	2	STRONGLY BINDING
3	CID_11535974	59	MODERATE BINDING
4	CID_11608403	0.58	STRONGLY BINDING
5	CID_11623136	1.8	STRONGLY BINDING
6	CID_11658763	230	MODERATE

			BINDING
7	CID_11673089	100	MODERATE BINDING
8	CID_11694324	2.1	STRONGLY BINDING
9	CID_16006089	3.3	STRONGLY BINDING
10	CID_24855949	229	MODERATE BINDING
11	CID_24855953	2.9	STRONGLY BINDING
12	CID_24855981	94	MODERATE BINDING
13	CID_24856012	68	MODERATE BINDING
14	CID_24856046	702	MODERATE BINDING
15	CID_24856107	5	STRONGLY BINDING
16	CID_24947569	5	STRONGLY BINDING
17	CID_24947939	37	MODERATE BINDING
18	CID_24964158	0.8	STRONGLY BINDING
19	CID_44351345	0.24	STRONGLY BINDING
20	CID_44390396	4	STRONGLY BINDING
21	CID_44456154	3.7	STRONGLY BINDING

Table 3: Unknown compounds with Ki values for hSERT.

Below is another table with 28 ligands that were included in my dataset (Table 4). These were known SSRI's, SNRI's, TCA's, TECA's and some serotonin modulators that are widely used in clinical practice and treat or manage depressive symptoms. The majority of Ki values were available in ChEMBL database and the experiments from which they were extracted, were radioligand binding

assays with [^{3}H] paroxetine or [^{3}H] citalopram, except for trimipramine and chlorpheniramine where [^{3}H] imipramine was used. Experimental conditions were not exactly the same across the compounds, but they all referred to human SERT protein transporter. For some ligands the Ki values were obtained from other sources like DrugBank or PDSP Ki DATABASE. For Dextromethorphan which is a cough suppressant, a value of 40nM was considered based on the research paper titled "Pharmacology of dextromethorphan: Relevance to dextromethorphan/quinidine (Nuedexta®) clinical use", were [^{3}H]paroxetine was used as radioligand (Taylor et al., 2016).

ID	COMPOUND	Ki	SERT Inhibition
22	PAROXETINE	0.043	STRONGLY BINDING
23	FLUOXETINE	0.271	STRONGLY BINDING
24	FLUVOXAMINE, PDSP KI	1.95	STRONGLY BINDING
25	CITALOPRAM	0.479	STRONGLY BINDING
26	ESCITALOPRAM	1.1	STRONGLY BINDING
27	SERTRALINE	0.075	STRONGLY BINDING
28	MILNACIPRAN, PDSP KI	8.44	MODERATE BINDING
29	LEVOMILNACIPRAN(DRUGBANK)	11	MODERATE BINDING
30	DOXEPIN	22.0	MODERATE BINDING
31	DULOXETINE, PDSP KI	0.8	STRONGLY BINDING
32	AMITRIPTYLINE	0.882	STRONGLY BINDING
33	NORTRIPTYLINE	6.977	MODERATE BINDING
34	PROTRIPTYLINE (DRUGBANK)	20	MODERATE BINDING
35	IMIPRAMINE, PDSP KI	1.3	STRONGLY

			BINDING
36	CLOMIPRAMINE	0.047	STRONGLY BINDING
37	DESIPRAMINE (DRUGBANK)	18	MODERATE BINDING
38	VENLAFAXINE (DRUGBANK)	82	MODERATE BINDING
39	DESVENLAFAXINE	15	MODERATE BINDING
40	MAZINDOL	MINIMUM:45	MODERATE BINDING
41	VILAZODONE	0.5	STRONGLY BINDING
42	VORTIOXETINE	1.6	STRONGLY BINDING
43	INDATRALINE (DRUGBANK)	4.8	STRONGLY BINDING
44	NEFAZODONE	290	MODERATE BINDING
45	REBOXETINE	1400	MODERATE BINDING
46	TRIMIPRAMINE (3h imipramine),PDSP KI	149	MODERATE BINDING
47	ATOMOXETINE	77	MODERATE BINDING
48	DEXTROMETHORPHAN	40	MODERATE BINDING
49	CHLORPHENIRAMINE	15.2	MODERATE BINDING

Table 4: Known SERT binders with Ki values.

The remaining 25 compounds that formed the dataset were randomly picked from a variety of drug classes and it was proved that they had minimal to zero affinity with hSERT and thus they were considered having zero Ki value with 5-HT protein. The list of them included:

"Acetaminophen", *"Aspirin"*, *"Bupropion"*, *"Cimetidine"*, *"Fexofenadine"*, *"Ibuprofen"*, *"Maprotiline"*, *"Metformin"*, *"Naproxen"*, *"Probenecid"*, *"Ranitidine"*, *"Amlodipine"*, *"Bromocriptine"*, *"Clozapine"*, *"Dantrolene"*, *"Diphenhydramine"*, *"Flutamide"*, *"Iprindole"*,

"Losartan", *"Metoclopramide"*, *"Mirtazapine"*, *"Naltrexone"*, *"Piroxicam"*, *"Quetiapine"* and *"Zolpidem"*.

3.2.3 Ligand Preparation

After choosing all samples next step was to download the 3D structure data file (SDF) for each ligand from PubChem. Docking was conducted automatically with the aid of AutoDock Vina on an Ubuntu-based environment. This procedure was facilitated by the use of Open Babel software, which is a chemical informatics software that handles chemical files data and molecular details (O'Boyle et al., 2011). Energy minimization was applied with Merck Molecular Force Field **(**MMFF94), which was developed by Merck and is a very effective force field mainly for organic molecules, such as the ligands that were used in this dataset. It includes parameters for the majority of atoms and ions and it is compatible with the Open Babel environment. In addition, since it takes into account all the different types of interactions that occur, refinement of structural geometry was achievable. The minimized SDF files were then converted into PDBQT format which stands for Protein Data Bank (PDB), partial charge (Q) and atom Type (T). This is the right readable format in order to perform docking simulations in Autodock Vina. Finally, all files were listed together.

3.2.4 Definition of Grid Box Size and final options prior to docking

One of the most substantial part for docking, is the configuration of the grid box size. An accurate grid box with punctual dimensions ensures that the docking will occur in a biologically relevant region of SERT protein. The coordinates and dimensions of the box are shown in Table 5, while the box is visualized in image 22. Center x, y and z values indicate the center point of the whole box, while size x, y and z are proportional to the size of the box in all axons. Obviously, a slightly enlarged box was constructed in order to capture a wider part of the protein, although higher computational cost was required and accuracy was reduced. However, the present study focused mainly on the central site of SERT protein, where paroxetine was bound, namely in transmembrane regions TM1, TM3, TM6, and TM8 of the SERT protein 5I6X crystal structure. All dimensions were recorded in a configuration file called *"conf.txt"*, where also a seed number *"1234"* was applied in order to get reproducible results in multiple ligand docking simulations.

center_x	-34,5132
center_y	-20,9471
center_z	3,38718
size_x	47,826 Å
size_y	30,9402 Å
size_z	28,9013 Å

Table 5: Grid Box Coordinates and Dimensions.

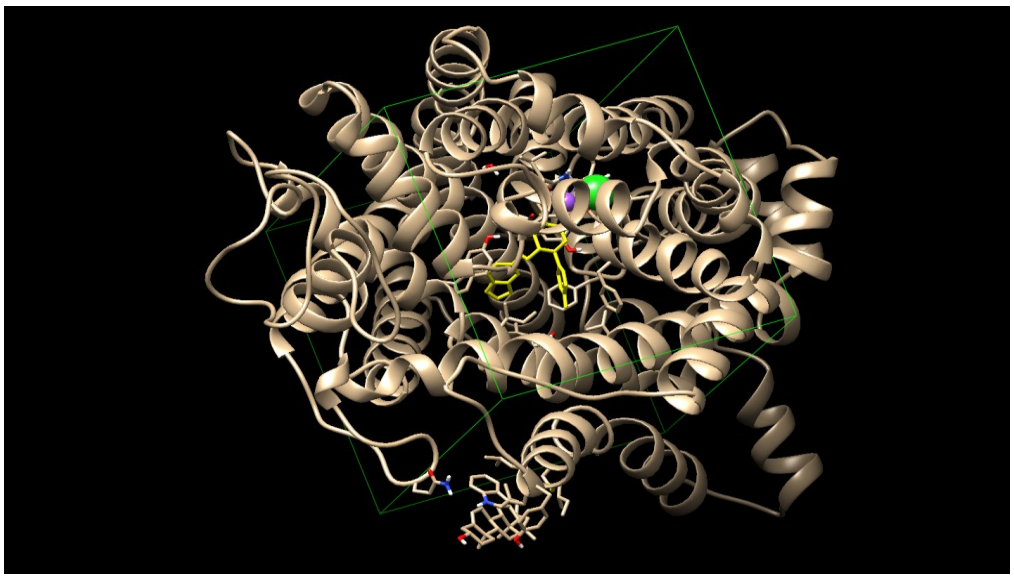


Image 22: Grid Box Configuration in known SERT binding pocket.

As for the final docking settings hydrogen atoms were placed in the structure. All the other options were set to false both in receptor and in ligand. This means that polar hydrogen atoms were retained in order to find possible hydrogen bonding interactions. Furthermore, water molecules and ions were kept in order to study their behaviour upon docking and whether they played a critical role in ligand-binding. Non-standard residues were also kept for structural integrity. In the advanced section, binding modes were set at 10 in order for AutoDock Vina to generate ten poses per ligand. Exhaustiveness of search was set at 32 manually in Open Babel to increase the probability to find the optimal pose for a given ligand. Last option referred to the maximum energy difference between the best and the last pose and this was set to 4 kcal/mol manually, in order unfavourable conformations with low binding affinity as an absolute value to be excluded (Image 23).

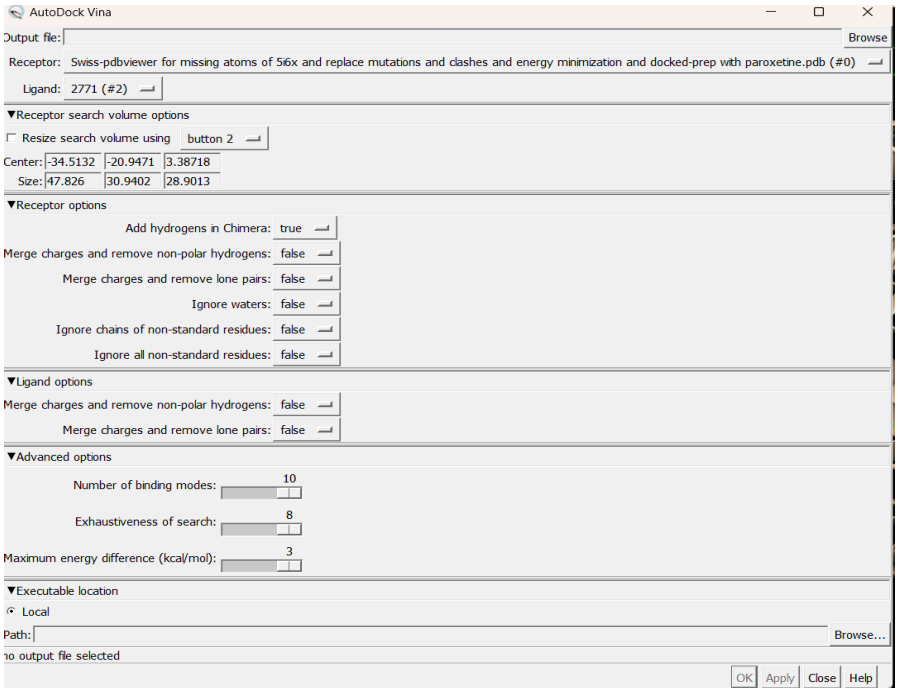


Image 23: Final Docking Settings for set up.

3.2.5 Batch Docking with Autodock Vina

To automate[1] the docking process of these 74 ligands **"***Vina_linux.pl"* command was applied. This command initially reads the receptor.pdbqt file from Chimera. Then it iterates over the *"ligand.txt"* which contains all pdbqt files from ligands and it adjusts its search based on the *"conf.txt"* file that includes the dimensions and size of the grid box, exhaustiveness, energy range, seed and number of poses. Finally, execution was made and all results were gathered with the command "*tail -n 11 *.log > results.txt"* which shows best poses and their respective binding affinities and RMSD values.

3.3 Docking Pose Evaluation and Selection

Autodock Vina generated the best ten poses for each ligand with their binding affinity and RMSD values. The former is proportional to the magnitude of ligand binding with SERT protein and the latter is separated into upper and lower bound. However, in the present study a scoring script-based approach was applied in Python in order to identify and classify the best 5 poses out of ten. This method outweights other alternatives, because of the fact that it counts in not only the binding affinity,

[1] Appendix, page 135-136

but also the RMSD values. This combination added value and integrity in the final results. Since the three variables were on a different scale, normalization was done based on the following formulas:

For normalized binding affinity: $\frac{A-Amax}{Amin-Amax}$

, where A is the binding affinity for a specific pose of a given ligand and Amin and Amax the minimum and maximum affinity among the 5 poses of the same ligand

For normalized RMSD lower value: $\frac{Rmax-R}{Rmax-Rmin}$

, where R is the RMSD value for a specific pose of a given ligand and Rmin and Rmax the minimum and maximum RMSD among the 5 poses of the same ligand. Same formula was followed for RMSD upper value. This formula has opposite signs in comparison to normalized binding affinity formula, because lower RMSD indicates closer geometrical structure with reference pose, namely the first one.

In summary, the formula used in the script was:

Score = 1/3(Normalized Affinity + Normalized RMSD lower value + Normalized RMSD Upper value)

By using this function, each parameter from the 3 contributed equally to the final score, where higher score was linked to better and more favourable pose. First 5 pose were selected for further post docking analysis and feature extraction in BIOVIA Discovery Studio.

3.4 Interaction Analysis and Feature Extraction using BIOVIA Discovery Studio Visualizer

Once the top 5 poses for each ligand were chosen as it is described in Chapter 3.3, namely 370 different poses, next step was to conduct deeper interaction analysis by using the BIOVIA Discovery Studio Visualizer software. This computational tool enables users to have comprehensive insights into the protein-ligand interactions and bonds by providing several features and descriptors. The procedure

for this task was simple. Firstly, the PDB structure of the complex SERT-ligand was loaded, then receptor and ligand were set and the 2D diagram with the interactions included as depicted in image 24 were generated. As shown in figure, different types of bonds and interactions are represented, such as Pi-Sigma bonds, Pi-Pi T-shaped, Amide-Pi Stacked, Alkyl, Pi-Alkyl, hydrogen bonds, hydrophobic contacts, van der Waals interactions, π-π stacking, and other relevant non-covalent forces. BIOVIA offers also the ability to calculate specific angles and distances between certain atoms of amino acids. In addition, it highlights the most critical residues for binding. This process was repeated for all the 370 poses of the 74 distinct ligands.

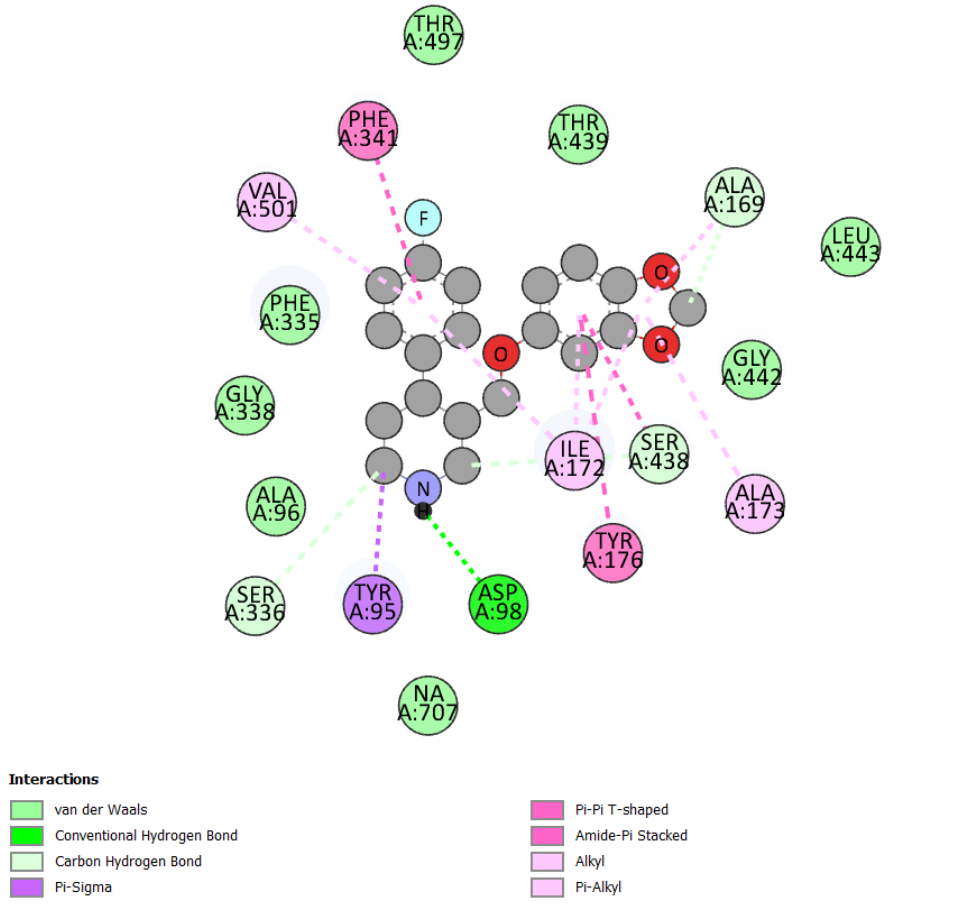


Image 24: Paroxetine-SERT 2D diagram pose 1.

3.4.1 Interaction analysis features and molecular characteristics[2]

[2] Appendix, page 137-138 (Image 57)

A curated dataset was constructed to gather all essential features needed before starting the preprocessing steps in Python in order to build the supervised machine learning model. These features derived from several sources:

- Chimera and AutoDock Vina: *"Binding Affinity (kcal/mol)"*, *"RMSD Upper value"* and *"RMSD Lower value"*. The first one refers to the free energy required for a ligand to bind with SERT protein. The more negative values indicate stronger interaction. RMSD measures the conformational deviation between a specific pose and the optimal pose with the lowest binding energy.
- PubChem Database: *"logP"*, *"Molecular_Weight"* and *"Polar_Surface_Area"* (PSA): LogP is an indicator of hydrophobicity for a molecule and it is the logarithm of the partition coefficient between octanol and water. Molecular weight is the mass of the ligand in Daltons. PSA shows the capacity of polar atoms in the structure, such as oxygen, nitrogen etc.
- BIOVIA Discovery Studio Visualizer tool: All the rest features except for *"Ligand_Distance_to_Grid_Box_Center"*. *"Molecular_Volume*" is proportional to the capacity of the whole ligand in 3D space. *"Surface_Area"* is the part of the ligand that can have access with the solvent and it is measured in Å² (square angstroms). "*NAME"* and "*CATEGORY"*, "*INTERACTION TYPES", "FROM", "FROM CHEMISTRY", "TO", "TO CHEMISTRY"* columns show which atoms were involved, in what kind of interactions (hydrogen bonding, hydrophobic, electrostatic and more) and the direction of electrons and protons. "*Hydrogen Bonds count"*, "*Hydrophobic Bonds count"*, "*Van_Der_Waals_Interaction_Count"*, "*Other types interaction count"* revealed the respective number of interactions between ligand and the receptor. In the other bonds, less common interactions, like halogen reactions, salt bridges and electrostatic interactions were included. More specifically, in hydrogen bonds a hydrogen donor sends the H to a hydrogen acceptor (like O, N). This happens because of the electronegativity of the atoms, where more electronegative atoms in a molecule tend to attract more electrons. Typically these bonds are moderate to strong and are vital for binding. On the other hand, the hydrophobic contacts are caused when non-polar parts of the ligand come into close proximity with the non-polar residues of SERT protein, typically involving alkyl or aromatic groups. Non polar atoms do not carry any electrical dipoles momentarily, meaning that non polar molecules have atoms with similar electronegativity. Hydrophobic bonds stabilize the binding process since it depends on the size of the receptor and the ligand. Van der Waals interactions, which are weak, are provoked by the dipoles that are created because of the redistribution of electrons. They act on atoms that are in close distance (~3–4 Å). For the other bonds, electrostatic interactions are generated due to the attraction of a positively charged atom with a negatively charged atom. They can act over longer ranges by forming salt bridges. In the excel file there were also "*Mean_Hydrogen_Bond_Length_Distance"* and

"*Mean_Hydrophobic_Bond_Length_Distance*" columns that calculated the average distance of all hydrogen and hydrophobic bonds that were included in a specific pose. "*CLOSEST ATOM DISTANCE"* feature is a metric distance for the shortest atom to SERT protein. Apart from these classical descriptors, several angular details that capture spatial geometry and orientation were extracted. These included:

- *"Angle DHA"* (Donor hydrogen acceptor): It calculates the angle between the donor hydrogen atom and the acceptor hydrogen atom. Values closer to 180° are highly connected with strong interactions.
- "*Angles HAY"* (Hydrogen acceptor Y) and *"DAY"* (Donor acceptor Y): These angles assess how the acceptor aligns not only with the donor and hydrogen atom but also with adjacent environment. They reveal possible steric effects due to the overall geometry.
- "*Angle XDA"* (X donor acceptor): This feature is correlated to the donor's orientation.
- *"THETA"* and *"THETA 2"*: These are torsional angles that are formed in a Pi-orbital system with the involvement of aromatic rings.
- *"GAMMA"*: This angle is configured solely by the interaction of Pi systems.
- *"ANGLE DEVIATION"*: Quantifies the deviation of the actual value of an angle with the ideal to prevent steric hindrances.
- *"Ligand_Distance_to_Grid_Box_Center"*: This molecular descriptor quantifies the position of each ligand based on the predefined grid box center (Table 5). It was calculated as the Euclidean distance between the centroid of the ligand which was measured by its three-dimensional atomic coordinates and the stable grid box. To automate this analysis, a custom Python script was designed. This metric was very informative for docking simulations, since it offered an accurate spatial orientation of each ligand in comparison to a known binding pocket as the initial PDB structure with Paroxetine revealed. All docking poses from all ligands were uploaded and the script extracted the atomic coordinates from *HEATM* entries and *UNL* which represent the different ligands of the SERT-ligand complex.

3.4.2 Residue-Level Interaction Features

After molecular interaction details, another sheet[3] in Excel file was constructed that contained residue-level information for the 370 poses. This file added high value to the present thesis as it highlighted critical residues for strong, moderate and non-binders with SERT protein. Furthermore, it delved into the kind of interaction each residue participated in, either in hydrogen, hydrophobic, van der Waals, or other interactions and their frequency of occurrence. Residues gathered through BIOVIA tool among

[3] Appendix, page 137 (Image 56)

these 74 ligands and 370 poses, were 68 which constituted the columns of the second excel file (Table 6). For samples there were the 370 poses (the best 5 for each ligand).

ID	Residue Name	Residue Number
1	TYR	95
2	ALA	96
3	VAL	97
4	ASP	98
5	LEU	99
6	GLY	100
7	ASN	101
8	TRP	103
9	ARG	104
10	TYR	107
11	ILE	108
12	ALA	110
13	GLN	111
14	ASN	112
15	GLY	113
16	GLY	114
17	ALA	169
18	PHE	170
19	ILE	172
20	ALA	173
21	TYR	175
22	TYR	176
23	ASN	177
24	ILE	179
25	PHE	263
26	PHE	311
27	LYS	314
28	PRO	315
29	ASN	316
30	LYS	319
31	ILE	327
32	ASP	328
33	ALA	331
34	GLN	332
35	PHE	334
36	PHE	335
37	SER	336
38	LEU	337
39	GLY	338
40	PHE	341
41	VAL	343
42	ASN	368
43	ASP	400

44	ALA	401
45	PRO	403
46	LEU	406
47	PHE	407
48	SER	438
49	THR	439
50	GLY	442
51	LEU	443
52	VAL	446
53	LYS	490
54	GLU	493
55	GLU	494
56	TYR	495
57	THR	497
58	GLY	498
59	PRO	499
60	VAL	501
61	SER	555
62	PHE	556
63	SER	559
64	PRO	560
65	PRO	561
66	GLN	562
67	LEU	563
68	ARG	564

Table 6: Residues included in the 370 poses of dataset.

The residue-level interactions were encoded as a 4-component vector. More specifically, in the example of Paroxetine pose 1 (Image 24) for *"TYR95"* residue there was only one hydrophobic bond and zero in the other 3 categories (hydrogen, van der Waals, Other interactions), so the final vector will be →

0,1,0,0

, while *"ALA169"* participated in one hydrogen and one hydrophobic bond so the final vector for Paroxetine pose 1 and *"ALA169"* residue will be →

1,1,0,0

Same procedure was followed for all 68 residues and 370 poses in the second excel sheet file. An important note to all this process, is the fact that in the hydrogen interactions category, conventional hydrogen bond, carbon hydrogen bonds, hydrogen-halogen bonds and hydrogen-electrostatic bonds were included. As for hydrophobic bonds, the interactions included are those that contain at least one Pi system or aromatic ring system or alkyl. Van der Waals bonds were appearing as circles with the green colour (Image 24). Other bonds included halogen bonds, salt bridges, electrostatic bonds and Pi-

sulfur. First component matches hydrogen bonds, second hydrophobic, third van der Waals and last the other category of bonds.

3.5 Data Preprocessing and Machine Learning Pipeline

3.5.1 Ligand labelling and final dataset

To build the supervised classification model that will predict the affinity of a ligand with SERT protein transporter, each ligand in the dataset was assigned a binding class label based on its experimentally measured inhibition constant (Table 3 and Table 4). Given the variability in assay conditions and the limited size of the dataset, a slightly wider threshold range was adjusted to ensure sufficient data points across all three classes. For strong binders with SERT the limits were Ki values smaller than 5nM, for moderate binders the limits were placed between 5 nM and 1000 nM and above 1000 nM or not known values no binding labels were attributed. In other words, in first category the majority were known drugs with high potency with SERT, mainly SSRI's, in the moderate binders there were several antidepressant drugs and a few others showing mid-affinity behaviour and finally the non-binders were those with minimal or zero inhibition of 5-HT receptor, like ibuprofen, aspirin etc.

To prepare the final dataset for the classification task, first step was to merge the 2 excel sheets in one. First source is named *"Molecular Descriptors"* and contained all interaction details derived from Autodock Vina, Chimera, PubChem and BIOVIA Discovery Studio. The second file, named "Residue-Analysis", on the other hand included all specific residue information in a format *"H"*, *"HYD"*, *"VDW"*, *"OTHER"* that showed the number of hydrogen, hydrophobic, van der Waals and other bonds for each of the 68 residues respectively. From the first file, *"NAME"*, *"DISTANCE"*, *"CATEGORY"*, *"INTERACTION TYPES"*, *"FROM"*, *"FROM CHEMISTRY"*, *"TO"*, *"TO CHEMISTRY"*, *"ANGLE DEVIATION"* columns were removed, since the present thesis did not aim to focus on deep atomic interactions. *"DISTANCE"* column was incorporated to the "*Mean Hydrogen Bond Length Distance*" and "*Mean Hydrophobic Bond Length Distance*" features. Across the same pose the hydrophobic interactions were averaged and constituted the "*Mean Hydrophobic Bond Length Distance*" feature, similarly for hydrogen bonds. *"ANGLE DEVIATION"* was removed since it was appeared in few samples, so there would be many missing values to handle. Next important step was to group the docking poses based on their ligand names and label them accordingly. In addition, missing values in all the interaction file were replaced with 0 since in all cases it had as biological meaning the absence of that particular feature. As for the residue analysis, the excel sheet name is "Residue-Analysis" and the missing values it contained had the

default vector 0,0,0,0 meaning absence of all the 4 categories of bonds. These 2 sheets were then merged and the resulting dataset is shown as snapshots in images 25 and 26 below. All vectors were then split into four separate numerical columns. All this process resulted in a dataset with shape (370,301), namely 370 samples or poses and 301 features. An important note in this part was that during the merging of the 2 excel sheets, the mean values of all angular metrics were calculated to yield a single representative value per pose for consistency and later aggregation.

index	Ligand Distance to Grid Box Center	Molecular Volume	Surface Area	Binding Affinity(kcal/mole)	RMSD Upper value	RMSD Lower value	Molecular Weight	logP	Polar Surface Area	Hydrogen Bonds count
std	2.238444736423042	68.15729972229299	71.67838871146613	1.1747581348629363	2.6437319886570938	1.4570532674038215	87.32188820711792	1.2731853278009875	30.523344307569715	2.008653366532777
min	0.4266718434233948	111.653075	152.341	-11.08	0.0	0.0	129.16	-1.3	3.2	0.0
mean	2.956945118417455	262.95790257567563	337.67057567567565	-8.209113513513513	3.640956756756757	2.030544864864865	313.24256756756756	3.1986486486486485	46.456756756756754	2.6
max	16.50690099709 86	487.927142	566.164	-4.749	13.07	8.765	654.6	5.8	121.0	11.0
count	370.0	370.0	370.0	370.0	370.0	370.0	370.0	370.0	370.0	370.0
75%	3.3565134251201396	315.2767805	380.12125000000003	-7.402	5.5035	2.76025	379.5	4.2	68.6	4.0
50%	2.387997532742529	248.564842	318.7495	-8.229	3.3819999999999997	1.9769999999999999	293.29999999999995	3.25	42.8	2.0
25%	1.769986468449996	219.84768275	291.817	-8.96025	1.87875	1.2774999999999999	255.35	2.6	19.0	1.0

Image 25: Merged Dataset with interaction details and residue analysis 1st image.

Hydrophobic bonds count	Van_Der_Waals_Interaction_Count	Other types interaction count	ANGLE DHA	ANGLE HAY	THETA	THETA 2	GAMMA	CLOSEST ATOM DISTANCE	ANGLE XDA
2.690131055874 55	2.6805573724229883	0.8593457164763683	63.622681795294625	58.42578418166932	17.496636351307906	24.615891068967 45	26.941451564169025	1.4239265284430658	56.696362735669915
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
5.424324324324324	9.143243243243242	0.4972972972972973	85.44509765444015	77.96708264800515	26.982727403474907	36.801404954954954	36.223558855855856	3.007947837837838	69.34935268339768
17.0	18.0	6.0	173.236	175.083	86.57	89.56	89.129	4.342	165.056
370.0	370.0	370.0	370.0	370.0	370.0	370.0	370.0	370.0	370.0
7.0	11.0	1.0	135.24303125	122.44112500000001	35.385125	52.896166666666666	59.286249999999995	3.7446875	113.03030714285714
5.0	9.0	0.0	119.515	106.04725	24.762999999999998	40.35516666666666	35.0925	3.6135	98.18241666666665
3.0	7.0	0.0	0.0	0.0	17.009	17.560499999999998	12.26625	3.46325	0.0

Image 26: Merged Dataset with interaction details and residue analysis 2nd image.

3.5.2 Aggregation of Docking Poses and Final Dataset Configuration

Since for each ligand there was variation in the features across the 5 poses, an alternative approach was applied and that was the aggregation per ligand in order to minimize this variation for the present classification task. More specifically, the group of ligands was achieved through the isolation of the basic names from the full names of pose identifiers (e.g., "Paroxetine with Sert dock pose 1") that was isolated into Paroxetine. The features were separated into 4 main categories:

- Descriptor_columns: *"Ligand_Distance_to_Grid_Box_Center", "Molecular_Volume", "Surface_Area", "Binding_Affinity(kcal/mole)", "logP", "Polar_Surface_Area", "Mean_Hydrogen_Bond_Length_Distance", "Mean Hydrophobic_Bond_Length_Distance", "ANGLE_DHA", "ANGLE_HAY", "THETA", "THETA 2", "GAMMA", "CLOSEST_ATOM*

DISTANCE", "ANGLE_XDA", "ANGLE_DAY". For these features, the median values were kept across the 5 poses, since they are not sensitive to outliers and also in other metrics, such as the mean values, the model outputs lower accuracy. *"Molecular_Weight"* was deleted since it was similar to *"Molecular_Volume"* and it did not add extra biological essence to the model.

- Interaction_columns: *"Hydrogen_Bonds_count", "Hydrophobic_bonds_count", "Van_Der_Waals_Interaction_Count", "Other types interaction count"*. For these columns, the sum of values was used in order to obtain a cumulative interaction profile for each ligand.
- Residue_columns: These included all columns that contained a format like *"H"*, *"HYD"*, *"VDW"*, *"OTHER"* which is described in section 3.5.1 and the sum values were used across 5 poses in order to involve the frequency factor in the model.
- Target variable: This is the column "*Labelled_class"* which represents the binding class for all ligands. In order to be manageable for the model to recognize this variable, it was encoded into 3 numerical values, 0 was mapped to *"MODERATE BINDING"*, 1 was mapped to *"NO BINDING"* and 2 to *"STRONGLY BINDING*". This was achieved by using the label encoder technique in scikit-learn , which is a free and open-source machine learning library for the Python programming language.

After these steps the aggregated dataset had a shape of (74,293), where 74 were the ligands-samples and 293 the features. Then, the constant features that showed no variance between ligands were deleted since they did not contribute to the training process of the model. The remaining features were 160.

3.5.3 Correlation and statistical analysis of dataset

Statistical analysis was conducted on the refined dataset (74,160) for better understanding of the variance across features and their curve distributions. Mean, min, max and standard deviation were calculated for each column and the correlation matrix was computed between the top 30 features and the target variable, either with positive or negative correlation. To visualize all these, pair-plots were created between features and labelled class. To avoid redundancy and multicollinearity in the dataset, interaction columns (e.g., *"Hydrogen_Bonds_count", "Hydrophobic_bonds_count", "Van_Der_Waals_Interaction_Count", "Other types interaction count"*) were compared against the sum of their corresponding residue-level interactions (e.g., *"TYR95_H"*, *"TYR95_HYD"*, *"TYR95_VDW"*, *"TYR95_OTHER"*) and it was confirmed that they were highly correlated, so they were removed from the dataset. This led to a more optimal version of 156 features instead of previous 160. Below in image 27 the Pearson correlation coefficients among features were calculated in order

to identify potential multicollinearity. Values closer to 1 indicated strong positive correlation, while values closer to -1 negative correlation. As illustrated, there were 3 pairs of features that showed high correlations (above 0.80 threshold). These were:

- *"ANGLE_XDA"* with *"ANGLE_DAY"* with 0.98 correlation
- *"ANGLE_HAY"* with *"ANGLE_DHA"* with 0.97 correlation
- *"Surface_Area"* with *"Molecular_Volume"* with 0.89 correlation

All other features were below this threshold. This was very important, especially during training of models, because algorithms like Logistic Regression and Support Vector Machines cannot handle multicollinearity, while tree-based algorithms, such as Random Forest and XGBoost are robust to such correlations.

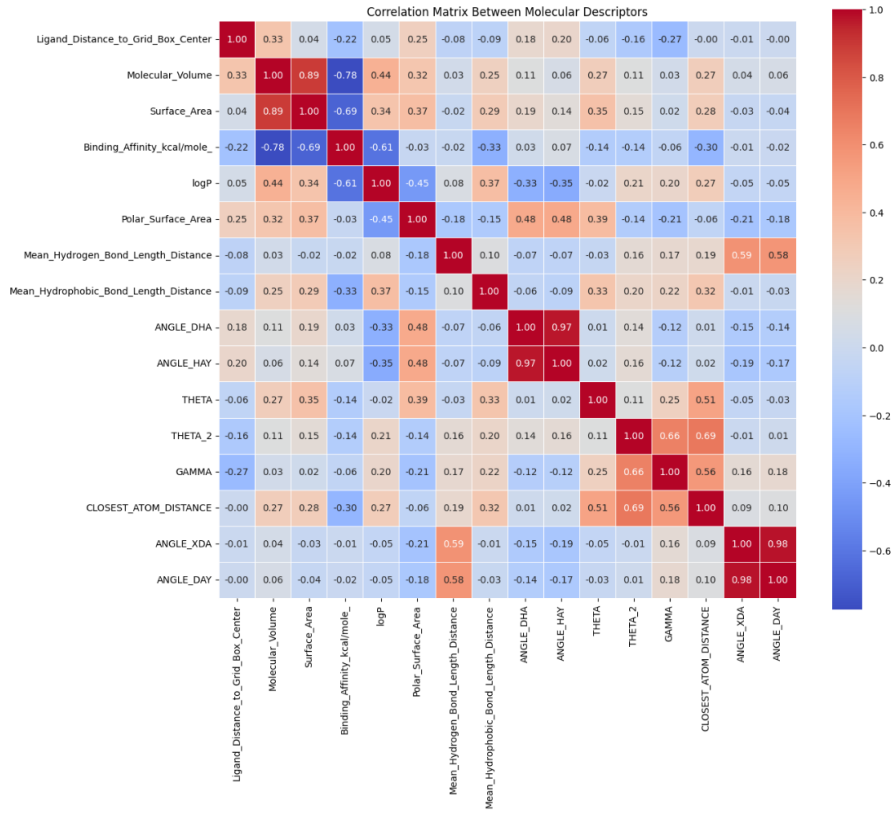


Image 27: Pearson correlation coefficients among features of dataset.

3.5.4 Model training[4]

The correlation analysis was then followed by the implementation of the model training. Due to the limited size of dataset, several methods were applied to handle this situation. The majority of them showed overfitting with moderate to low promising results. However, nested cross validation technique performed moderately and at the same time it minimized the risk of overfitting. It consists of two loops, the outer loop and the inner loop. The former splits the entire dataset into k training and

[4] Appendix, page 138-140

test folds and predicts the evaluation metrics, while the latter splits further the training set into n folds in order to find the optimal features and tune the hyperparameters for each model. Nested cross validation offers several advantages that are commented below (Varma & Simon, 2006), (Varoquaux et al., 2017).

- Decreased risk for overfitting, since model and feature selection arise solely from training set
- Prevention of data leakage, since the evaluation of model is done prior to the model selection which is achieved in the inner loop.
- High generalizability, because model selection and parameters are selected from different folds in comparison to final evaluation. This leads to reliable performance in unseen data, which is the final scope of the present thesis.

More specifically, a 5-Fold stratified cross validation was applied in the outer loop to test the model's performance. With this way, the dataset was split into 5 parts, 4 for training and 1 for test which calculated precision, accuracy, recall, F1-score and ROC curve. For the inner loop 3-Fold stratified cross validation was applied in order to extract the most predictive features and tune the parameters of the models trained. The process of feature extraction started from selecting the best 30 features out of 155 by using feature importance scores on the respective model (meaning XGBoost with xgboost feature importance, random forest with random forest feature importance, etc. and absolute coefficients for Logistic Regression and Support Vector Machines) and then Recursive Feature Elimination (RFE) followed. RFE was applied on the 30 features starting from 15 and proceeding with step 5 until 30 to search for the optimal number and subset of features. At each number of features, GridSearchCV was integrated to find the best hyperparameters and subset of characteristics based on the f1 macro score. It surpasses over other approaches, such as RandomizedSearchCV, because it searches exhaustively all the possible combinations of parameters and different subsets of features, though it required high computational cost and time. RFE was not applied in LightGBM algorithm, because it required high computational time, but remaining training process was similar to the other models. Lastly, the frequency of features across the 5 folds was printed in addition with the robust ones which referred to the features that were present in 3 or more folds in the nested structure. These robust features were used later in the explainable section.

Several classifier algorithms were trained and evaluated in the present study. Initially, XGBoost is a scalable and fast gradient boosted classifier. It acts by creating an ensemble of decision trees sequentially, where each new trees aim to minimize the errors made by the previous ones. Regularization (both L1 and L2) techniques were incorporated, which were useful for reducing potential overfitting and improving generalization (Chen & Guestrin, 2016). The parameters that were set for this classifier were the following:

Grid parameters for XGBoost	Possible values in GridsearchCV
Max depth	4
Learning rate	0.05, 0.1
Reg alpha (L1 Regularization)	1, 2
Reg lamda (L2 Regularization)	4, 5
Gamma	0.5, 0.8
Subsample	0.7
Colsample bytree	0.7

Table 7: Grid parameters for XGBoost classifier.

Default value of max depth is 3 and was increased it by one, in order to capture more complex patterns. Learning rate is proportional to how quickly will the model converge to a final decision. The difficulty lies in the fact that the model may reach a local minimum instead of a global minimum which results in poorer performance. However, lower rates are better for generalization. Regularization parameters L1 and L2 act by reducing the models complexity and alleviating the overfitting that usually occurs in small and high dimensional datasets. Gamma parameter determines the splits of the tree nodes required to improve the accuracy of the model. Last 2 parameters involve the fraction of samples and features respectively that are sampled for each tree and both are defined to minimize the risk for overfitting.

Another algorithm that was applied is the Random Forest Classifier (RF), which is an ensemble learning method that incorporates several decision trees. Each tree is trained on a different subset of training samples and features according to the Bootstrapped Sampling technique and feature subsampling respectively. With this way, the randomness in the RF model is enhanced, dragging equally the generalization at the same time. The operation of RF is depicted in Image 27. The final output of the model lies on the majority voting of the different trees, so in the following example the model would predict class c. This aggregated mechanism ensures robustness especially when the dataset has few samples and many features as in the present thesis, because the variance across trees is minimized (Scikit-learn, 2009).

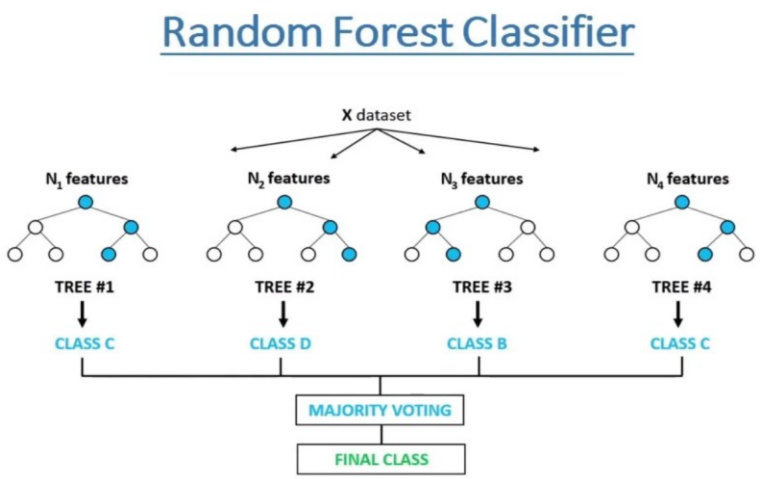


Image 28: Structure of Random Forest Classifier (Chauhan, A., 2021).

Random Forest offers several advantages and drawbacks. It is user friendly and easy to adjust, because it does not require standard scaling, since it consists of several decision trees that are split by features and not by other distance metrics or thresholds. Another positive aspect of RF is the risk of overfitting which is much less in comparison to single decision tree classifier or other traditional algorithms. It also enables feature importance analysis in order to find the most predictive features for the classification task. Last but not least, it is widely used in the scientific community for several purposes, like healthcare, finance etc. On the other hand, an issue in RF classifier is the high computational time needed to process all the number of trees. The higher the number of trees is, the longer the run time will be. In addition, the structure is more complex than that of a decision tree, so careful tuning of hyperparameters is essential (Coursera, 2024).

As now for the hyperparameters, these included:

Grid parameters for Random Forest	Possible values in Random Forest
Max depth	5
Minimum samples split	25
Minimum samples leaf	5
Max features	‘log2’,‘sqrt’
Class Weight	‘balanced’

Table 8: Grid parameters for Random Forest classifier.

Max depth value was set at 5 (default value is ‘None’) in order to capture more complex patterns, but it was safe to handle the risk of overfitting. Minimum samples split and leaf were much higher than the default values (2 and 1 respectively). This was a logical option, because having more samples before splitting a node and more in leaf, increases the validity of results. Max features parameter added randomness in the model since it determines the size of available features at every split. Finally,

class weight parameter was responsible for alleviating the small class imbalance that existed in the dataset.

A more sophisticated algorithm employed in this thesis was the Light Gradient Boosting Machine (LightGBM) which is an advanced gradient boosting algorithm that utilizes lower memory while at the same time it surpasses in speed and accuracy in comparison to other algorithms. Based on LightGBM's documentation the unique structure is based on the leaf-wise (best-first) tree growth. This means that LightGBM expands in a leaf level that minimizes the error, regardless of the layer level (Image 29). For example in XGBoost classifier with default parameters, all nodes from the same depth should expand firstly and then the model goes deeper (LightGBM 4.6.0 documentation, 2017).

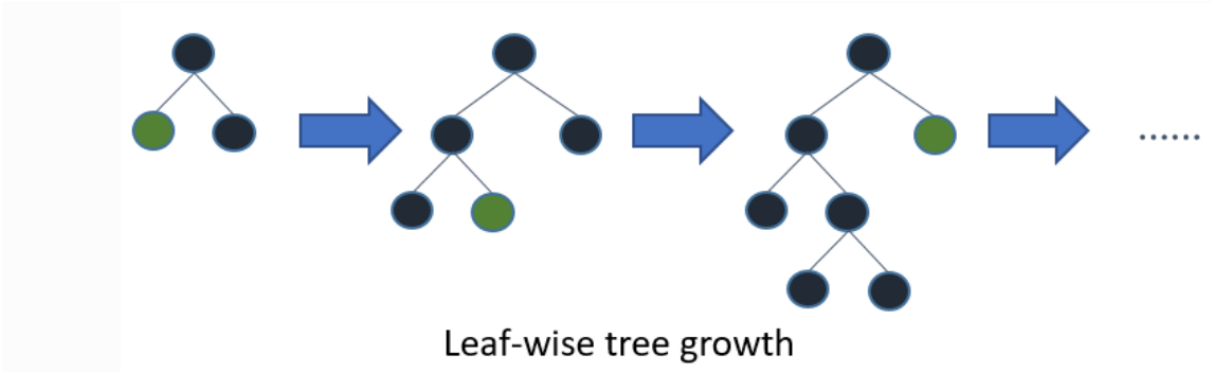


Image 29: Leaf-wise tree growth strategy in LightGBM classifier (LightGBM 4.6.0 documentation., 2017).

In parallel to the leaf-wise method, another approach is applied, called histogram-based learning. This method discretizes the continuous variables into fixed bins, saving computational time and reducing model complexity. Then, it decides the optimal place to split the leaf. This process is accelerated through the histogram subtraction, which does not reconstruct the histogram for every node, instead it is derived from the parent nodes (LightGBM 4.6.0 documentation, 2017).

As for the hyperparameters in LightGBM learning rate was tested into 3 possible values, mainly focus on a smoother learning process instead of a more aggressive approach, because the dataset is small and generalization was the final goal of the present thesis. Number of leaves per tree were tested in values 15 or 31 (31 is the default value), because more leaves usually drive the model to learn more complex patterns of the dataset. However, in this case overfitting was a possible threat due to the limited size. Similar to number of leaves is the minimum number of samples per node, where more balanced values were tuned. L1 and L2 regularization were applied to decrease the probability of overfitting (Table 9).

Grid parameters for LightGBM	Possible values in LightGBM
Learning rate	0.01,0.05,0.1
Number of Leaves	15,31
Min child samples	10,20,30

Reg alpha	3,4,5
Reg lamda	4,5

Table 9: Grid parameters for LightGBM classifier.

In conclusion, based on LightGBM's documentation LightGBM offers as advantages the high speed in process without consuming much memory units with the aid of histogram-based learning. It ensures a higher accuracy due to the leaf-wise method and lastly it does not require encoding categorical variables as other algorithms. Nevertheless, overfitting is considered a realistic situation, because it contains several hyperparamaters that need to be tuned effectively. Furthermore, it lacks in interpretability because of the complex structure it has (LightGBM 4.6.0 documentation, 2017).

Logistic regression was also applied in this thesis to test its performance. Logistic Regression classifier (LR) is a simpler structured linear algorithm that handles both binary and multiclass problems. Main formula behind LR is the sigmoid function, which maps the predicted values that arise from the combination of features into probabilities (Image 30). These probabilities are then translated into values between 0 and 1, depending on where they are closer. This is clear especially in binary categorical cased where 1 is "Yes" and 0 is "No" (GeeksforGeeks, 2024).

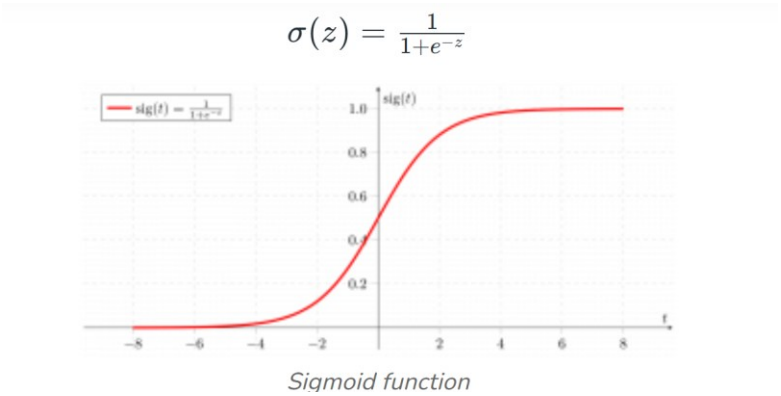


Image 30: Sigmoid function of Logistic Regression (GeeksforGeeks., 2017).

However, in multiclass tasks like in the present thesis, scikit learn supports the One-vs-Rest (OvR) strategy by default, where a separate binary classifier is trained for each class against the rest (Scikit-learn, 2014).

For the hyperparameters tuning process (Table 10), increased regularization strength was introduced to manipulate overfitting. In addition, 2 types l1 and l2 of regularization were tested and the solver is the algorithm used to minimize the loss function. In this case liblinear solver supports both penalty types and can be implemented in multiclass problems (Scikit-learn, 2014).

Grid parameters for Logistic Regression	Possible values in Logistic Regression
Regularization Strength (C)	0.0001 – 100 (log-spaced)

Penalty	l1, l2
Solver	liblinear

Table 10: Grid Parameters for Logistic Regression.

For further analysis, LR can be useful in cases where the dataset is simple and the features are linearly separated, though in real world scenarios these are not linearly distributed. This means that, it cannot capture any complex patterns within the dataset. Overfitting occurs regularly, especially in small sample size and high dimensional datasets (like in this thesis). It is a very fast algorithm and very interpretable to identify critical features that define the class label either in binary or in multiclass models again with the major assumption that the correlation of independent variables and target variable is linear (GeeksforGeeks, 2023).

In addition, Support Vector Machine Classifier (SVM) was also employed in this thesis. When data points are non-linearly separable, the main mechanism behind SVM is that it maps the input vectors into higher dimensional feature space, called "hyperplane" in order to separate the classes with the largest possible margin (Image 31). This is achieved through some non-linear functions which are called "kernel" functions and affect the decision boundaries. The points closer to these margins are the support vectors (Cortes & Vapnik, 1995).

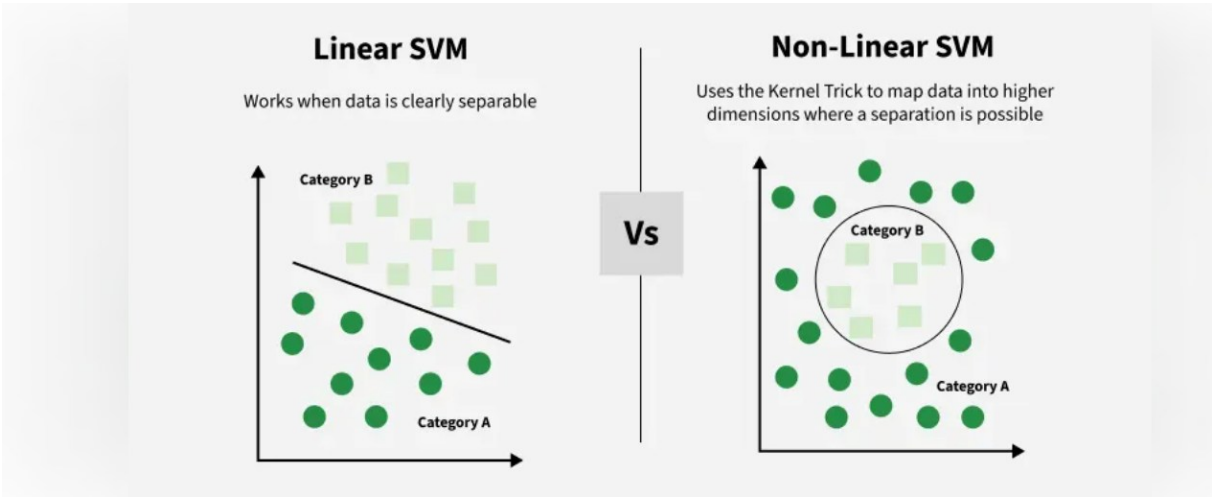


Image 31: Mechanism of Support Vector Machines Classifier (GeeksforGeeks., 2021).

On the other hand, if features are linearly correlated, then the formula behind the structure of SVM is: $F(x) = w \cdot x + b$, where w is the weight vector and defines the orientation of the hyperplane and b is the bias term. Since the goal is to increase the margin, $2 \div \|w\|$ should be maximized (Cortes & Vapnik, 1995).

The hyperparameters optimized for SVM are the following:

Grid parameters for Support Vector Machines	Possible values in Support Vector Machines
Regularization Strength (C)	0.01, 0.1, 1, 10
kernel	'linear', 'rbf'
gamma	'scale', 'auto'

Table 11: Grid Parameters for SVM.

Regularization strength is inversely proportional to C parameter, where lower C values allow wider margins. Several values were tested for their performance. The kernel types that were tested in SVM were 'linear' and 'rbf' and kernel coefficient is the gamma parameter (Scikit-learn, 2019).

The most important aspect of SVM is that it has the ability to handle high dimensional datasets, either linearly or non-linearly separable. The outliers are easily excluded with the accurate adaptation of the margins, ensuring robustness in the model. Finally, it supports multiclassification problems without consuming too much memory in the system. As for the limitations of SVM, it lacks in interpretability, because its structure of hyperplane is difficult to understand. Furthermore, right scaling is necessary, otherwise SVM performs poorly in comparison to other algorithms. Also, it cannot handle overlapping classes (GeeksforGeeks, 2021).

Last algorithm that was implemented is the Voting Classifier, an ensemble meta-estimator that combines the predictions of multiple base models with their optimal parameters. This works by averaging the probabilities of each model and for each class. There are 2 different voting strategies. The first, which was applied in present thesis is, the soft voting strategy. Its core function is based on the fact that each model estimates a probability for each of the 3 classes, then the average probability across all models is measured and the voting classifier outputs the class with the highest averaged probability. This means that the confidence intervals for each model are counted in the final prediction. On the other hand, there is the hard voting, where each model outputs a specific class without probabilities and then the ensemble classifier votes based on the majority class (Scikit-learn ; Pedregosa et al., 2011).

Based on the code shell below, the best parameters from Random Forest and XGBoost were used, since these 2 algorithms performed better individually as shown in Chapter 4.

```
VotingClassifier(estimators= [('rf', best_rf), ('xgb', best_xgb)], voting='soft')
```

Soft voting seems to stabilize predictions, because it reduces the variances. In other words, it is a smoother vote that relies on consistency and not on the absolute prediction. This ensemble strategy is grounded in the core philosophy of ensemble learning and that is that the overall is greater than the sum of its parts, because it enables to keep the strengths of each algorithm and minimize their weaknesses. Random Forest is known for its robustness to noise due to bootstrap aggregation, while XGBoost excels in handling complex patterns through gradient-boosted additive trees with regularization. Their complementary nature enhances the effectiveness of ensemble techniques like the

Voting Classifier and promotes a better generalization which was the main goal of the present study (Rokach, 2010).

As for the advantages of this ensemble method, most important one, is that it ensures robustness, since it integrates the structures of 2 base models. Similarly, accuracy and generalizability are improved and there is less risk for overfitting. However, this Voting Classifier lacks in the field of interpretability in comparison to the individual models, due to the more complex structure that includes the combination of other models (Scikit-learn ; Pedregosa et al., 2011).

3.5.5 Model evaluation

All the aforementioned models were assessed using a diverse set of evaluation metrics from the sklearn.metrics module in Scikit-learn (Pedregosa et al., 2011). This variety ensured different perspective of the performance of each model and all metrics together act complementarily. This set includes (Pedregosa et al., 2011):

- **Accuracy:** Shows the number of correct predictions over total number of samples. It is applied across all models, since accuracy is the most common used metric and easy to interpret. However, it is not reliable, because in imbalanced datasets the minority classes may be falsely labelled.
- **Precision:** It represents how many of the positive predicted class are actually positive. For example how many samples that are predicted as class 2 are actually in class 2.This metric reveals the false positive samples and it is more valuable especially in imbalanced and small datasets. The formula that corresponds to precision is ⇨ **TP/ (TP+FP),** where FP are the false positive samples (those that were assigned a specific class but it is actually the wrong class). The goal is to get closer to 1 by reducing the FP term.
- **Recall:** This metric known as sensitivity, is the ratio of the correctly positive predicted cases to all the actual positive samples and the formula is ⇨ **TP/ (TP+FN**), where FN are the false negative samples (those that failed to be assigned in the correct class). Again the goal is to get closer to 1 by reducing the number of FN.
- **F1-Score:** It is attributed as the harmonic mean of precision and recall ⇨ **2*TP/ [(2*TP) + FN+ FP]** and it is the most recognizable and reliable metric since it captures both types of errors. Values closer to 1 indicate better performance with fewer wrong predictions.
- **Confusion Matrix:** This is a matrix that is plotted in order to visualize per class the predicted samples in comparison to the actual labels. It shows which class each model struggles to identify and which samples mislabels.

- **Classification Report:** It aggregates all aforementioned metrics, such as precision, recall and F1-score for each class.
- **ROC Curve and AUC (One-Vs-Rest):** The Receiver Operating Characteristic (ROC) curve is a graphical representation that depicts the true positive rate against the false positive rate. The Area under the Curve (AUC) is proportional to the models ability to identify a positive class among other classes. Values of AUC closer to 1 mean that the model discriminates well, while values closer to 0.5 are in the randomness region. Since in the present thesis there are 3 classes labelled the ROC analysis was performed in combination to the One-Vs-Rest technique. This method converts the multiclassification task into a binary task, where one class is considered the positive case and the other 2 classes the negative. As a result, three individual ROC curves, one per class, were generated. This strategy is very useful to assess a models performance across the 3 classes in the present thesis and it is not influenced by the imbalanced dataset (Pedregosa et al., 2011).

3.5.6 Key Libraries and modules used from Python language

Several Python libraries were utilized by the present thesis for different purposes. “Numpy” and “Pandas” libraries were used for their effectiveness in handling large datasets with a variety of features as in this study. They also facilitated loading CSV or excel files, unzipping them, transforming them, aggregating the 5 poses for each ligand into one single vector and finally filtering the descriptors in order to keep the most meaningful (McKinney, 2010; Harris et al., 2020). Regular expressions (“re”) module enabled matching samples between the 2 files according to their docking pose names and assigning the correct number of bonds for each residue and type of bond (Python Software Foundation, 2009). For visualization analysis**, “**Matplotlib” and **“**Seaborn” were employed. With the aid of these libraries, correlation matrix, pair plots analysis, ROC curves, feature importances plots etc were plotted (Hunter, 2007; Waskom, 2021). For the application of machine learning algorithms (LR, SVC, RF), scaling of data (Standard scaler), hyperpararameter tuning (“StratifiedKFold”, “GridSearchCV”), evaluation metrics that are mentioned in chapter 3.5.5, nested cross validation and selection of a specific number of features through “RFE”, scikit-learn library was responsible (Pedregosa et al., 2011). To handle imbalances in the dataset the imbalanced-learn library was installed (Lemaître et al., 2017). Last but not least, the “counter” module was essential for tracking the most frequent features across the folds in the nested cross validation, therefore the most influential features that added value to a models decision (Python Software Foundation, 2024).

3.6 Explainable AI Tools (XAI)

To further validate results from predefined models and enhance their interpretability for future purposes, 2 additional explainable AI techniques were incorporated: SHAP (SHapley Additive Explanations) and Partial Dependence Plots (PDP). These tools shed light on a models prediction process, meaning that they highlighted the most influential features that determined a models decision in order to classify a ligand into strong binder, moderate binder and non-binder with human SERT protein transporter.

SHAP is a recent AI strategy based on the game theory approach. It calculates the marginal contribution of all possible combinations of features in order to come up with a final prediction. It is easily adapted in tree-based models, such as Random Forest Classifier and XGB Classifier with the aid of SHAP-Tree-Explainer method (Lundberg & Lee, 2017). The SHAP values were computed to the top 2 models from 3.5.5 chapter evaluation metrics. The models with their optimal hyperparameters from their best fold in the nested cross validation technique were retrained on the entire dataset with the robust features obtained from each model. The SHAP summary was plotted for both models, which performed the contribution of each feature for each of the three classes of this classification task, while also a heatmap was printed that matched this contribution into percentages for a clearer view. A comparison was conducted between the 2 models to identify how each model prioritized and ranked each feature, while also a mislabeling analysis was conducted by using SHAP waterfall plots to analyze which features pushed models towards wrong classes (Molnar C., 2025).

On the other hand, Partial Dependence Plots (PDPs) are another interpretability technique that visualizes the effect of a specific feature on a models prediction. In contrast to SHAP, PDPs provide a global perspective on feature–response relationships, enabling users to capture non-linear relationships among features and final predictions. They are especially useful in multiclassification problems, because they represent the marginal values of features that can alter the models behaviour. PDPs were computed in the present thesis for both models and they were compared to SHAP features for further validation. However, the most negative characteristic of PDPs, is the fact that they assume independency across features, so if the correlation between some features is high, misleading results may occur (Molnar C., 2025).

Chapter 4: Results

4.1 Docking Results

Firstly, in order to evaluate whether docking simulation was successful and proceed with the results, a comparison between docked paroxetine pose 1 with the real human SERT (5I6X) structure from RCSB protein data bank was conducted. As the following image depicts, paroxetine from 5I6X structure with yellow colour is parallelized with the docked paroxetine pose 1 with the procedure mentioned in chapter 3 (blue colour). Obviously, there is almost complete alignment, indicating that all the docking regulations and options followed, are consistent with reality to a great extent and enhance the validity of the docking results and thesis (Image 32).

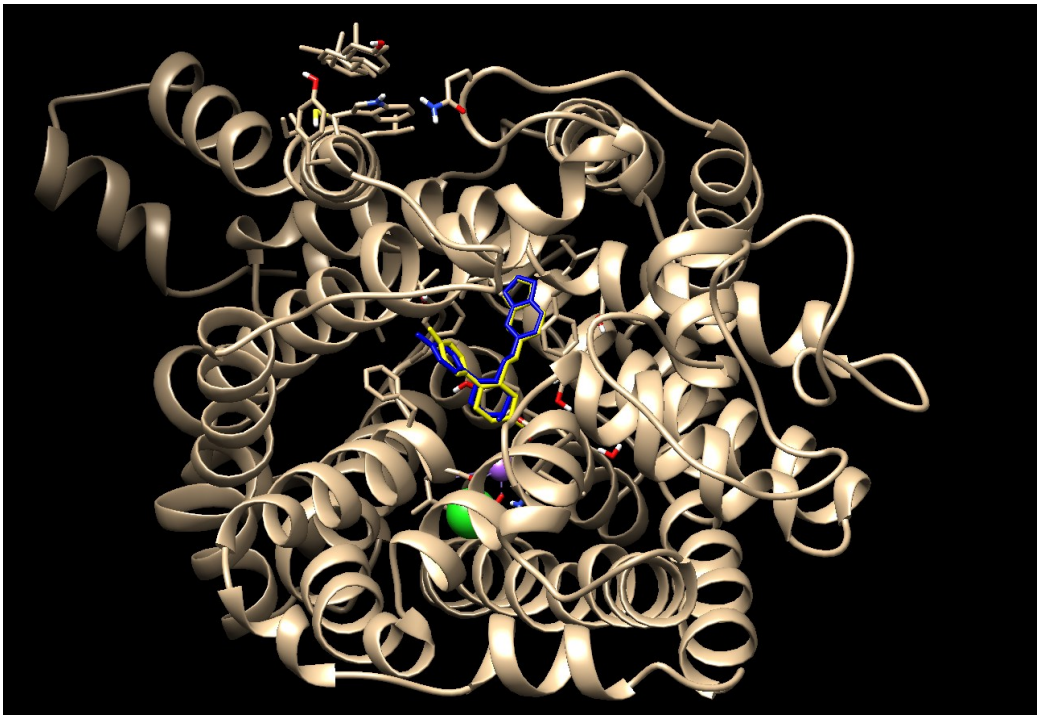


Image 32: Docking of Paroxetine Pose 1 in comparison to 5I6X structure.

As mentioned in Chapter 3.3, ten poses were extracted from Chimera software and AutoDock Vina and through a custom script in Python the best 5 poses were retained based again on the mathematical formulas described in 3.3 chapter. For example, below the results for the ten poses for paroxetine are represented with decreasing values of binding affinities with SERT protein. Third and fourth columns are the RMSD lower and upper bounds respectively with first pose being zero. In table 12 the best 5 were kept from the initial 10 through the Python script. Similar procedure was conducted for all 370 poses and 74 ligands and the best 5 are presented in appendix below[5].

==> PAROXETINE_COMPOUND_CID_43815<== Chimera and Autodock Vina

1	-10.09	0	0
2	-8.475	2.841	3.936

[5] Appendix, page 114-135

3	-8.439	2.912	3.813
4	-8.432	4.73	7.021
5	-8.39	3.148	4.01
6	-8.35	3.472	6.18
7	-8.277	3.496	5.825
8	-8.247	6.714	9.165
9	-8.246	3.741	6.649
10	-8.186	3.72	6.801

Top 5 Poses for PAROXETINE_COMPOUND_CID_43815 through Python custom script:

POSE	BINDING AFFINITY	RMSD LOWER BOUND	RMSD UPPER BOUND	COMPOSITE SCORE
1	-10.090	0.000	0.000	1.000000
2	-8.475	2.841	3.936	0.433060
3	-8.439	2.912	3.813	0.427706
5	-8.390	3.148	4.010	0.400246
6	-8.350	3.472	6.180	0.298234

Table 12: Best 5 poses chosen out of 10 based on Python script.

For the predefined classes mentioned in chapter 3.2.2, the binding affinities distributions were analysed by taking into account only the first pose for each ligand, namely the best one. As anticipated, strong binders exhibit lower median value and generally more negative values across the dataset with the majority of them falling between -8.5 and -9.5 kcal/mol. The interquartile range (IQR) is relatively narrow, however there is obvious overlap between strong and moderate binding classes. As for *"MODERATE BINDING"* class, the IQR is narrower than that of the strong class and the majority of values lie between -8 and -9 kcal/mol with lowest binding affinity around -7 kcal/mol, similar to strong binders. There is a huge overlap between moderate and non-binders and moderate with strong binders, which is a logical behaviour since it represents the intermediate condition. "*NO BINDING*" class shows an irrelevant behaviour in molecular docking with SERT, because the IQR reveals high overlap not only with moderate class but also with the strong class. It falls between -7 kcal/mol and -9 kcal/mol, which are values typical of strong and moderate binders. A distinct behaviour, however, is that "*NO BINDING"* class has several extreme values both on the "*STRONGLY BINDING*" spectrum and in weak binding ranges, like -5 kcal/mol to -6 kcal/mol (Image 33).

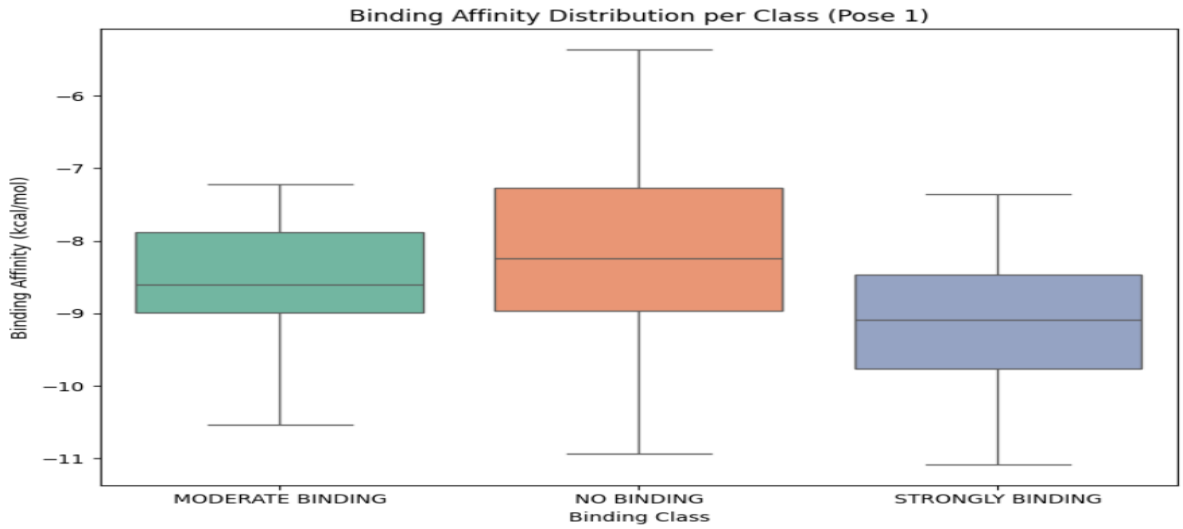


Image 33: Box Plots of Binding Affinity across the 3 classes.

Although the overall pattern reflects a realistic behaviour to some extent and that is for example that strong binders show more negative values, moderate binders show an intermediate step between non-binders and strong binders etc., several limitations exist in class labeling of human SERT inhibitors with "Chimera" and "Autodock Vina" that cannot identify the exact boundaries between these 3 categories. This fact, suggests a need for complementary validation methods beyond docking simulations.

Another useful plot is the Kernel Density Estimate (KDE) plot which estimates the probability density function for binding affinities in the 3 classes. As the following image depicts, in the strong binders, there is a narrow distribution curve with a peak among -9 kcal/mol and -10 kcal/mol. This suggests that strong SERT inhibition in the present dataset shows consistency to a great extent. However, as the box plots reveal (Image 33), the distribution of the moderate SERT binders overlaps with the other 2 classes and peaks around -8.5 kcal/mol proving that it balances between strong and non-binders. Surprisingly again, class *"NO BINDING"* has the main peak around -9 kcal/mol and overlaps significantly with moderate and strong binders. Nevertheless, the distribution is much more flattened than in the other 2 classes and that enhances the perception, that such a classification task with potential SERT inhibitors cannot be interpreted only by molecular docking techniques and several samples may easily be mislabelled (Image 34).

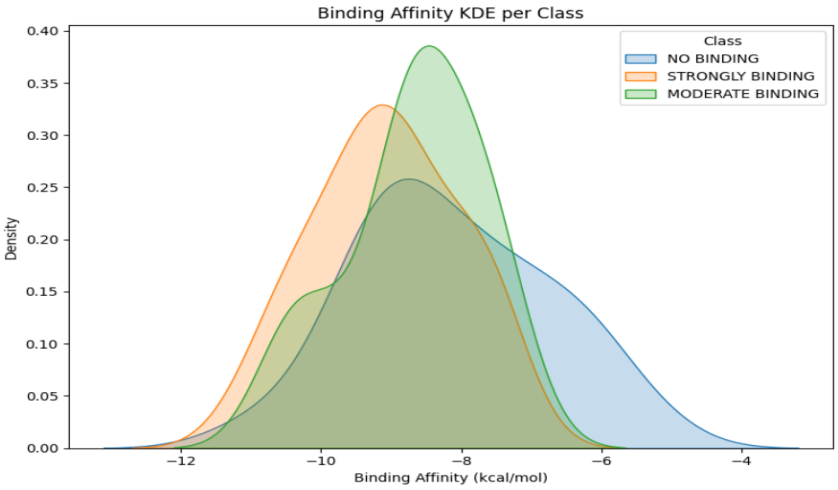


Image 34: KDE Plot of Binding Affinities across Classes.

4.2 BIOVIA Discovery Studio Results

All the features that are displayed in the machine learning pipeline and described extensively in Chapter 3.4 have been extracted from BIOVIA Discovery Studio. Each pose of each ligand was uploaded in the BIOVIA software and the features were presented as in the following box. This process was repeated for all 370 poses and the molecular details dataset was formed (Image 35). In addition, residue analysis dataset was formed by collecting all 370 images from all poses (as shown in Image 24), which depicts not only the type of interactions between human SERT and each ligand, but also distance metrics of the bonds participating in each interaction.

	Name	Visible	Color	Parent	Distance	Category	Types	From	From Chemistry	To	To Chemistry	Angle DHA	Angle HAY	Angle XDA	Angle DAY	Angle Deviation	Theta	Theta 2	Gamma	Closest Atom Distance
1	:UNL1:H -...	Yes		Ligand No...	2,19995	Hydrogen Bond	Conventi...	:UNL1:H	H-Donor	A:AS...	H-Acceptor	125,954	134,377							
2	:UNL1:C -...	Yes		Ligand No...	3,44996	Hydrogen Bond	Carbon H...	:UNL1:C	H-Donor	A:SER...	H-Acceptor			102,477	97,026					
3	:UNL1:C -...	Yes		Ligand No...	3,52996	Hydrogen Bond	Carbon H...	:UNL1:C	H-Donor	A:SER...	H-Acceptor			90,529	112,29					
4	:UNL1:C -...	Yes		Ligand No...	3,23738	Hydrogen Bond	Carbon H...	:UNL1:C	H-Donor	A:AL...	H-Acceptor			110,661	113,391					
5	:UNL1:C -...	Yes		Ligand No...	3,85325	Hydrophobic	Pi-Sigma	:UNL1:C	C-H	A:TY...	Pi-Orbitals					2,993	23,475			
6	A:TYR17...	Yes		Ligand No...	4,82745	Hydrophobic	Pi-Pi T-sh...	A:TYR176	Pi-Orbitals	:UNL1	Pi-Orbitals						9,195	65,693	58,935	3,798
7	A:PHE34...	Yes		Ligand No...	5,13944	Hydrophobic	Pi-Pi T-sh...	A:PHE341	Pi-Orbitals	:UNL1	Pi-Orbitals						11,736	87,944	82,991	3,747
8	A:SER438...	Yes		Ligand No...	4,24433	Hydrophobic	Amide-Pi ...	A:SER43...	Amide	:UNL1	Pi-Orbitals						33,297	36,011	3,694	3,514
9	A:ALA16...	Yes		Ligand No...	4,81083	Hydrophobic	Alkyl	A:ALA169	Alkyl	:UNL1	Alkyl									
10	A:ALA17...	Yes		Ligand No...	4,86172	Hydrophobic	Alkyl	A:ALA173	Alkyl	:UNL1	Alkyl									
11	:UNL1 - A...	Yes		Ligand No...	4,16633	Hydrophobic	Alkyl	:UNL1	Alkyl	A:ILE...	Alkyl									
12	:UNL1 - A...	Yes		Ligand No...	5,46013	Hydrophobic	Pi-Alkyl	:UNL1	Pi-Orbitals	A:ILE...	Alkyl									
13	:UNL1 - A...	Yes		Ligand No...	5,14885	Hydrophobic	Pi-Alkyl	:UNL1	Pi-Orbitals	A:VA...	Alkyl									
14	:UNL1 - A...	Yes		Ligand No...	4,48346	Hydrophobic	Pi-Alkyl	:UNL1	Pi-Orbitals	A:ILE...	Alkyl									

Image 35: Molecular Details for Paroxetine Pose 1 with SERT protein from BIOVIA Discovery Studio Visualizer.

4.3 Machine Learning Results

4.3.1 Molecular Descriptor Distribution Analysis Across 3 Classes

A comprehensive analysis is conducted in this section in order to outline the distribution patterns between molecular descriptors and specific residues with the target variable, namely binding classes. This analysis sheds light on which features dominate in each class, facilitating interpretability and explanation of the model performance.

In Image 36 there is clear evidence that higher molecular volumes and surface areas are correlated to stronger binding with SERT protein transporter. This occurs because the contact interface is maximized. Moderate binders show overlap with the adjacent classes, while non-binders have lower values of these features. In addition, in *"NO BINDING"* class, several outliers appear in the image, which might struggle model distinguishing binding classes.

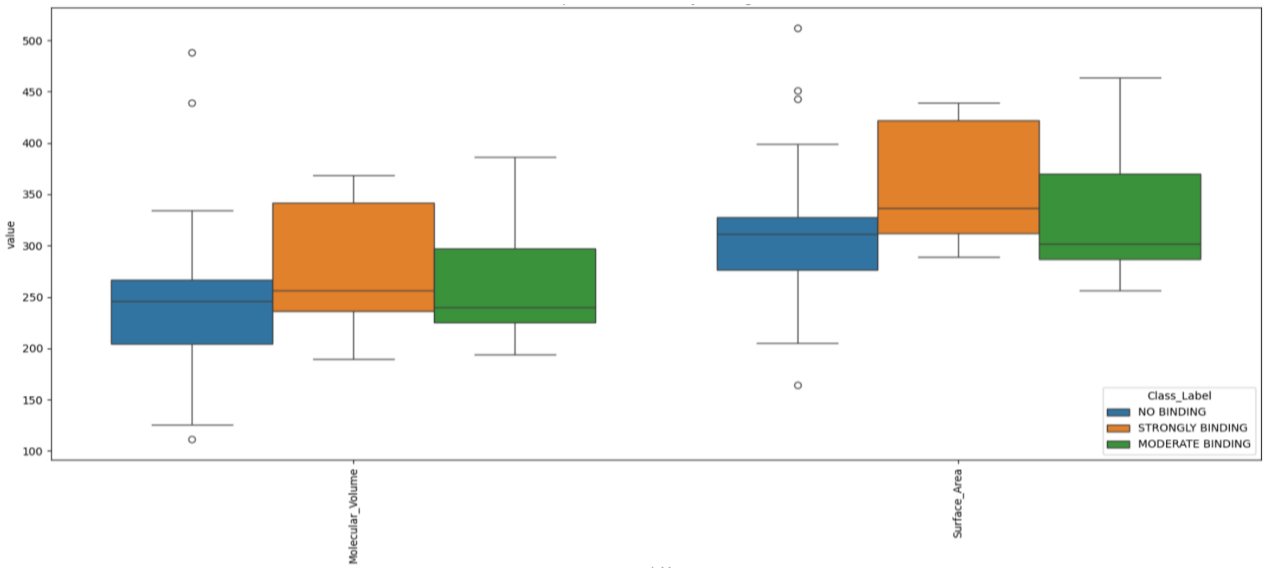


Image 36: "Molecular_Volume" and "Surface_Area" distributions across binding classes.

Image 37 depicts the distributions of several more features important for inhibition of SERT protein. *"Polar Surface Area"* is notably high in non-binders, while in the rest classes there is a significant percentage of overlapping. An explanation for this behaviour could be that excess of polar atoms may decrease the probability of a ligand to enter inside the binding pocket of SERT protein transporter which consists of several hydrophobic bonds. As a result, moderate values of PSA can facilitate this SERT inhibition. The angular descriptors *"ANGLE_DHA"* and *"ANGLE_HAY"* capture critical geometric relationships involved in hydrogen bonding. Strong binders tend to exhibit a high variability in values ranging from 0°- ~130°, while moderate binders from about 50° to ~135°. Interestingly, the *"NO BINDING"* class shows higher angular values that might be more ideal for binding. This irrelevant behaviour suggests that these 2 angular features cannot together effectively classify ligands based on SERT inhibition. Furthermore, image 37 shows that strong binders possess a compact and narrow distribution in *"THETA"*, *"THETA_2"* and *"GAMMA"* angular metrics in contrast to moderate and non-binders which have broader distributions. In strong binders, *"THETA"* falls between 25–35°, ~ 20–30° for moderate and ~ 20–35° for *"NO BINDING"* class. For *"THETA_2"*, 35–45° for strong binders, ~ 30–50° for moderate and 10–45° for non-binders. For *"GAMMA"* feature, the IQR of strong binders is ~ 20–40°, ~ 30–70° for moderate and ~ 10–50° for non-binders. Lastly, for "*ANGLE_XDA"* and "*ANGLE_DAY"* descriptors, huge overlap exists across the 3 classes. The median values, though smaller in non-binders with higher variance, the interquartile ranges are largely comparable, indicating that these angular features may not distinctly differentiate across the 3 classes. This suggests that while these angles may add value to the overall geometry, they do not contribute to the final model's estimation.

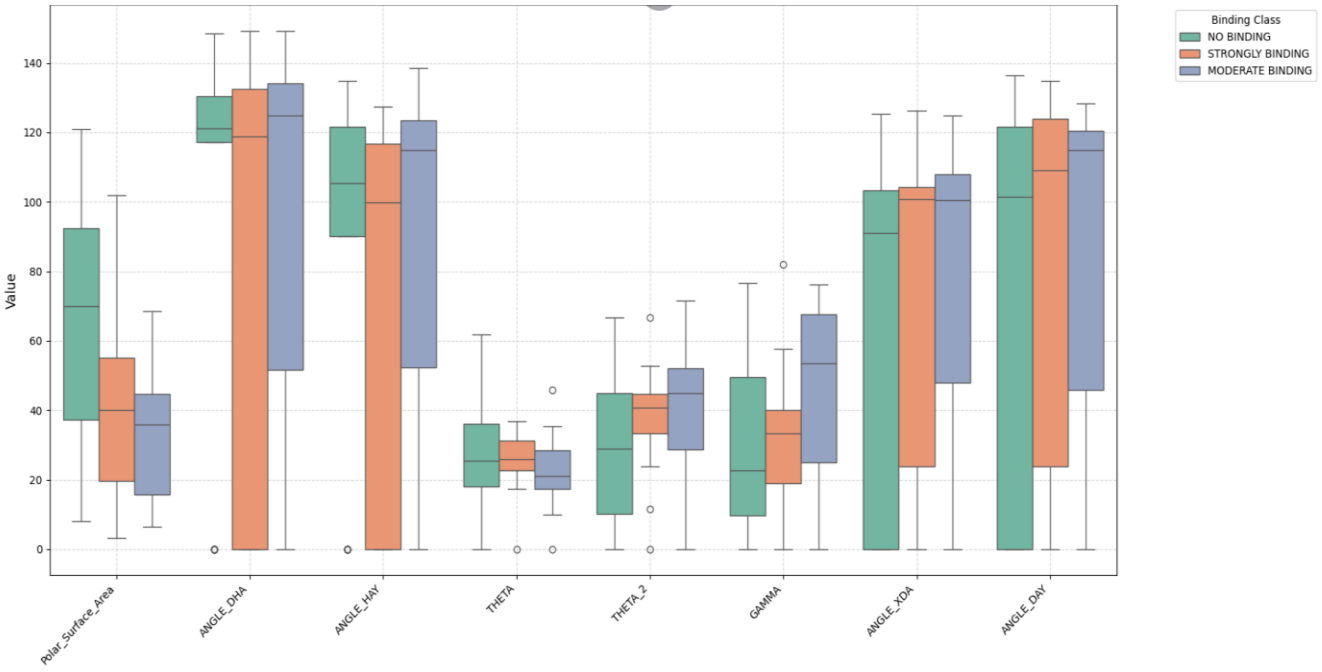


Image 37: Polar Surface Area and other angular features distributions across binding classes.

Last 5 features for molecular descriptors are the following presented in Image 38. In *"Ligand_Distance_to_Grid_Box_Center"*, the strong binders have a median value between moderate and non-binders, indicating that the boundaries that determine the binding class of a ligand with SERT protein based on this feature are relatively small. *"NO BINDING"* class varies more than the other 2 classes, while at the same time in all classes several outliers appear. The second feature, namely *"logP"*, indicates that higher median values of hydrophobicity, typically around 3.5 are correlated to strong binders, compared to moderate and non-binders that exhibit values closer to 3. This means, that, increased hydrophobicity of a ligand favours the ligand binding into SERT pocket. The descriptor "*Mean_Hydrogen_Bond_Length_Distance"* shows that *"STRONGLY BINDING"* ligands tend to exhibit slightly longer hydrogen bond distances with higher median values than non-binders and moderate binders, the last of which are the intermediate step. For "*Mean_Hydrophobic_Bond_Length_Distance"* there are no obvious differences across the 3 classes, neither in the interquartile ranges, nor in the median values, suggesting that this feature does not help classify ligands for SERT inhibition. Finally, "*CLOSEST_ATOM_DISTANCE"* shows only subtle median shifts across the 3 classes.

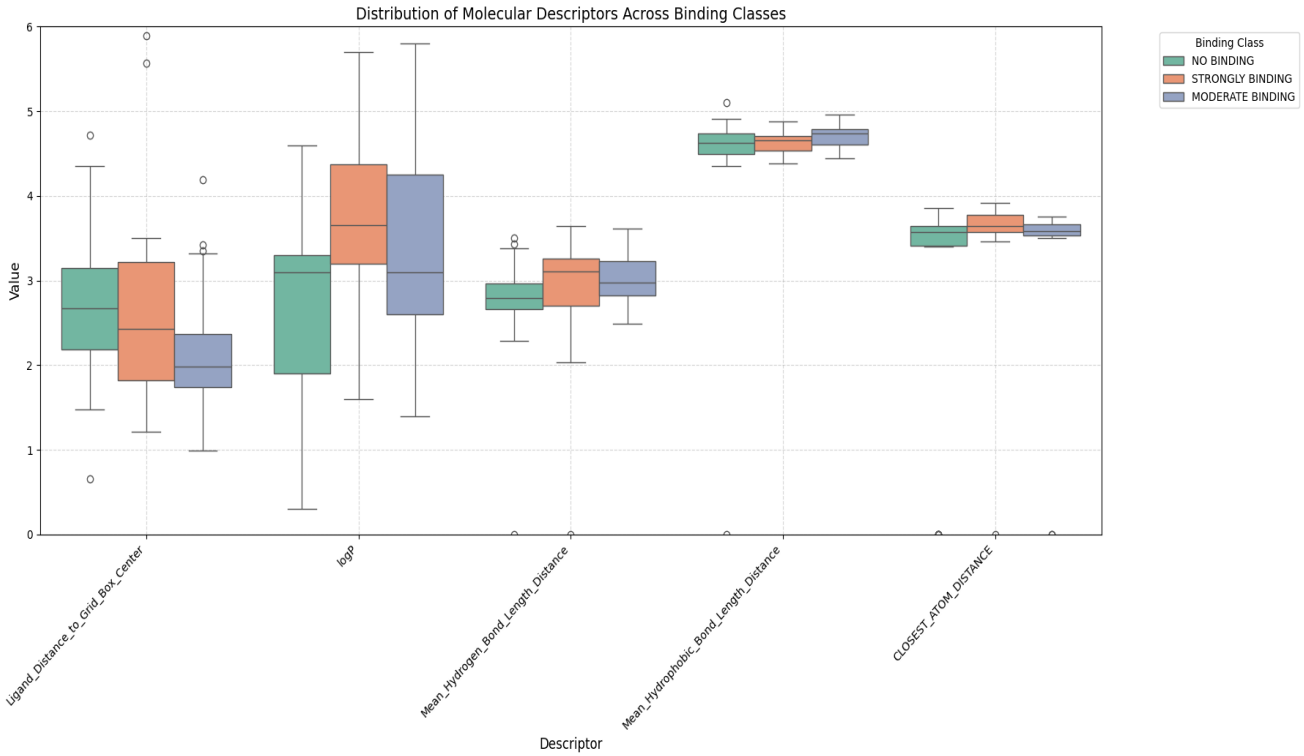


Image 38: More molecular descriptors across the 3 binding classes.

4.3.2 Residue-Level Distribution Analysis across 3 classes

Residue-level frequency analysis was then conducted through Python programming language. The top twenty residue interaction counts across the three classes are illustrated in Image 39. Notably, the residue *"ILE_172_HYD"* (Isoleucine 172) is the most dominant, particularly in strong binders, across the 3 classes and the "HYD" suffix suggests that it participates in hydrophobic interactions, which play crucial role in stabilizing the ligand binding. 225 interactions appear totally in strong binders, 176 in moderate and 138 in non-binders. Similarly, "*TYR_176_HYD", "PHE_335_HYD", "TYR_175_VDW", "THR_439_VDW"* and *"THR_497_VDW"* show progressively increasing interaction counts from *"NO BINDING"* to *"STRONGLY BINDING"* class. Interestingly, *"ALA_173_HYD"*, *"SER_438_HYD"*, *"THR_439_HYD"* and *"TYR_95_H"* appear predominantly in the top 20 residues in strong binders which are compatible with literature review to some extent. Other residues, such as *"ASP_98_H"*, *"TYR_176_VDW"*, *"VAL_501_HYD"* and *"VAL_501_VDW"* are observed frequently in moderate and non-binders, however they are also present, albeit with lower counts, in *"STRONGLY BINDING"* class.

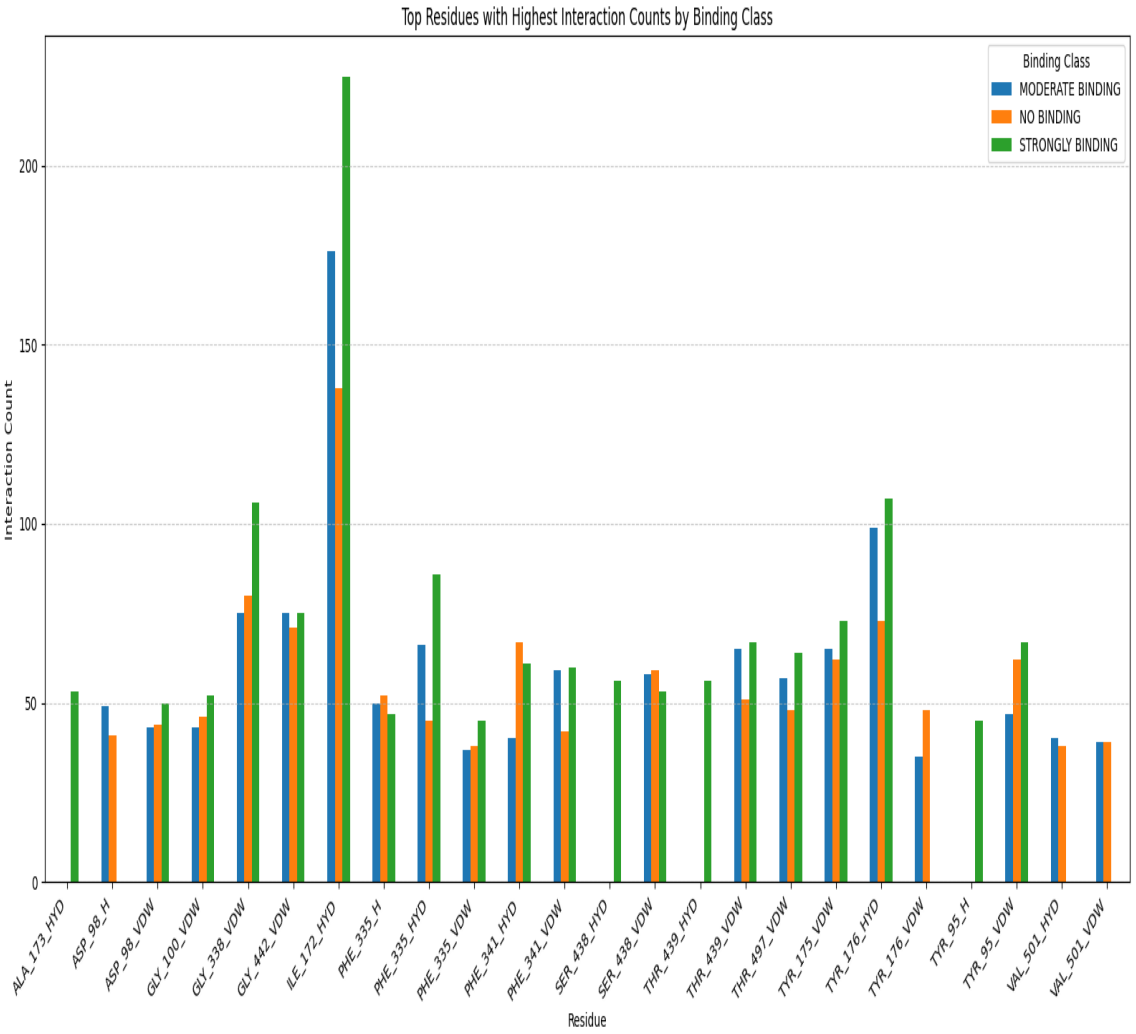


Image 39: Top 20 most frequent residues across the three binding classes.

4.3.3 Machine learning pipeline outputs

As mentioned in Chapter 3.5.3 in the present thesis, the remaining features were 156. From the correlation matrix below, the thirty features with the highest contribution to the target variable *"Labelled class"* are highlighted based on absolute values (Image 40). Since the encoding was applied based on the 3.5.2 chapter where 0 is *"MODERATE BINDING"* class, 1 is the *"NO BINDING"* class and 2 the *"STRONGLY BINDING"*, this means that higher values in the correlation matrix are possibly connected to class 2 with a linear correlation. For this reason, *"ALA_173_HYD"* with a correlation 0.30 indicates that in more interactions this residue is involved, the higher the chances are to classify this ligand as strong binder to SERT protein transporter. In the same way, *"Hydrophobic_bonds_count"*, *"ALA_169_HYD"* and more residues are correlated to strong binders. On the other hand, *"TYR_175_HYD"*, *"GAMMA"* and *"ALA_96_VDW"* have negative values of correlation, meaning that as these increase, the less probable is, for a ligand to be assigned as

"STRONGLY BINDING" class. However, all values in this matrix are quite small ranging from -0.28-0.30, so no clear conclusions can be made for the strength and the kind of the relationship between features and target variable. For example, in the negative correlations, it is uncertain whether a more negative value drives the model towards class 1 which is non-binders or class 0 which is the moderate binders.

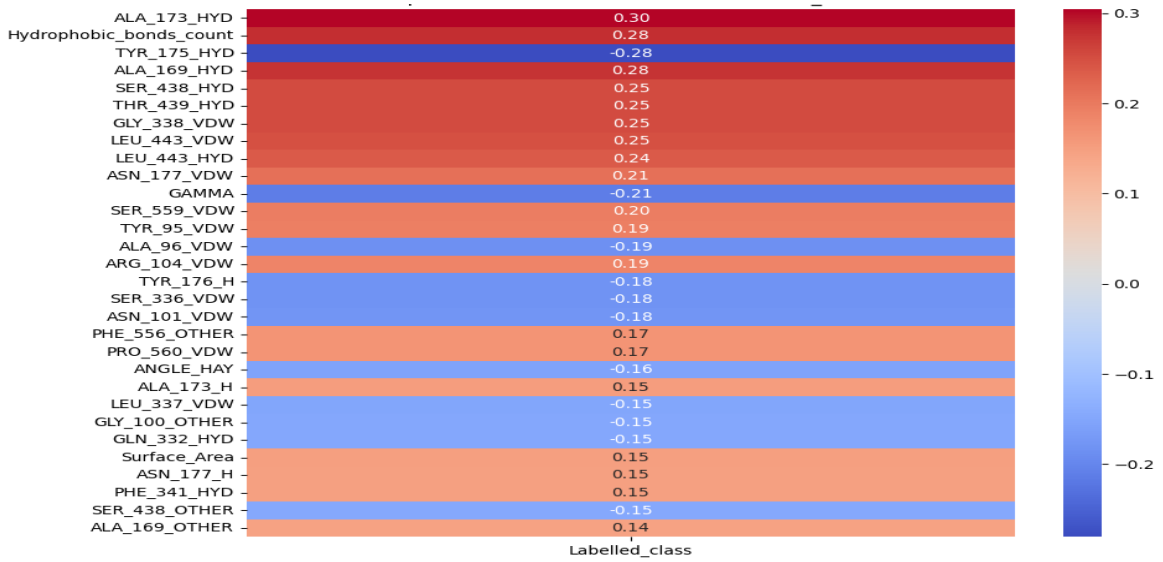


Image 40: Correlation of the top thirty features with the target variable Labelled class.

After all preprocessing steps and correlation analysis of features mentioned both in chapter 3 and chapter 4, next step was to evaluate the 5 classical machine learning algorithms (XGBoost, Random Forest, LightGBM, Logistic Regression, Support Vector Machines and the ensemble method Voting Classifier). These algorithms were trained and tuned as mentioned in chapter 3.5.4 and an estimation performance was conducted based on the evaluation metrics from chapter 3.5.5. Table 13 presents the train and test accuracies for each model across the five outer folds of nested approach. The greater the difference between train and test fold, the higher the probability for overfitting is, since the model "memorizes" the cases instead of learning them. For instance, XGBoost shows up to 98.3% in training fold while the mean accuracy in test is only 52.6%, which means that there is a high risk for overfitting. Similarly, LightGBM exhibits lower training and test accuracy than XGBoost, but again the gap is huge. Logistic Regression algorithm and Support Vector Machines underperform, since their training accuracies are more than 90% and their test accuracies 47.3% and 40.7% respectively. On the other hand, the Random Forest classifier has approximately 84% training accuracy and achieves the best mean test accuracy with 61.9% and even 80% to one fold. This results in a more balanced condition where the model generalizes better than the other algorithms, while at the same time offers room for improvement in training, possibly with increasingly number of samples. After

Random Forest, Voting Classifier seems to perform also quite satisfactorily with 92.2% training and 60.6% test accuracy. All results are shown in table 13 below.

Fold	XGBoost	Random Forest	LightGBM	Logistic Regression	Support Vector Machines	Voting Classifier
1 Train	0.983	0.881	0.780	0.949	0.966	0.966
1 Test	0.600	0.667	0.467	0.60	0.533	0.667
2 Train	0.949	0.864	0.847	0.898	0.898	0.898
2 Test	0.600	0.667	0.533	0.333	0.333	0.733
3 Train	0.814	0.814	0.763	0.966	0.898	0.898
3 Test	0.400	0.533	0.467	0.400	0.400	0.533
4 Train	0.915	0.814	0.814	0.966	1.000	0.932
4 Test	0.600	0.800	0.467	0.533	0.267	0.667
5 Train	0.900	0.833	0.850	1.000	1.000	0.917
5 Test	0.429	0.429	0.500	0.500	0.500	0.429

Table 13: Per fold accuracy in train and test for each algorithm.

Besides table 13, below in Tables 14-20 there are other additional metrics that are useful for forming a comprehensive view about how models perform. These include precision, recall, F1-score and area under the curve analysis (via one-Vs-rest strategy). Random Forest and Voting Classifier achieved the highest area under the curve (AUC) for all three classes, reflecting strong discriminative capability with both models being more able to identify the strong binders. This view is reinforced by the fact that F1-score for class 2 is 0.68 in RF and 0.65 in Voting Classifier. Both perform quite effectively in *"NO-BINDING"* class, while they struggle to identify the moderate binders with 0.52 and 0.50 F1-score respectively. XGBoost, although it performs moderately, it cannot easily recognize class 1 with 0.48 F1-score, but it shows better results for class 0. LightGBM seems to struggle to identify the moderate binders. However it underperforms in general. Last 2 algorithms, namely LR and SVM show limited effectiveness with low accuracies, recall and F1-scores, especially SVM where AUC scores are close to 0.5 and all evaluation metrics close to 0.40 that are near the boundaries of random guess in a multiclassification task with 3 classes.

Model	Nested Accuracy	Macro Precision	Macro Recall	Macro F1	AUC Class 0	AUC Class 1	AUC Class 2
XGBoost	0.526	0.537	0.526	0.529	0.72	0.69	0.62
Random Forest	0.619	0.618	0.617	0.615	0.69	0.71	0.73
LightGBM	0.487	0.49	0.483	0.482	0.73	0.71	0.62

Logistic Regression	0.473	0.475	0.474	0.473	0.60	0.65	0.59
SVM	0.407	0.396	0.411	0.401	0.64	0.61	0.51
Voting Ensemble	0.606	0.605	0.602	0.599	0.73	0.74	0.74

Table 14: Average evaluation metrics of all algorithms.

XGBoost	**Class 0**	**Class 1**	**Class 2**
Precision	0.63	0.48	0.50
Recall	0.52	0.48	0.58
F1-score	0.57	0.48	0.54
Support	23	25	26

Table 15: Evaluation metrics per class for XGBoost.

Random Forest	**Class 0**	**Class 1**	**Class 2**
Precision	0.58	0.61	0.67
Recall	0.48	0.68	0.69
F1-score	0.52	0.64	0.68
Support	23	25	26

Table 16: Evaluation metrics per class for Random Forest.

LightGBM	**Class 0**	**Class 1**	**Class 2**
Precision	0.50	0.50	0.47
Recall	0.39	0.48	0.58
F1-score	0.44	0.49	0.52
Support	23	25	26

Table 17: Evaluation metrics per class for LightGBM.

Logistic Regression	**Class 0**	**Class 1**	**Class 2**
Precision	0.44	0.50	0.48
Recall	0.52	0.44	0.46
F1-score	0.48	0.47	0.47
Support	23	25	26

Table 18: Evaluation metrics per class for Logistic Regression.

Support Vector Machines	**Class 0**	**Class 1**	**Class 2**
Precision	0.44	0.44	0.30
Recall	0.52	0.48	0.23

F1-score	0.48	0.46	0.26
Support	23	25	26

Table 19: Evaluation metrics per class for Support Vector Machines.

Voting Classifier	**Class 0**	**Class 1**	**Class 2**
Precision	0.59	0.61	0.62
Recall	0.43	0.68	0.69
F1-score	0.50	0.64	0.65
Support	23	25	26

Table 20: Evaluation metrics per class for Voting Classifier.

The confusion matrixes are presented below, where first row refers to class 0, second to class 1 and third to class 2 and that is similar for columns.

Confusion Matrix for XGBoost: Here, XGBoost out of 23 samples in class 0, the model finds 12 and mislabels 5 into class 1 and 6 into class 2. For the second row, out of the 25 non-binders, the model identifies correctly 12 samples and mislabels 4 into moderate binders and 9 into strong binders, which is quite confusing why the majority of wrong predictions are strong binders, because theoretically speaking, the moderate binders are an intermediate category between non-binders and strong binders. Lastly, 15 samples are correctly identified as class 2 while 8 are mislabelled into class 1 and 3 as class 0, providing similar behaviour as in the second row.

[12, 5, 6]
[4, 12, 9]
[3, 8, 15]

Confusion Matrix for Random Forest: This model shows the best performance, since it correctly finds 18 samples belonging to class 2 and mislabels 3 into moderate binders and 5 into non-binders. Furthermore, it identifies 17 non-binders and misclassifies 3 as strong binders and 5 as moderate binders, which is a more logical condition. And finally, as all models struggle, it predicts correctly 11 *"MODERATE BINDING"* samples out of the 23 and the rest 12 are shared equally falsely into the other 2 classes.

[11, 6, 6]
[5, 17, 3]
[3, 5, 18]

Confusion Matrix for LightGBM: LightGBM performs better in *"STRONG BINDING"* class, where it finds 15 out of the 26 samples and misclassifies 7 into non-binders and 4 into moderate binders. It also correctly identifies 12 samples of class 1, however it mislabels 8 into class 2 and 5 into class 0 following similar irrelevant behaviour as the XGBoost algorithm. In class 0, it offers limited success, since more than half of samples are misclassified.

[9, 5, 9]
[5, 12, 8]

[4, 7, 15]

Confusion Matrix for Logistic Regression: Logistic Regression algorithm performs moderately in *"MODERATE BINDING"* class with 12 correct predictions out of 23, while in the other 2 classes the model struggles to recognize them with almost 45% correct predictions.

[12, 4, 7]
[8, 11, 6]
[7, 7, 12]

Confusion Matrix for Support Vector Machines: SVM is the weakest model with moderate performance in class 0 and class 1 and poor performance in the *"STRONG BINDING"* class with 6 correct predictions out of the 26.

[12 5 6]
[5 12 8]
[10 10 6]

Confusion Matrix for Voting Classifier: This ensemble classifier performs almost similar to RF. Class 2 is identical to RF, while class 1 has only 1 difference in one misclassified sample, where in RF 5 were mislabelled to class 0 and 3 to class 2, while in this case 4 were mislabelled to class 0 and 4 to class 2. Another small difference is in moderate binders where RF correctly identifies 11, while voting classifier finds 10.

[10, 6, 7]
[4, 17, 4]
[3, 5, 18]

Robust Features per Model

For interpretability and robustness purposes, each model produced a list of features that appeared in at least 3 of the 5 outer folds of the nested cross validation and are depicted below. From the best 3 models analyzed previously, which are Random Forest, Voting Classifier and XGBoost, the overlapping robust features include residues *"ALA_173_HYD"*, *"PHE_341_HYD"*, *"PHE_341_VDW"* and for molecular descriptors **"***ANGLE_HAY"*, *"Ligand_Distance_to_Grid_Box_Center"*, *"Mean_Hydrogen_Bond_Length_Distance"*, *"Mean_Hydrophobic_Bond_Length_Distance"*, *"Polar_Surface_Area"*, *"Surface_Area*" and *"THETA"* (Image 41).

```
# Robust Features
feature_counter = Counter([f for fold in fold_best_feature_lists for f in fold])
print("\n Feature selection frequency across folds:")
for feat, count in feature_counter.most_common():
    print(f"{feat}: {count}")
robust_features = [f for f, c in feature_counter.items() if c >= 3]
print(f"\n Robust features (selected in ≥3 folds): {robust_features}")
```

Image 41: Code cell in Python for picking the robust features.

Below the robust features for each model are mentioned:

- **XGBoost:**
  - *"Polar_Surface_Area"*, *"ILE_172_HYD"*, *"GAMMA"*, *"ALA_173_HYD"*, *"ALA_96_VDW"*, *"ANGLE_HAY"*, *"Surface_Area"* and *"TYR_175_HYD"*
- **Random Forest:**
  - *"Polar_Surface_Area"*, *"ILE_172_HYD"*, *"logP"*, *"ANGLE_HAY"*, *"Mean_Hydrophobic_Bond_Length_Distance"*, *"GAMMA"*, *"Surface_Area"*, *"THETA_2"*, *"Ligand_Distance_to_Grid_Box_Center"*, *"CLOSEST_ATOM_DISTANCE"*, *"ALA_96_VDW"*, *"Binding_Affinity_kcal/mole_"*, *"THETA"*, *"Mean_Hydrogen_Bond_Length_Distance"*, *"THR_439_VDW"*, *"TYR_95_H"*, *"Molecular_Volume"*, *"ANGLE_XDA"*, *"PHE_335_HYD"*, *"PHE_341_HYD"* and *"ALA_173_HYD"*
- **LightGBM:**
  - *"ALA_96_VDW"*, *"THETA_2"*, *"Polar_Surface_Area"*, *"ANGLE_HAY"*, *"CLOSEST_ATOM_DISTANCE"*, *"Mean_Hydrophobic_Bond_Length_Distance"*, *"ASN_177_VDW"*, *"PHE_335_HYD"*, *"logP"*, *"THETA"*, *"Surface_Area"*, *"ALA_173_HYD"*, *"GLY_338_VDW"*, *"PHE_341_HYD"*, *"SER_336_VDW"*, *"TYR_95_VDW"*, *"GAMMA"*, *"Ligand_Distance_to_Grid_Box_Center"*, *"Mean_Hydrogen_Bond_Length_Distance"*, *"ANGLE_DHA"*, *"TYR_95_H"*, *"PHE_334_VDW"*, *"SER_336_H"*, *"PHE_335_VDW"*, *"ANGLE_XDA"*, *"Molecular_Volume"*, *"SER_438_HYD"* and *"ARG_104_VDW"*
- **Logistic Regression:**
  - *"ALA_96_VDW"*, *"TYR_175_HYD"*, *"GLY_338_VDW"*, *"Polar_Surface_Area"*, *"ANGLE_HAY"*, *"PHE_335_HYD"*, *"SER_438_H"*, *"SER_336_HYD"*, *"ALA_169_HYD"* and *"ASP_98_OTHER"*
- **SVM:**
  - *"ALA_96_VDW"*, *"GLY_338_VDW"*, *"TYR_175_HYD"*, *"ANGLE_HAY"*, *"Polar_Surface_Area"*, *"THR_497_VDW"*, *"SER_438_H"*, *"PHE_335_HYD"*, *"SER_336_HYD"*, *"SER_336_VDW"*, *"PHE_334_VDW"*
- **Voting Classifier:**
  - *"ILE_172_HYD"*, *"GAMMA"*, *"ALA_96_VDW"*, *"Polar_Surface_Area"*, *"ANGLE_HAY"*, *"Surface_Area"*, *"Mean_Hydrogen_Bond_Length_Distance"*, *"Mean_Hydrophobic_Bond_Length_Distance"*, *"Binding_Affinity_kcal/mole_"*, *"Ligand_Distance_to_Grid_Box_Center"*, *"THETA"*,

"CLOSEST_ATOM_DISTANCE", *"logP"*, *"TYR_176_HYD"*, *"ANGLE_XDA"*, *"THETA_2"*, *"ALA_173_HYD"* and *"Molecular_Volume"*

The hyperparameters from the best fold are the following:

Random Forest:

Best Outer Fold Index: 3, Accuracy: 0.800

Best Hyperparameters for that fold: {'class_weight': 'balanced', 'max_depth': 5, 'max_features': 'log2', 'min_samples_leaf': 5, 'min_samples_split': 25}

XGBoost:

Best Outer Fold Index: 0, Accuracy: 0.600

Best Hyperparameters for that fold: {'colsample_bytree': 0.7, 'gamma': 0.5, 'learning_rate': 0.1, 'max_depth': 4, 'reg_alpha': 1, 'reg_lambda': 5, 'subsample': 0.7}

4.4 Explainability Analysis with SHAP Values

For validity and interpretability of the previous results, SHAP (Shapley Additive exPlanations) values were applied for the optimal 2 algorithms, which are Random Forest and XGBoost. Since Voting Classifier has not an internal decision structure and combines the outputs of the other 2 models, SHAP analysis cannot be conducted for this ensemble method. For this reason, the 2 models were retrained on the robust features that are described in chapter 4.3.3 and on the best hyperparameters from the best fold of nested cross validation approach. This means that the final dataset size is (74, 21) for Random Forest and (74, 8) for XGBoost algorithm and the SHAP summary plots were generated as shown below in Image 42 and Image 43. The graphs illustrate that *"Polar_Surface_Area"* in both models is the top contributor and affects the decision more in samples belonging to class 1 and class 0. Among the 5 most predictive features in both models is *"GAMMA"* angle, which is equally useful for discriminating the strong and the moderate binders in both models. In addition, *"ANGLE_HAY"* feature is very decisive in XGBoost algorithm, especially in class 0 and class 2, while in RF it is not that critical. *"Surface_Area"* is important in XGBoost algorithm to discriminate class 1 and class 2 samples, while in RF is less vital and is more predictive towards moderate and strong binders. As for residues *"ALA_173_HYD"* and *"ILE_172_HYD"* they are considered highly important, especially in the Random Forest Classifier, and facilitate decisions towards class 1 and class 2. Furthermore, the models show strong agreement in the *"ALA_96_VDW"* residue and that is that it contributes to moderate and non-binders. Another highlighted fact is that the binding affinity in Random Forest

contributes almost equally to all classes, meaning that this feature exclusively, although it is very important in molecular docking simulations and computational biology in general, it does not achieve proper classification of these 74 ligands in the 3 classes. This is also confirmed by the fact that the SHAP value of binding affinity is close to zero. Notably, SHAP values derived from Random Forest were significantly lower in magnitude than those from XGBoost. This difference stems from their underlying architectures. XGBoost is a gradient-boosting algorithm that builds trees in a sequential manner and optimizes the loss function at each stage, resulting in more strict and confident predictions. In contrast, Random Forest mechanism as described in section 3.5.4, averages the decisions of several trees and that led to the production of more distributed and conservative SHAP values. Despite these scale differences, the overall proportions of decisions towards the 3 classes indicate similar patterns and the most critical features are observed.

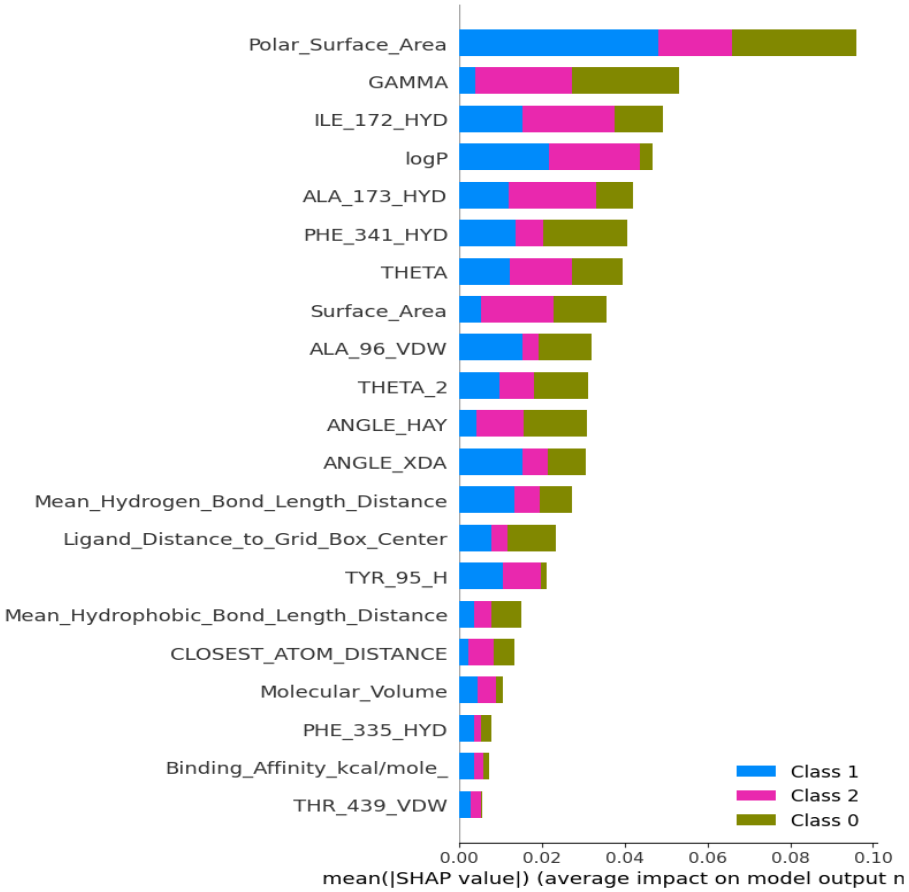


Image 42: SHAP summary plot for Random Forest.

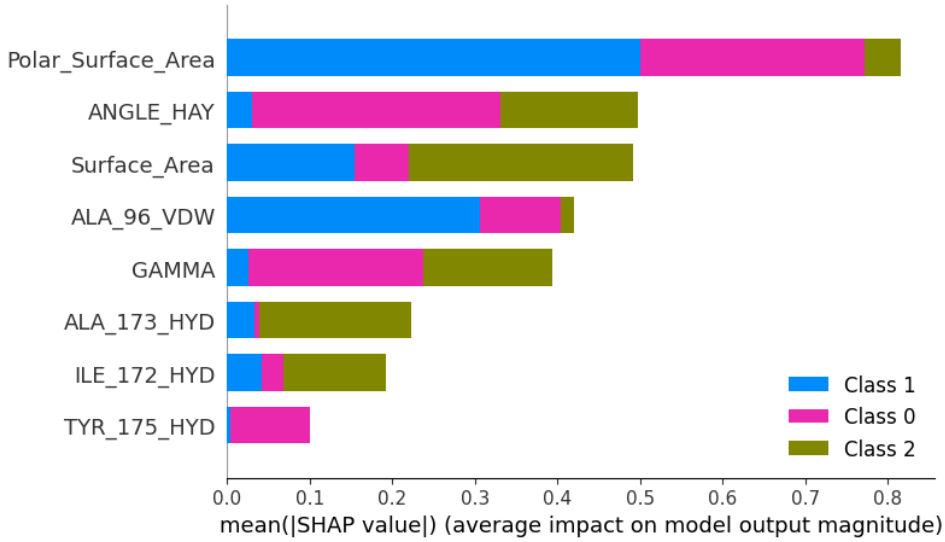


Image 43: SHAP summary plot for XGBoost algorithm.

For better understanding of the predictive power of each feature in the 2 previous algorithms, SHAP summary plots were converted into class-wise percentages as the following heatmaps illustrate. As discussed in chapter 4.4, it is clear that random forest operates in a more conservative manner since none of the 21 robust features contribute more than 50% towards a class. On the other hand, in XGBoost algorithm, the differences in percentages across classes are wider, enhancing the strict and confident character of this model.

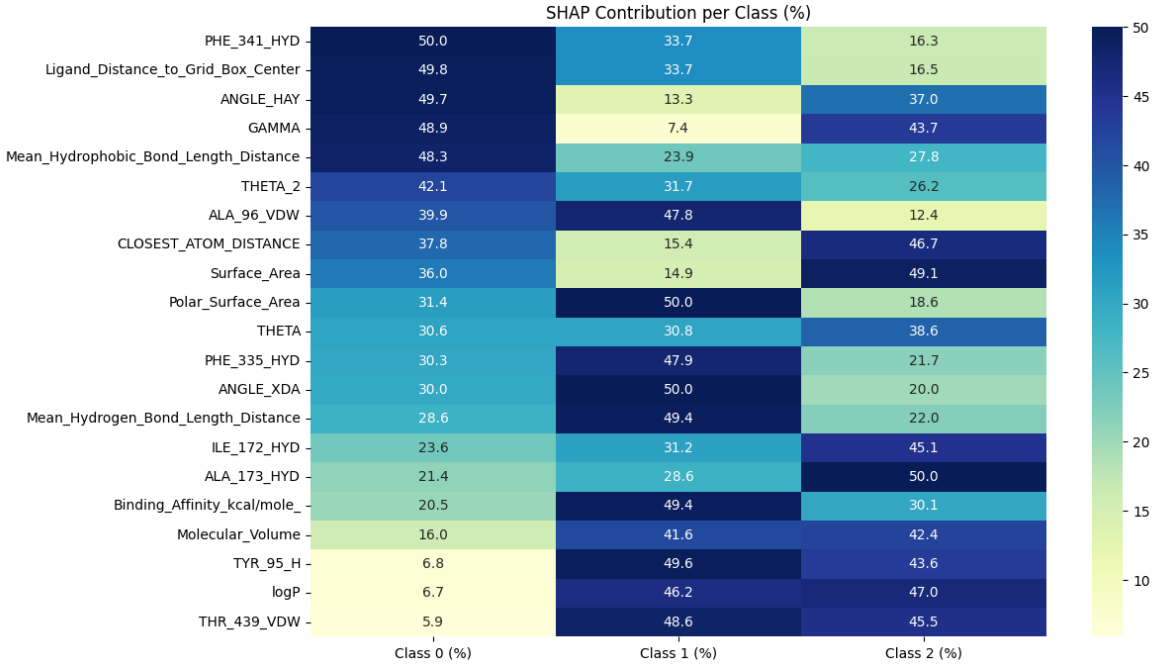


Image 44: SHAP contribution percentage per class for Random Forest.

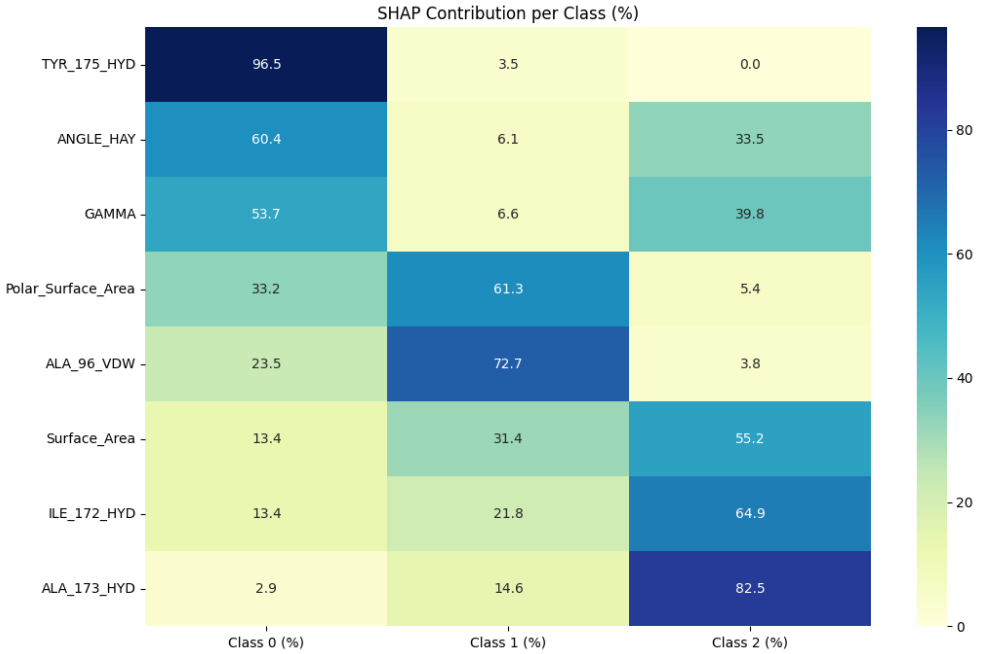


Image 45: SHAP contribution percentage per class for XGBoost.

To complement and validate the class-specific SHAP breakdowns, a cross-model comparison was performed by averaging the absolute SHAP values across all samples and classes for both Random Forest and XGBoost models. All features are shown below in Image 46 with their overall contribution to model predictions. More specifically, features such as "*Polar_Surface_Area"* and "*GAMMA"* were found to be the top contributors in both models. In conclusion, XGBoost exhibited generally higher SHAP magnitudes than the Random Forest. However, the relative rankings between the two models show strong agreement. This alignment enhances the robustness and the repeatability of the results, which are totally compatible with what previous SHAP analysis revealed.

	Feature	RF_SHAP	XGB_SHAP	RF_Rank	XGB_Rank	Avg_Rank
14	Polar_Surface_Area	0.032032	0.271862	1.0	1.0	1.0
6	GAMMA	0.017721	0.131412	2.0	5.0	3.5
7	ILE_172_HYD	0.016449	0.064228	3.0	7.0	5.0
0	ALA_173_HYD	0.014010	0.074482	5.0	6.0	5.5
15	Surface_Area	0.011911	0.163660	8.0	3.0	5.5
1	ALA_96_VDW	0.010638	0.140102	9.0	4.0	6.5
2	ANGLE_HAY	0.010273	0.165883	11.0	2.0	6.5
3	ANGLE_XDA	0.010152	NaN	12.0	NaN	NaN
4	Binding_Affinity_kcal/mole_	0.002381	NaN	20.0	NaN	NaN
5	CLOSEST_ATOM_DISTANCE	0.004472	NaN	17.0	NaN	NaN
8	Ligand_Distance_to_Grid_Box_Center	0.007749	NaN	14.0	NaN	NaN
9	Mean_Hydrogen_Bond_Length_Distance	0.009042	NaN	13.0	NaN	NaN
10	Mean_Hydrophobic_Bond_Length_Distance	0.004994	NaN	16.0	NaN	NaN
11	Molecular_Volume	0.003506	NaN	18.0	NaN	NaN
12	PHE_335_HYD	0.002572	NaN	19.0	NaN	NaN
13	PHE_341_HYD	0.013507	NaN	6.0	NaN	NaN
16	THETA	0.013146	NaN	7.0	NaN	NaN
17	THETA_2	0.010355	NaN	10.0	NaN	NaN
18	THR_439_VDW	0.001845	NaN	21.0	NaN	NaN
19	TYR_175_HYD	NaN	0.033282	NaN	8.0	NaN
20	TYR_95_H	0.007015	NaN	15.0	NaN	NaN
21	logP	0.015598	NaN	4.0	NaN	NaN

Image 46: Absolute SHAP values of robust features across the 2 models.

4.4.1 Misclassification Analysis

To further identify and assess the limitations of the models applied, an in-depth study was conducted into the causes behind them based on the following misclassification analysis. Figure 47 points out that Random Forest algorithm mislabelled 10 samples, while XGBoost 8 and the common incorrect predictions were 4 which is inevitably a high percentage. This strong agreement between the two models highlights that although their core architecture is different, they can effectively detect possible mistakes in the initial dataset configuration and question directly the binding behaviour of these compounds. More specifically, the four common misclassifications were:

- *"Maprotiline"* was predicted as a strong binder by both models, though it has an actual label 1. This could be a reasonable condition, because it is a tetracyclic antidepressant, but it has very weak SERT inhibition with 5800 Ki value from PDSP.
- *"Nefazodone"* was mislabelled into strong binder but its actual label is moderate binder.
- *"Trimipramine"* showed a controversial behaviour, because Random Forest predicted it as a strong binder, while XGBoost as non-binder. However, the true label is *"MODERATE BINDING"* class.
- *"Vilazodone"*, which was predicted as a non-binder again by both models, but its actual label is strong binder with Ki value approximately 0.5 and it is a serotonin modulator.

Based on these results, it is obvious that they reinforce earlier findings from confusion matrices and SHAP values, where the most challenging group to identify, is the moderate binders. Two out of the 4

misclassifications were between *"MODERATE BINDING"* class and one of the remaining two. This suggests, that class 0 shares potentially an overlapping feature space and constitutes the intermediate condition. Nevertheless, in two cases, like *"Vilazodone"* and *"Maprotiline"*, the models do not follow an expected prediction, indicating potential overlap even in the distinct *"STRONG BINDING"* and *"NO BINDING"* classes. For the remaining six samples of Random Forest, 5 of them were again mislabelled between moderate binders and another class, while in 1 case, which was *"CID_24856107"* between strong and non-binders. As for XGBoost algorithm, for the remaining four samples, two of them were between moderate binders and one of the adjacent classes, one was classified as strong binder but it was a non-binder and finally one was labelled as a non-binder but it was actually belonging to *"STRONGLY BINDING"* class.

	Ligand Name	True Class_RF	Predicted Class_RF	Predicted Class_XGB
0	CID_11310988	2.0	0.0	NaN
1	CID_24855981	NaN	NaN	2.0
2	CID_24856046	0.0	2.0	NaN
3	CID_24856107	2.0	1.0	NaN
4	CID_24947939	0.0	2.0	NaN
5	CID_44351345	NaN	NaN	1.0
6	Diphenhydramine	1.0	0.0	NaN
7	Doxepin	NaN	NaN	1.0
8	Iprindole	1.0	0.0	NaN
9	Maprotiline	1.0	2.0	2.0
10	Nefazodone	0.0	2.0	2.0
11	Trimipramine	0.0	2.0	1.0
12	Vilazodone	2.0	1.0	1.0
13	Zolpidem	NaN	NaN	2.0

Image 47: Misclassification analysis of Random Forest and XGBoost algorithms.

To better understand the underlying path by which each model forms a final prediction, waterfall plots were generated. They consist of a base value denoted as *"E[f(X)]"* and it is the mean model output, namely what the model expects based on all samples of the dataset. Then the actual model's output is *"f(x)"* and its value depends on whether the contribution of the additional feature pushes the model towards the predicted class or not. If the SHAP value is positive (red colour in Image 48), it indicates that this specific feature helps the model to get this particular prediction, while if it is negative (blue colour in Image 48), it means that the specific feature introduced is preventing the model from getting this predicted class. For the short waterfall analysis below, 4 compounds are depicted which represent 4 different cases.

"Paroxetine" is an example of a positive predicted sample in both models that belongs in the *"STRONG BINDING"* class. For Random Forest classifier, initially the base value was relatively low with a value of 0.331. The most positive contributors were *"ALA_173_HYD"*, *"Surface_Area"*, *"Polar_Surface_Area"*, *"THETA"*, *"logP"* and *"THETA_2"* while the rest 15 features did not affect the model significantly. Final *"f(x)"* increased to 0.478, so the overall contribution was relatively small. On the other hand, XGBoost again recognized *"ALA_173_HYD"* as the most positive influential feature for predicting class 2, followed by *"Surface_Area"* and *"GAMMA"*. Their SHAP values were 0.32, 0.28 and 0.16 respectively. The final output *"f(x)"* went from 0.545 to 1.404, highlighting the importance of these features. However, there was 1 feature, *"ILE_172_HYD"* that pushed the model towards the opposite direction of class 2 prediction (Image 48).

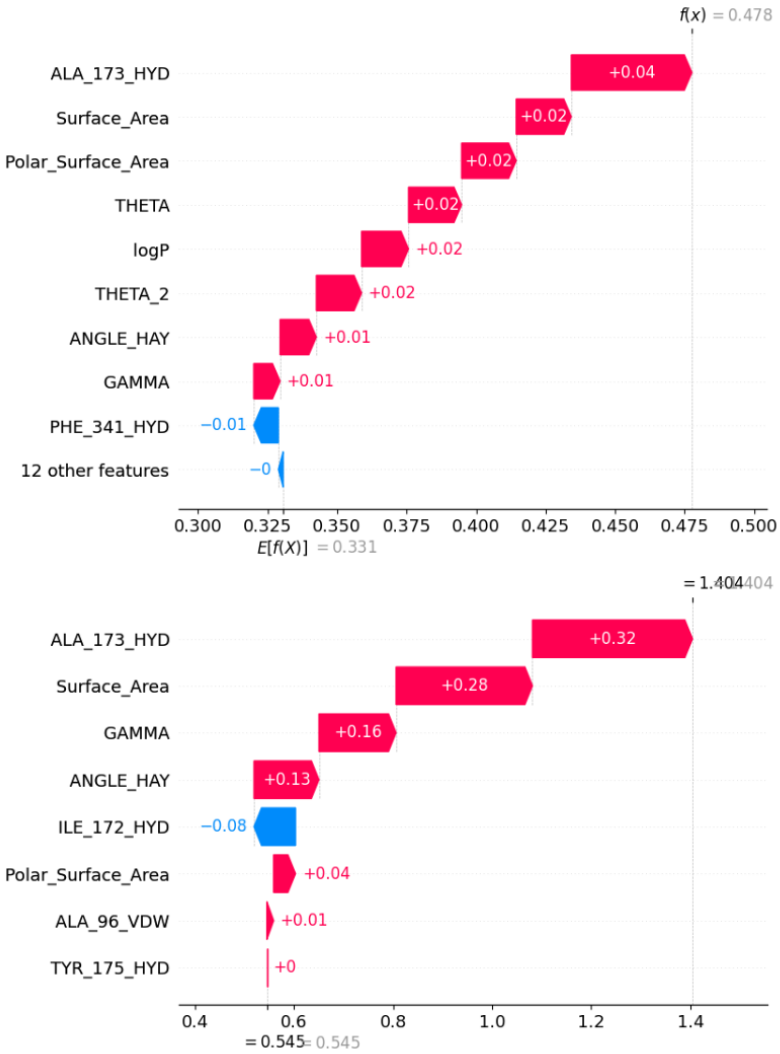


Image 48: Waterfall plots for "Paroxetine".

In *"Maprotiline"* both models failed to predict the actual *"NO BINDING"* class and misclassified it as a strong binder. For RF baseline *"E[f(X)]"* was 0.331 and it slightly increased to *"f(x)"*=0.371. *"logP"* and *"GAMMA"* guided this prediction, but mainly *"THETA"* opposed this outcome. Since the impact of all these features was small, the final output of the model was governed by a high percentage of uncertainty and that was the most probable reason for the wrong prediction. XGBoost had 3 features, like *"Surface_Area"*, *"ANGLE_HAY"*, and *"GAMMA"* that dragged the model towards class 2 with high values, while only the residues *"ILE_172_HYD"* and *"ALA_173_HYD"* contributed to the opposite direction. As a result, the final model output increased from 0.545 to 0.843 and this concludes that XGBoost classifier was confident in its incorrect prediction, in contrast to Random Forest (Image 49).

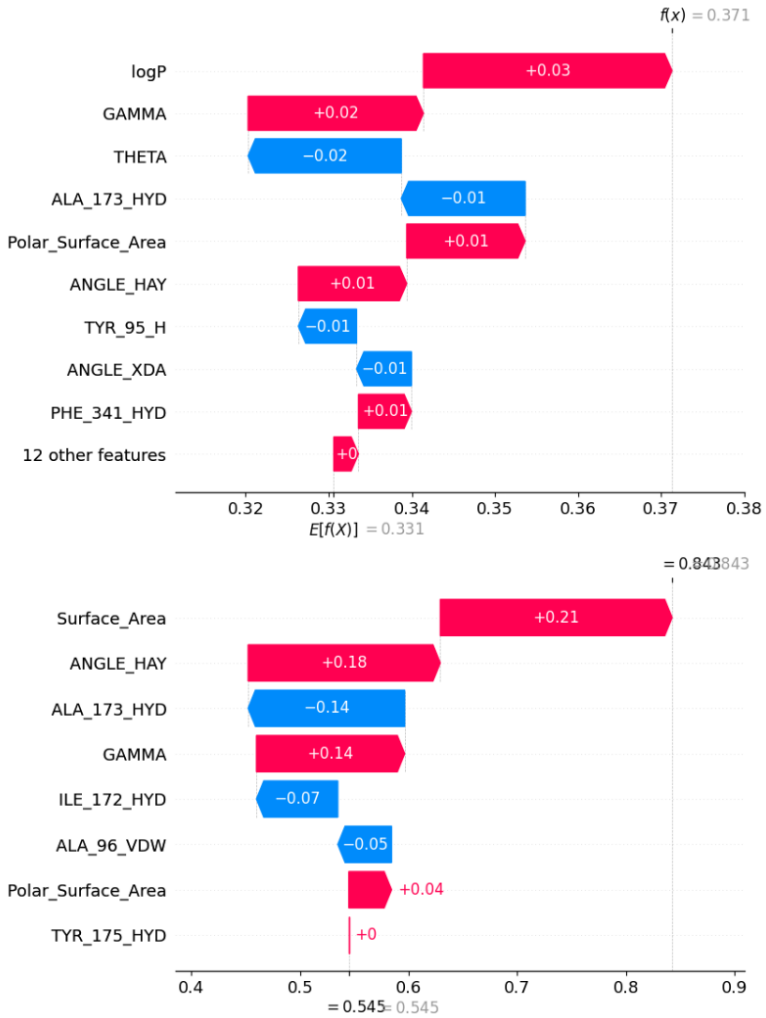


Image 49: Waterfall plots for "Maprotiline".

Another unexpected misclassification case is that of *"Trimipramine"*, which was incorrectly predicted by RF as a strong binder and as a non-binder by XGBoost, although its actual label is moderate. In Random Forest algorithm, the majority of features possessed relatively low SHAP contributions to either in favour of class 2 or against and that is the reason why the model's output was almost similar

to the baseline value. *"logP"* had the highest contribution though, with a small 0.04 value. In contrast, for XGBoost, *"ALA_96_VDW"* contributed with a positive value of 0.28 to a class 1 prediction, however, *"Polar_Surface_Area"* performed a significant contribution to the opposite direction with a value of -0.48. This led to an *"f(x)"*=0.348 from the baseline value *"E[f(X)]"* = 0.469, indicating high uncertainty of prediction towards class 1 (Image 50).

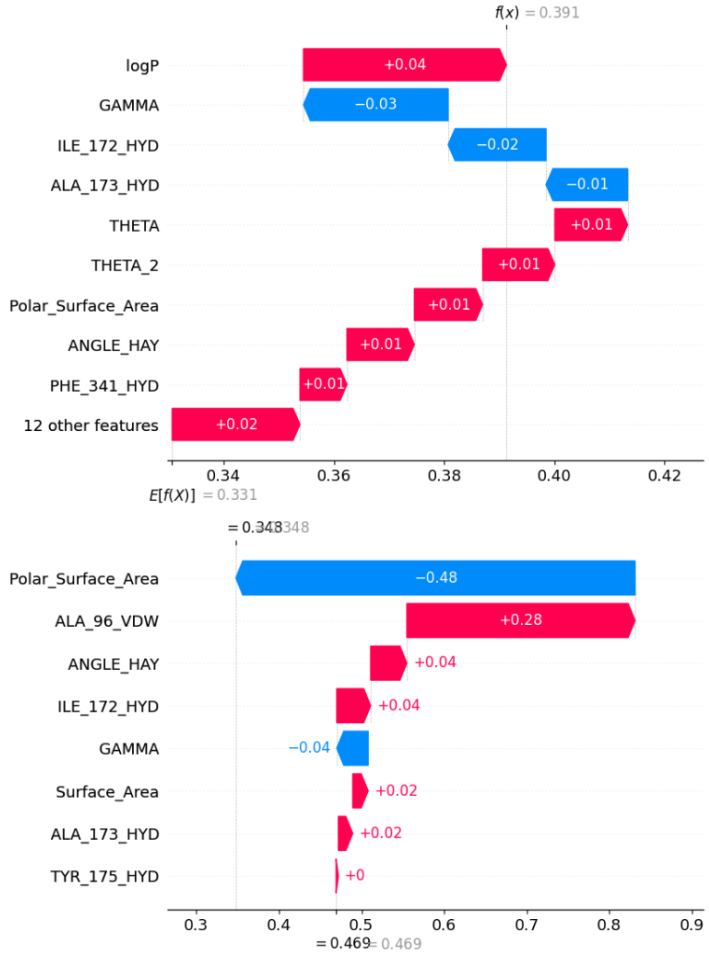


Image 50: Waterfall plots for "Trimipramine".

Last misclassified case that is discussed in the present thesis refers to *"Vilazodone"*. It was recently launched in the US in 2011 and it is a serotonin modulator with very strong inhibition of SERT protein transporter. However, the 2 models labelled it as a non-binder which is an alarming condition since it indicates minimal to zero interaction with human SERT protein transporter. RF showed moderate certainty with output value *"f(x)"*=0.41 from *"E[f(X)]"* = 0.334, while XGBoost exhibited strong confidence in the false prediction with an output value *"f(x)"*=0.83 from *"E[f(X)]"* = 0.469. Both models' decisions were driven mainly by *"Polar_Surface_Area"* feature, especially in XGBoost where it had a SHAP value of 0.9. To conclude, both models struggled to categorize *"Vilazodone"* in the *"STRONG BINDING"* class, because they relied heavily on one specific feature and that is the

reason why in the present thesis there are several misclassifications that result in low accuracy in nested cross validation approach (Image 51).

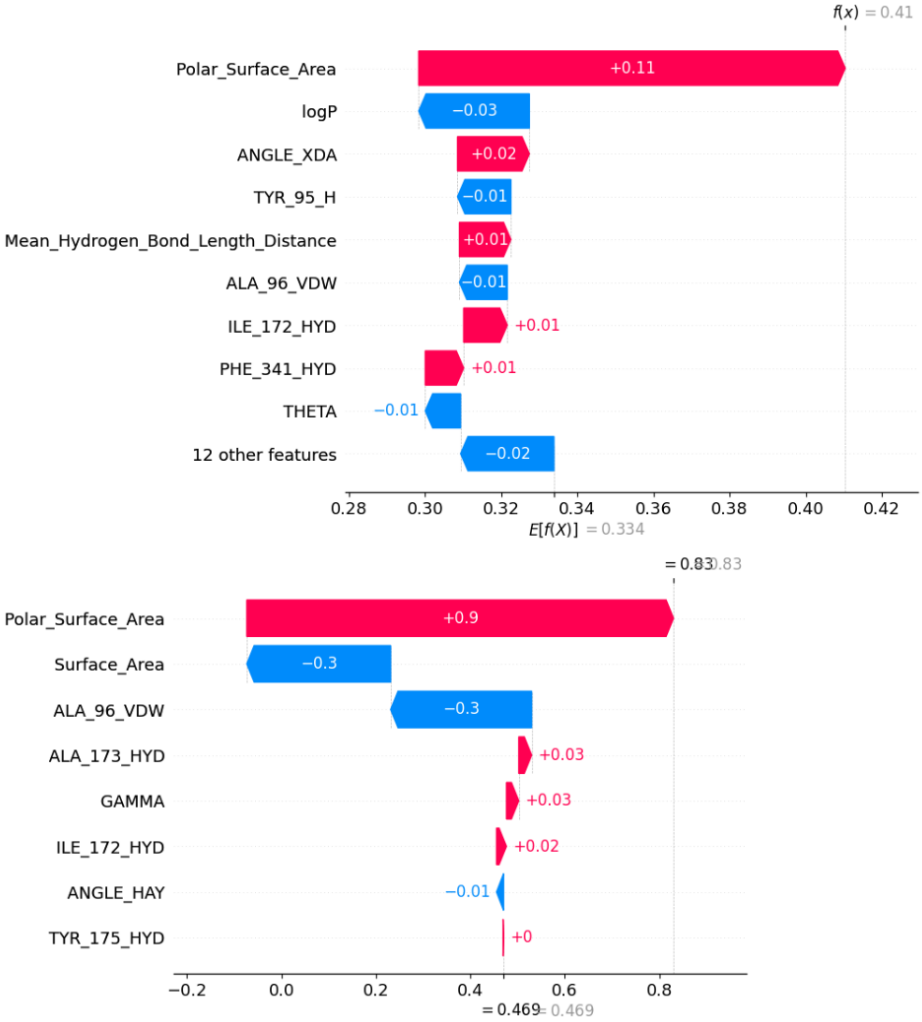


Image 51: Waterfall plots for "Vilazodone".

In summary, the misclassification analysis conducted reinforces the complexity of predicting ligand binding affinity, particularly for compounds positioned in the intermediate *"MODERATE BINDING"* class. The details of evaluation metrics insights, confusion matrices, SHAP interpretability and individual case studies mentioned in waterfall plots, illustrate that a large portion of the errors occurred at the boundaries between moderate and adjacent classes. Random Forest displayed more conservative predictions with smaller SHAP values, while XGBoost produced, in general, more confident outputs, even in wrong cases. This behaviour reflects the inherent architectural differences of how each model interprets feature space. *"Paroxetine"* and drugs like these facilitate the scientific community to identify which features are common in strong binders, since this compound is widely recognized as an SSRI with strong affinity with SERT protein. It highlighted specific features, such as *"ALA_173_HYD"* and *"Surface_Area"* that are marked as potential strong binding indicators. In contrast, cases like *"Maprotiline"* and *"Trimipramine"* revealed that even expected predictions can

contradict dataset labels, hinting at either biological nuance or inconsistencies in the ground truth. Lastly, *"Vilazodone"*, for which both models failed to detect its actual label, highlights the danger of over-reliance on dominant features with extreme SHAP values, such as *"Polar_Surface_Area"*. Ultimately, these findings underscore the limitations of the present thesis and how essential the explainable AI tools are in the drug discovery process, as they enable researchers not only to quantify possible errors but also modify an underdevelopment drug.

4.4.2 Partial Dependence Plot Interpretation

The last tool used to validate all the previous results was the generation of partial dependence plots (PDPs). These plots illustrate how the distribution of each feature affects the predicted probability for a given class label. To maintain clarity and avoid redundancy, the partial dependence analysis presented here focused exclusively on classes 2 and 0 which represent the strong and the moderate binders, since these are the most probable candidates for potential antidepressant activity. The features shown in the following graphs (Images 52 and 53) were the most influential based on the previous SHAP analysis, though there were many more from Random Forest that were not integrated in this PDP analysis. Most predictive include *"THETA"*, *"Polar_Surface_Area"*, *"Surface_Area"*, *"ANGLE_HAY"*, *"PHE_341_HYD"*, *"ALA_173_HYD"*, *"ILE_172_HYD"*, *"ALA_96_VDW"*, *"GAMMA"*, *"TYR_175_HYD"*, *"logP"* and *"Binding_Affinity_kcal/mole_"*. The following plots display the contribution of the robust features to the *"STRONG BINDING"* class. More specifically, for *"Polar_Surface_Area"* both models converged. Values above approximately 50Å² tend to reduce the probability for a given ligand to be labelled as class 2, suggesting that highly polar drugs cannot effectively inhibit SERT protein transporter. As for the angle *"THETA"*, which was presented only by Random Forest, it was clear that values between 20° to 40° increased the chances for a class 2 label and then the curve plateaued. *"ANGLE_HAY"* and *"GAMMA"* depicted steep drops in their curves in XGBoost after ~100° and ~50° respectively, while similar pattern was followed by RF, though with much smoother decline. Major rise was observed in *"Surface_Area"* after the value of ~300 Å², particularly in XGBoost, but the curve was stabilized and slightly decreased after the value of ~450Å². This indicates that the ideal contact area of a molecule should possess an intermediate condition, meaning that very small area is not a probable SERT inhibitor, but also very large areas may not fit properly in the binding pocket. In addition, *"logP"* was a positive contributor for a *"STRONG BINDING"* case after a value of 3 in Random Forest Classifier. Regarding the most critical residues involved in interactions with SERT protein based on the SHAP summary plots and robust features, the 2 models showed strong alignment. *"Alanine_173_HYD"* showed a rise in predicting class 2 with

values up to ~3 hydrophobic interactions. However, this consists the saturated point where the addition of extra bonds did not contribute to the models' ability to label a compound as a strong binder. *"Isoleucine_172_HYD"* which appeared over 500 times in the present dataset indicated that when it formed more than 10 bonds with a given ligand, representing the sum of the 5 individual poses, it was more likely to be labelled as strong binder. *"ALA_96_VDW"* performed a slight increase in probability of a strong binder in the range of 0 to 1 bonds and then curves in both models plateaued. *"PHE_341_HYD"* in RF had a plot that was mostly flat, but in the range of ~1-2 bonds a minor increase was observed in the probability of a class 2 occurrence. Interestingly, the binding affinity curve was mostly flat in RF, suggesting low discriminative power to classify strong binders in the current task. Lastly, *"TYR_175_HYD"* that was included only in robust features of XGBoost algorithm had a minimal decrease in strong binding probability between 0-1 bonds. Generally, it is substantial to note that the range on y-axis for RF was very narrow (0.30-0.40) indicating that the predictive power of the depicted features did not alter the output probability of class 2 to a great extent. Similar but less narrow (0.20-0.45) was the range for XGBoost algorithm. We can conclude then, that the binding affinity of a ligand to human SERT protein transporter depends on multiple factors and not one exclusively, making the binding process a difficult task to clarify.

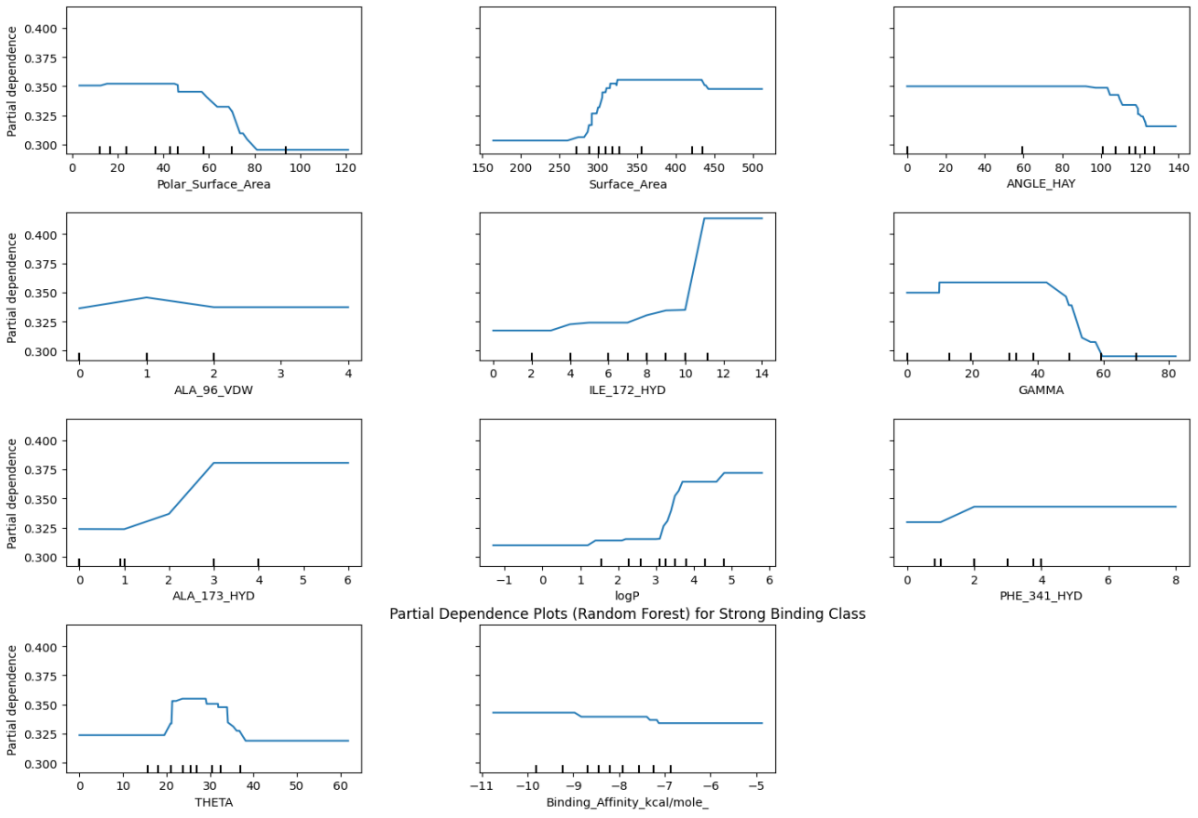


Image 52: Partial Dependence Plots for Random Forest for strong binders.

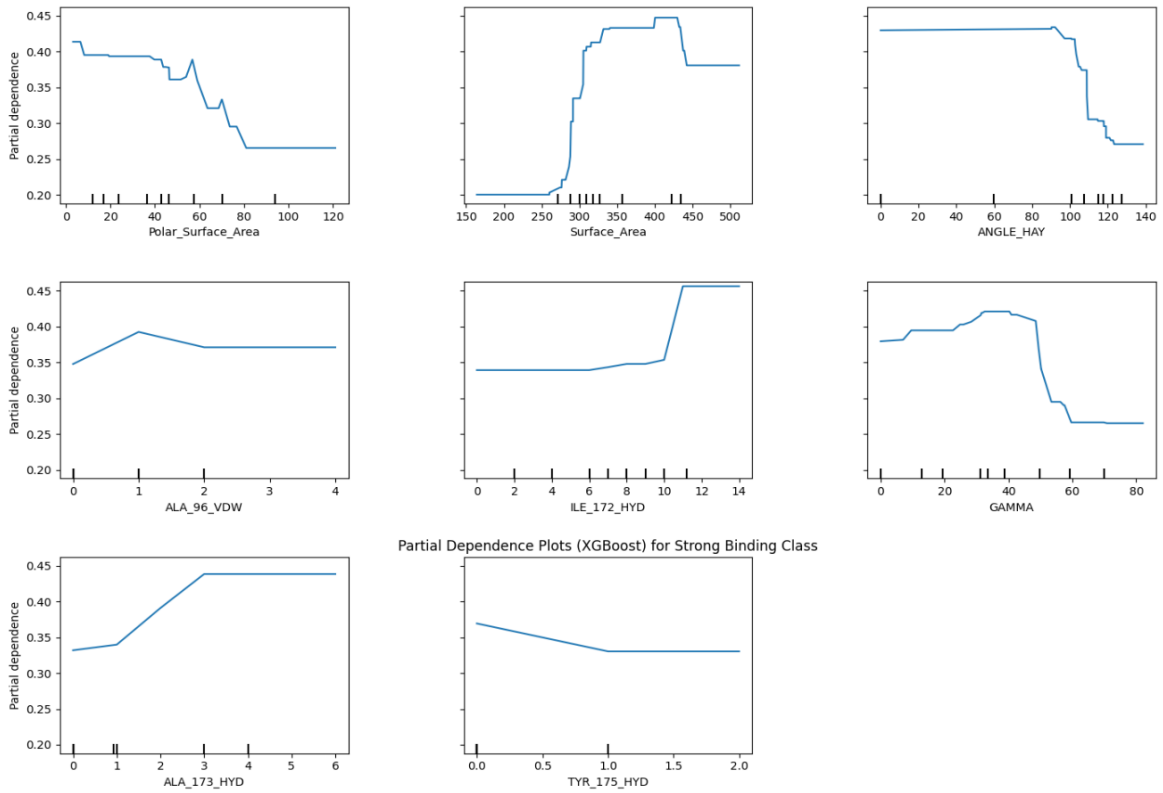


Image 53: Partial Dependence Plots for XGBoost for strong binders.

Images 54 and 55 represent the partial dependence plots across the two models for *"MODERATE BINDING"* class. Notably, *"Polar_Surface_Area"* aligned well with previous class 2 PDP plots as it demonstrated a steep decline after ~50 Å² in both models. This underlines that lower polarity favours both models, while higher values prevent a ligand from binding to SERT protein transporter. In contrast to class 2 PDP plots, *"ANGLE_HAY"* and *"GAMMA"* features showed increasing likelihood for *"MODERATE BINDING"* in the same ranges, where strong binding probability decreases. In addition, an opposite trend was observed for the *"Surface_Area"* feature, since in the range of ~300 Å²- ~450 Å² both models displayed a drop in their probability curves, whereas in the *"STRONG BINDING"* PDP plots a pronounced increase. As for *"THETA"*, which was depicted only in Random Forest Classifier, a slight decrease was observed around 20°–40°, opposite to the trend of strong binding class, indicating minimal discriminative power towards class 0. *"logP"* and *"Binding_Affinity_kcal/mole_"* had mostly flat curves for class 0 as image 55 revealed, indicating limited influence on predicting such compounds. Both models confirmed that for residues *"ALA_173_HYD"* and *"ILE_172_HYD"*, the behaviour was opposite to *"STRONG BINDING"* PDP plots. More specifically, for isoleucine the addition of more than 10 bonds weakened the model of predicting a potential moderate binder. The same was concluded for alanine, since the slope of the curve was negative. *"ALA_96_VDW"* exhibited positive contribution for moderate binders from 0 to 2 bonds and then the curves plateaued. *"PHE_341_HYD"* showed an initial slight drop between values 1-2 and then the curve flattened. Lastly, *"TYR_175_HYD"* increased the probability of a class 0

prediction in values 0 to 1 bonds, while in the same range the probabilities for a class 2 label were reduced as shown in image 53. Similarly to *"STRONG BINDING"* PDP plots, the ranges in y-axis were very narrow, which is correlated to minimal discriminative ability across the three binding classes.

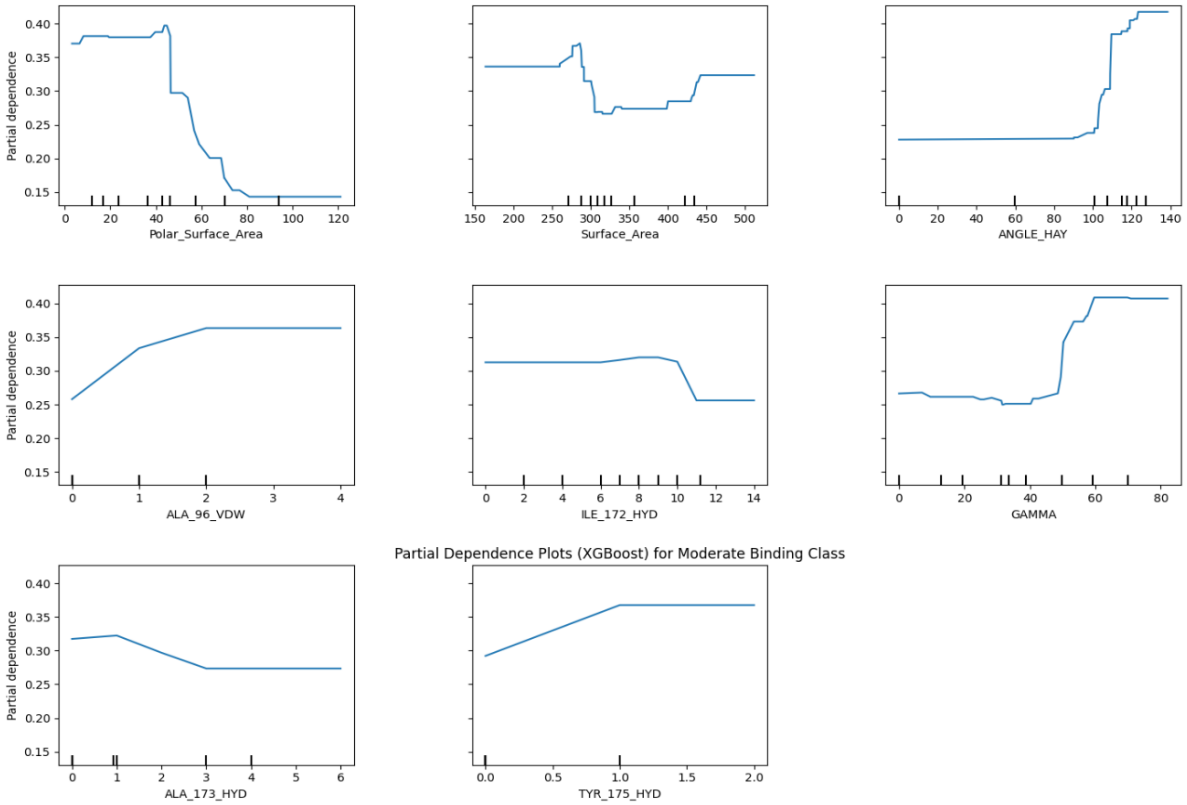


Image 54: Partial Dependence Plots for XGBoost for moderate binders.

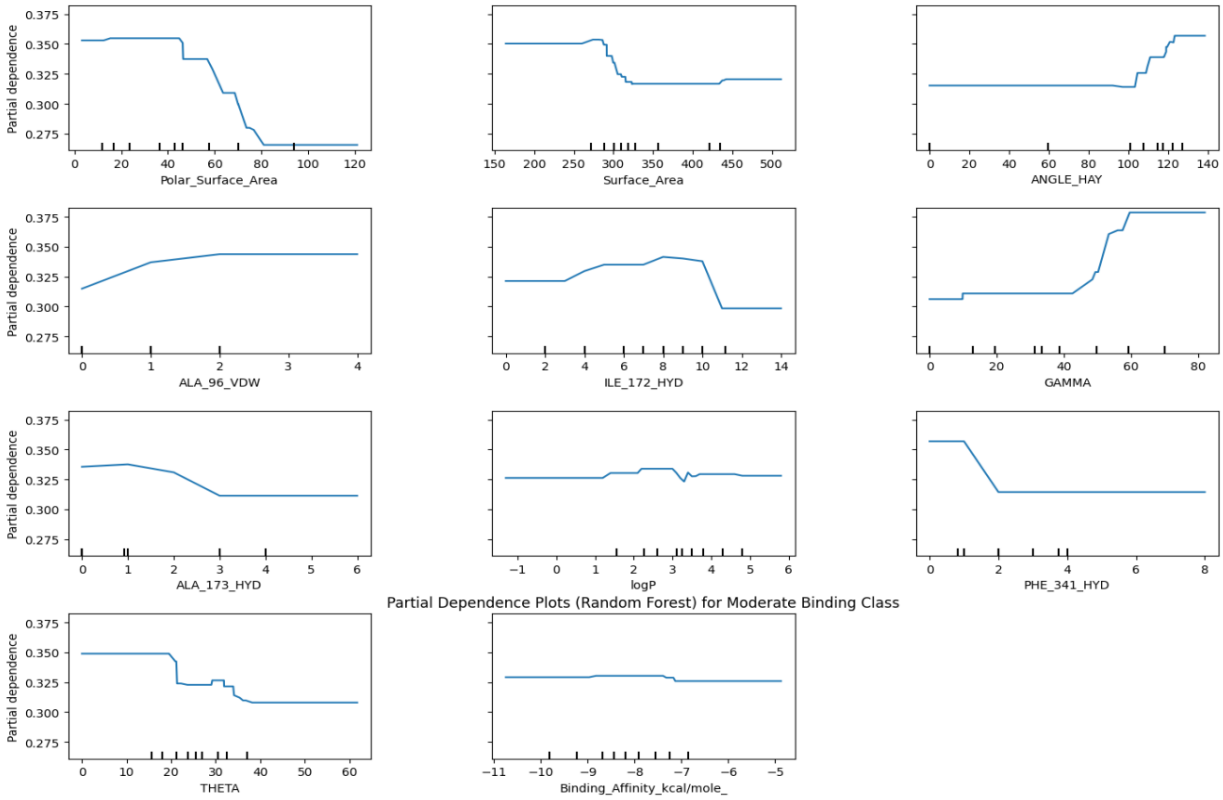


Image 55: Partial Dependence Plots for Random Forest for moderate binders.

In overall, the PDP analysis across both Random Forest and XGBoost models revealed that several SHAP-selected features, such as *"ALA_173_HYD"*, *"GAMMA"* and *"Surface_Area"*, were utilized by both models in predicting class 0 and class 2. However, their effects often appeared in opposite directions between the two classes, which aligns with the expected behaviour in a multiclassification task. Major exception was *"Polar_Surface_Area"*, which consistently decreased the probability of strong and moderate binding beyond ~50 Å², suggesting more distinct role in ligand-SERT binding process. Apart from these findings, the presence of flat and marginal curves highlighted the limited discriminative power for individual features, supporting the idea that SERT binding inhibition is governed by multiple factors or that all the assumptions and limitations of this thesis that are described in chapter 5.2 prevent the model from identifying potential discriminative patterns.

Chapter 5: Conclusion and Discussion

5.1 Findings Analysis

In the present study, an innovative and detailed approach was implemented that combined molecular docking simulations, extraction of molecular interaction descriptors, residue-level profiling and a machine learning pipeline to classify 74 ligands based on their affinity with human SERT protein transporter. *"AutoDock Vina"*, *"Chimera"* and *"BIOVIA Discovery Studio"* configured a final multidimensional dataset with over 150 features that captured different perspectives of the protein-ligand binding process including inherent molecular properties of ligands, physicochemical characteristics, residue interactions and other distance and geometry-based features.

The docking analysis conducted was first validated through the alignment with the 5I6X PDB structure (Image 32) and it showed high convergence. As expected, the KDE and box plots (Images 33-34, 36-38) illustrated that strong binders have the most negative binding affinities, followed by moderate binders and last the non-binders. However, significant overlapping was observed, which suggests that this feature, while vital for binding processes, was insufficient for effective binding classification of the 74 ligands with SERT protein. This was also confirmed by the fact that in Random Forest SHAP values and PDPs, the binding affinity did not show any discriminative power towards a specific class. The top ten poses were generated from the software and from them the best five were selected through a custom Python script that ranked them based on binding affinity and RMSD values.

Several machine learning models were tested and evaluated for their ability to classify ligands based on SERT inhibition into the three classes with a nested cross validation strategy. Among the algorithms, the most promising and robust one was Random Forest, with a mean training accuracy across folds ~84% and mean test accuracy 61.9%. The gap between train and test was acceptable, showing minimal overfitting possibly due to the small dataset size. However, it was still reliable and can possibly form the basis for future studies in this field. Second in rank, came the Voting Classifier which is an ensemble method that predicts based on the decisions of other known individual models. In the present thesis, for the Voting Classifier, Random Forest and XGBoost were integrated. It achieved mean train accuracy ~92% and mean test accuracy 60.4%. Notably, there was a high risk of overfitting since the difference between train and test was more than 30%. Another possible cause for this could be that the model memorizes the sample set instead of learning from them, again due to the small sample size. Finally, XGBoost performed moderately with 91.22% and 52.6% in train and test mean accuracy respectively, which was close to the Voting Classifier's behaviour. The remaining models underperformed. The ranking order based on the F1-score and precision was similar to the accuracy for the three models. Random Forest predicted correctly more strong binders followed by non-binders and lastly moderate binders, indicating that it can easier detect samples in the diametrically opposed categories, while in the intermediate class which is in a boundary condition

with the adjacent 2 classes, it underperformed. Voting classifier performed similarly to Random Forest, while XGBoost acted oppositely to the other two models by focusing mostly on the moderate to strong binding spectrum and underperformed in *"NO BINDING"* class. Apart from the evaluation metrics, the two models had as outputs the robust features involved in the nested cross validation technique, which represented features that were selected in 3 or more out of the 5 folds.

Several descriptors were examined statistically (Images 36-38) that reflect their distribution patterns across the three classes, while in combination with SHAP values and PDPs, an overall behaviour for each feature can be assessed. Notably, from SHAP summary plots it was clear that both Random Forest and XGBoost relied heavily on overlapping features with different weights. Across all features, *"Polar_Surface_Area"* was demonstrated as the top contributor in both models. Values exceeded ~50 Å² were linked to lower probabilities of strong and moderate binding cases, as the steep decline in PDPs revealed, highly compatible to the box plot depicted in Image 37. The interpretation from this lies on the fact that excess polarity atoms in ligands are less likely to interact with the hydrophobic part of the protein. To further confirm this interpretation, *"logP"*, which is a measure of hydrophobicity of a molecule, indicated that the probabilities for a strong binder case rose sharply in values more than 3, as shown in the PDPs (Image 52). For *"Surface_Area"*, there was a clear relationship among classes and that was, higher values tend to correlate easier to strong binders. This argument was enhanced by the increased discriminative power that this feature showed in the PDP curve, where around ~300-450 Å² performed a rise, highlighting simultaneously the optimal contact area for a potential strong SERT inhibitor. Box plot in image 36 confirmed these findings. As for the angular descriptors, most important and influential angles were *"GAMMA"* and "*ANGLE_HAY"* and last *"THETA"* which was only exhibited in Random Forest robust features. *"THETA"* supported strong binding to SERT protein between ~20° and 40°. This was compatible with the PDPs of moderate binders, where in the same range exhibited a decreasing curve and therefore lower probabilities. SHAP summary plots of the other two angles displayed positive contribution effect to class 0 and class 2. PDPs revealed decreasing probabilities for strong binders at values beyond ~50° in *"GAMMA"* and ~100° in *"ANGLE_HAY"*, while a reversed behavior was depicted in the *"MODERATE BINDING"* plots for the same values. Another important aspect of the present thesis was the residue-level analysis and which of them were decisive and observed in each class more frequently. Based on the robust features from the 2 main algorithms and their SHAP values, the most important residues include *"ALA_173_HYD"*, *"ILE_172_HYD"*, *"ALA_96_VDW"*, *"PHE_341_HYD"* and *"TYR_175_HYD"*. First 3 were demonstrated in the robust features of the two models applied, while *"PHE_341_HYD"* was found in RF and *"TYR_175_HYD"* in XGBoost algorithm and performed moderate SHAP values. The probability of strong binding increased in both models when *"Alanine_173_HYD"* participated in up to 3 hydrophobic interactions, while a reversed behaviour was performed in *"MODERATE BINDING"* as shown by the PDPs. *"ILE_172_HYD"*,

which was the most frequent residue in all classes with more than 500 records, enabled *"STRONG BINDING"* classification for number of bonds higher than 10, while at the same time in moderate binders PDPs had a negative slope. *"Alanine_96_VDW"* that is involved in van der Waals interactions was informative for moderate and non-binders based on SHAP summary plots. Higher probabilities for the *"MODERATE BINDING"* class, emerged when it was involved in 0-2 bonds. *"PHE_341_HYD"* was more critical in class 0 and class 1, however the probability of a strong binder sample was increased in the range of 1-2 bonds, while in moderate binders was reduced. Lastly, *"TYR_175_HYD"* proved to be significantly influential for class 0, in which the probabilities were increased when *"TYR_175_HYD"* formed 0-1 bonds. There were also additional features among those in Random Forest, for which a more comprehensive analysis was not conducted for space, however their SHAP summary plots and box plots added value to this thesis and can be integrated in future studies. Generally, both SHAP summary plots and PDPs demonstrated that the features extracted from Chimera, AutoDock Vina and BIOVIA Discovery Studio did not have enough predictive power to direct a models decision, instead they were more likely to show the tendency that these features followed. This statement was enhanced by the low SHAP values in the SHAP summary plots, especially in the Random Forest model and by the y-axis in all PDPs, where the ranges were narrow (~0.3-0.4 in RF and 0.15-0.45 in XGBoost). Apart from that, y-axis represents the average predicted probability of a ligand belonging to a given class as a function of a feature X. Notably the ranges are close to the random guess in a multiclassification task with 3 categories which is 0.33, since several curves were flat around this range.

Misclassification analysis was conducted to prove that both models underperformed due to limitations in the whole pipeline followed in the present thesis. Both models over-relied on some features and this led to wrong predictions. For example, in *"Vilazodone"* which is known as strong SSRI, both models failed to confirm this, and they labelled it as a non-binder due to its dominant value of *"Polar_Surface_Area"*, as image 51 illustrates. Similarly, *"Maprotiline"* was falsely labelled as strong binder instead of a non-binder, because both models over-relied on features like *"Surface_Area"*, *"logP"* and *"GAMMA"*. However, both models managed in cases such as *"Paroxetine"* to combine the contributions stemmed from molecular descriptors, like *"Surface_Area"*, *"Polar_Surface_Area"* and residues, like *"ALA_173_HYD"* and predict correctly the binding class. All this analysis, though very informative for ligand binding to SERT protein, it underlines the complexity and the multifactorial nature of SERT inhibition, which is difficult to clarify based solely on these tools.

The findings of the present study aligned well with prior works related to SERT inhibition and computational modeling of ligand binding mentioned in Chapter 2. Key residues such as *"ILE_172_HYD"* and *"PHE_341_HYD"*, which consistently appeared in this thesis as critical

discriminators for *"STRONG BINDING"* class, were also identified in experimental studies by Andersen et al. (2009) and Adejoro & Adewara (2025) confirming their key role in SERT inhibition. Nencetti et al., 2011 and Andersen et al. (2009) also highlight the essential role of residue *"Tyrosine 95"* that was among the robust features of Random Forest in SHAP summary plots. Apart from key residues, *"Polar_Surface_Area"*, is emphasized by Crampon et al. (2022) as a meaningful molecular descriptor for ligand binding affinity, compatible with the present study, in which PSA was a key discriminator especially for a *"NO BINDING"* class label. In addition Kong et al. (2019), achieved high precision using Random Forest and Voting Classifier models on SERT inhibitors by applying a binary classification task. The present thesis confirmed this tendency, although the evaluation metrics were much lower, possibly due to the three number of classes instead of two. Another common conclusion with the study (Crampon et al., 2022) is that the accuracy of molecular docking scores is questionable and for this reason an integrated machine learning pipeline is considered a more optimal solution for classifying ligands binding interactions.

5.2 Assumptions- Limitations and Challenges

The present thesis which incorporated several computational techniques, like docking simulations, extraction of specific molecular and residue details and construction of a supervised machine learning model to classify ligands based on their SERT inhibition, was governed by several assumptions, limitations and challenges that required careful consideration. Assumptions were grouped into 4 main parts:

Docking Environment Assumptions:

- Protein structure source: It was assumed that the 5I6X crystallization structure from RCSB Protein Data Bank was suitable for docking purposes.
- It was also assumed that protein and ligand preprocessing steps (removal of unwanted atoms and ligands, addition of hydrogen atoms, assigning of charges, energy minimization of structure and steric effects and dock preparation) were done correctly.
- It was assumed that sodium, chloride ions, water molecule and cholesterol did not disturb the correct execution of the docking simulations.
- Construction of Grid Box: The grid box dimensions were defined manually in Chimera software, considering that the initial structure with the bound paroxetine at the central site of the protein was correct. In addition, it was assumed that the box had the suitable dimensions which included all the central site of the protein that the present thesis focused on.

- Docking scores validity: While docking is widely used in the drug discovery community and it reveals approximate biological aspects, it does not reflect a realistic ligand-binding simulation, since the receptor was considered rigid, the ligand was partially flexible and the scoring functions simplify the overall interactions that occur without considering potential steric effects.
- Another assumption was based on the formulas used in 3.3 chapter for the evaluation of best 5 poses, where binding affinity and RMSD values contributed equally to these. However, another approach, for example an increase in the weight of binding affinity could lead to alternative 5 best poses and possibly better and more realistic results.

Feature Extraction & Representation Assumptions:

- The selection of the best 5 poses through the custom Python script relied on a combination of binding affinities and RMSD values. Nevertheless, there is uncertainty in the fact that these poses may not reflect possible realistic conformations during ligand-SERT binding or they may not be functionally effective.
- Molecular interaction details and residue-level profiling: Similar to docking validity, all the descriptors and residues were extracted from BIOVIA Discovery Studio which represents static snapshots of the ligand with the SERT transporter, though in reality these interactions emerged from dynamic environments under specific conditions, for example pH or temperature.
- Median Aggregation: During aggregation of molecular descriptors with the median value, it was assumed that the median captures the most representative interaction profile of a ligand, effectively down-weighting outliers while still assuming that all 5 poses added value to the overall binding behaviour.
- The strategy used for residues was summation, meaning that for the five poses, the values for each residue were summed in order to get one value for each residue and for each ligand.
- For each ligand pose, the mean value of geometric properties, including the angular features *"ANGLE DHA"*, *"ANGLE HAY"*, *"ANGLE XDA"*, *"ANGLE DAY"*, *"THETA"*, *"THETA 2"*, *"GAMMA"*, as well as the feature *"CLOSEST ATOM DISTANCE"* was computed. These values represent averaged structural features per pose, irrespective of the specific residues involved in the interactions and are computed across all interaction entries associated with that pose.
- In cases where an interaction (e.g., *Amide–Pi Stacked*) involved multiple residues the interaction was classified based on its type (in this case, as hydrophobic) and contributed to the hydrophobic interaction count for each of the residues involved. This ensured that multi-residue contributions were integrated in the final model.

- While in the docking process, all ligands that support the functional part of the protein were kept, such as ions of sodium, chloride, water and cholesterol, they were excluded from machine learning model, since it focused on amino acids. However, among these ligands, only Na^+ ion was present with some van der Waals interactions.
- The mean values of hydrogen and hydrophobic bonds length distances emerged from the average distances of the individual bonds that participated in the respective category. This approach was quite confusing, because detailed information was put aside, but it was easier to integrate such feature in a machine learning pipeline.
- In the other interactions feature, electrostatic bonds and halogen bonds were included, although their action is different.

Machine Learning Pipeline Assumptions:

- The machine learning models used (Random Forest, XGBoost, Voting Classifier) assume independence among input features, even though molecular descriptors are often correlated.
- Collinear features described in chapter 3.5.3 were maintained in tree-based models that are more robust in handling multicollinearity, while in Logistic Regression and Support Vector Machines were dropped from the training process.
- Nested cross validation framework was applied, since it is often a more ideal approach when the sample size is small.
- For LightGBM algorithm RFE was not applied due to increased demands of computational time.
- Slight variations in SVM AUC metrics across runs were assumed to result from internal randomness during probability estimation and data splitting. This could possibly be fixed with the probability parameter.
- Explainable AI tools used here assumed that each feature had an individual discriminative power. However, it is possible that some features act cooperatively with other both in machine learning models and in realistic SERT-ligand binding environments.
- The strategy for robust features that were obtained from nested cross validation in 3 or more outer folds, while logical, there could be another more effective solution of picking a subset of features.
- In explainability analysis the thesis focused on all robust features from XGBoost algorithm that were also displayed in Random forest, with the addition of a few more robust features from RF.
- The model focused equally to all 3 classes, although the strong and moderate binders are more probable SERT inhibitors.

Biological Interpretation & Validation Assumptions:

- Label Accuracy: The most important fact for the present thesis is the correct labelling of the 74 ligands. It was assumed that the Ki values for these ligands were valid from the widely known sources and also relatable. More specifically, while the radioligand binding assays from which the Ki values were extracted were not the same (radioligand, conditions of experiments etc.), they were considered suitable for the present thesis.
- It was assumed that the structural and biological information associated with the selected CID compounds from the recognized sources was accurate, relevant, and suitable for docking and classification task in the present study.
- It was assumed that the custom scripts about selection of best 5 out of 10 poses and *"Ligand_Distance_to_Grid_Box_Center"* feature were correct, although they are not mentioned in the following thesis for clarity and space limitations.
- During the extraction of 2D images from BIOVIA Discovery Studio, several unfavourable interactions were revealed with red circles. While these might be correlated to severe steric effects, the specific poses were not excluded from the dataset or penalized somehow. The majority of them belong to the *"NO BINDING"* class.
- The class separation was based solely on real Ki values with human SERT protein transporter and not based on the clinical antidepressant activity itself. This means that a moderate binder can be more effective as an antidepressant than a strong binder.
- It was assumed that all ligands act via the same mechanism and binding site within SERT, without accounting for allosteric or atypical modes of action, although in literature there are several articles that claim for allosteric mechanism of SERT protein.
- There was zero reference in the pharmacokinetics and ADMET for these ligands, despite their importance in ligand binding and drug-discovery processes.

Apart from the previous assumptions there were several limitations and challenges that need to be mentioned. These include:

- Limited Dataset: Notably, the sample size of 74 ligands is very small. In addition, it is a high dimensional dataset with over 150 features, so there was overfitting to some extent. This was proved by the fact that training accuracies in nested cross validation were much higher than test accuracies. The models showed poor generalization since they tended to "memorize" more the samples than learning the distinct patterns from them. Only Random Forest performed decently.
- Aggregation issues: While the aggregation is a highly acceptable technique, especially in small datasets like in the present thesis, it may confuse the model, because it may take into account a potential outlier case. For example, in this thesis, where the median values were

used in pose aggregations, there was high uncertainty of whether the chosen sample was representative or not for a specific ligand.

- All descriptors were extracted from static docking poses, where the SERT protein was rigid and the ligands were partially flexible without integrating molecular dynamics phenomena.
- The thesis focused on all available residues that emerged from the static reaction of each ligand with SERT protein. This means that there might be more residues from SERT that were not observed in the 370 poses of the 74 different ligands, but may play a vital role in SERT inhibition. Furthermore, not all residues fall up or close to the binding pocket of SERT, so it may be redundant to study them. Last but not least, there are residues that are useful for the ligand-binding process and others for the stabilization of the whole complex. The challenge that emerged is the fact that this thesis did not separate them based on their realistic role, but it considered them equally useful for binding.
- Collinear features may be present in the thesis, since only the features related to the 4 types of bonds have been removed, because they were the sum of the individual residue bonds for that specific type. Collinear features may confuse the model and underperform, especially Logistic Regression and SVM that are not tree-based models.
- Lack of external validation set: The models were evaluated via internal nested cross-validation, but no fully independent external test validation was used, because it would show potential overfitting.
- The tuning of hyperparameters was limited due to low computational resources. Extended search for the optimal hyperparameters could benefit the model and increase evaluation metrics.
- SHAP interpretability and PDPs considered the power of features individually, without any synergistic effects.

5.3 Future Recommendations and Enhances

Building on the findings and limitations of the present thesis, future research could progress in several aspects. First and mandatory, is the expanding of the dataset with a larger number of ligands, more than 1000 ligands, with experimentally validated Ki values. By this, generalizability will be enhanced and more advanced models could be implemented. An innovative approach would be to integrate in docking simulations, apart from the serotonin transporter (SERT), other related monoamine transporters, such as the norepinephrine transporter (NET) and dopamine transporter (DAT), which are primarily targets for SNRIs and TCAs, resulting in the creation of poly-pharmacological models that will have the ability to distinguish broad-spectrum antidepressant profiles. The docking scores of all

tested compounds with the three receptors and the extraction of all molecular descriptors and residues will facilitate the classification task that is applied in this thesis. Additionally, the incorporation of molecular dynamics simulations is essential, since it captures protein flexibility and reflects a more realistic in-vivo environment. From docking and molecular dynamics simulations, much more valid features will emerge that will be useful for possible machine learning model. Another proposal would be to integrate molecular fingerprints, which encode atom-level substructures and represent detailed chemical structures that are more realistic than the traditional descriptors used in the present thesis. Furthermore, the inclusion of pharmacokinetic and ADMET (absorption, distribution, metabolism, excretion and toxicity) data, would bridge the gap between binding affinity, efficacy and clinical condition. Finally, the creation of hybrid models or advanced models, like Graph neural networks (GNNs) may capture more distinct patterns of the ligands and classify them, handling the overlapping of classes more effectively.

BIBLIOGRAPHY

1.11. Ensemble methods — scikit-learn 0.22.1 documentation. (2012). Scikit-Learn.org. Available at: https://scikit-learn.org/stable/modules/ensemble.html#forest

1.13. Feature selection. (n.d.). Scikit-Learn.
Availabe at: https://scikit-learn.org/stable/modules/feature_selection.html#rfe

A. Reyes-Chaparro, Flores-Lopez, N. S., F. Quintanilla-Guerrero, Nicolás-Álvarez, D. E., & Hernandez-Martinez, A. R. (2023). *Design of new reversible and selective inhibitors of monoamine oxidase A and a comparison with drugs already approved*. Bulletin of the National Research Centre/Bulletin of the National Research Center, 47(1). (IMAGE 5)

Adejoro, I. A., Adewara, I. J., Babatunde, D., & Johnson. (2025). *Predicting Selective Serotonin Re-Uptake Inhibitors Potency: Machine Learning and Molecular Docking Approach*. Research Square (Research Square).

Aghajani, J., Poopak Farnia, Parissa Farnia, Jalaledin Ghanavi, & Velayati, A. A. (2022). *Molecular Dynamic Simulations and Molecular Docking as a Potential Way for Designed New Inhibitor Drug without Resistance*. Tanaffos, 21(1), 1.

Agu, P. C., Afiukwa, C. A., Orji, O. U., Ezeh, E. M., Ofoke, I. H., Ogbu, C. O., Ugwuja, E. I., & Aja, P. M. (2023). *Molecular docking as a tool for the discovery of molecular targets of nutraceuticals in diseases management*. Scientific Reports, 13(1).

Andersen, J., Olsen, L., Hansen, K. B., Taboureau, O., Jørgensen, F. S., Jørgensen, A. M., Bang-Andersen, B., Egebjerg, J., Strømgaard, K., & Kristensen, A. S. (2010). *Mutational Mapping and Modeling of the Binding Site for (S)-Citalopram in the Human Serotonin Transporter**. Journal of Biological Chemistry, 285(3), 2051–2063.

Author, N. (2019). *Biolayer Interferometry and Surface Plasmon Resonance Comparison*. Nicoya. Nicoya - Improving Human Life by Helping Scientists Succeed. Available at: https://nicoyalife.com/blog/biolayer-interferometry-vs-surface-plasmon-resonance/

Ballester, P. J., & Mitchell, J. B. O. (2010). *A machine learning approach to predicting protein–ligand binding affinity with applications to molecular docking*. Bioinformatics, 26(9), 1169–1175.

Bank, R. P. D. (n.d.). *RCSB PDB - 5I6X: X-ray structure of the ts3 human serotonin transporter complexed with paroxetine at the central site*. Www.rcsb.org. Available at: https://www.rcsb.org/structure/5I6X

Barbier, A. J., Aluisio, L., Lord, B., Qu, Y., Wilson, S. J., Boggs, J. D., Bonaventure, P., Miller, K., Fraser, I., Dvorak, L., Pudiak, C., Dugovic, C., Shelton, J., Mazur, C., Letavic, M. A., Carruthers, N. I., & Lovenberg, T. W. (2007). *Pharmacological characterization of JNJ-28583867, a histamine H3 receptor antagonist and serotonin reuptake inhibitor*. European Journal of Pharmacology, *576*(1-3), 43–54.

BioinformaticsCopilot. (2024, August 31). *Fixing Missing Atoms in PDB file using Swiss PDB Viewer*. YouTube. https://www.youtube.com/watch?v=ixKVqmtexQE

Bjørn Olav Brandsdal, Fredrik Österberg, Almlöf, M., Feierberg, I., Luzhkov, V. B., & Johan Åqvist. (2003). *Free Energy Calculations and Ligand Binding*. 123–158.

Butt, S. S., Badshah, Y., Shabbir, M., & Rafiq, M. (2020). *Molecular Docking Using Chimera and Autodock Vina Software for Nonbioinformaticians*. JMIR Bioinformatics and Biotechnology, 1(1), e14232.

Canadian Society of Pharmacology and Therapeutics (CSPT). *Inhibitory constant (Ki)*. (n.d.). Pharmacologycanada.org. Available at: https://pharmacologycanada.org/Inhibitory-constant-ki

Chauhan, A. (2021). *Random Forest Classifier and its Hyperparameters*. Analytics Vidhya. Available at: https://medium.com/analytics-vidhya/random-forest-classifier-and-its-hyperparameters-8467bec755f6

Chen, T., & Guestrin, C. (2016). *XGBoost: a Scalable Tree Boosting System*. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16, 1(1), 785–794.

Christoph Molnar (2019). *Interpretable Machine Learning*. Github.io. Available at: https://christophm.github.io/interpretable-ml-book/.

Chu, A., & Wadhwa, R. (2023). *Selective Serotonin Reuptake Inhibitors*. PubMed; StatPearls Publishing. https://www.ncbi.nlm.nih.gov/books/NBK554406/

Cleveland Clinic. (2023). *Antidepressants*. Cleveland Clinic; Cleveland Clinic. Available at: https://my.clevelandclinic.org/health/treatments/9301-antidepressants-depression-medication

Cleveland Clinic. (2023). *Nervous system: What It is, types, Symptoms*. Cleveland Clinic; Cleveland Clinic. Available at: https://my.clevelandclinic.org/health/body/21202-nervous-system

Cleveland Clinic. (2023). *Treatment-resistant depression*. Cleveland Clinic. Available at: https://my.clevelandclinic.org/health/diseases/24991-treatment-resistant-depression

Coleman, J. A., Green, E. M., & Gouaux, E. (2016). *X-ray structures and mechanism of the human serotonin transporter*. Nature, 532(7599), 334–339.

Cortes, C., & Vapnik, V. (1995). *Support-vector networks*. Machine Learning, 20(3), 273–297.

courses.lumenlearning.com. (n.d.). *Protein Structure | Biology for Non-Majors I*. [online] Available at: https://courses.lumenlearning.com/wm-nmbiology1/chapter/reading-protein-structure/.

courses.lumenlearning.com. (n.d.). *Protein Structure | Biology for Non-Majors I*. [online] Available at: https://courses.lumenlearning.com/wm-nmbiology1/chapter/reading-protein-structure/.

Crampon, K., Giorkallos, A., Deldossi, M., Baud, S., & Steffenel, L. A. (2021). *Machine-learning methods for ligand–protein molecular docking*. Drug Discovery Today, 27(1).

Delgado, P.L. (2000). Depression: The Case for a Monoamine Deficiency. The Journal of clinical psychiatry. Available at: https://pubmed.ncbi.nlm.nih.gov/10775018/.

Duman, R. S., & Aghajanian, G. K. (2012). *Synaptic Dysfunction in Depression: Potential Therapeutic Targets*. Science, 338(6103), 68–72.

Eberhardt, J., Santos-Martins, D., Tillack, A. F., & Forli, S. (2021). *AutoDock Vina 1.2.0: New Docking Methods, Expanded Force Field, and Python Bindings*. Journal of Chemical Information and Modeling, 61(8).

eli. (2024). *What Are The Restrictions Of Elisa Tests? Unlocking The Enigma*. Merkel Technologies Ltd. Merkel Technologies Ltd. Available at: https://merkel.co.il/what-are-the-restrictions-of-elisa-tests-unlocking-the-enigma/

Fan, J., Fu, A., & Zhang, L. (2019). *Progress in molecular docking*. Quantitative Biology, 7(2), 83–89.

Features — LightGBM 4.6.0 documentation. (2017). Readthedocs.io. Available at: https://lightgbm.readthedocs.io/en/stable/Features.html#references

Gasteiger, J., & Marsili, M. (1980). *Iterative partial equalization of orbital electronegativity—a rapid access to atomic charges*. Tetrahedron, *36*(22), 3219–3228.

GeeksforGeeks. (2017). *Logistic Regression in Machine Learning*. GeeksforGeeks. https://www.geeksforgeeks.org/machine-learning/understanding-logistic-regression/

Geeksforgeeks. (2024). *Understanding Logistic Regression*. GeeksforGeeks. Available at: https://www.geeksforgeeks.org/understanding-logistic-regression/

GeeksforGeeks. (2021). *Support Vector Machine (SVM) Algorithm*. GeeksforGeeks. https://www.geeksforgeeks.org/machine-learning/support-vector-machine-algorithm/

Glossary. (2024). Python Documentation. Available at: https://docs.python.org/3/glossary.html#term-hashable

Guedes, I. A., Pereira, F. S. S., & Dardenne, L. E. (2018). *Empirical Scoring Functions for Structure-Based Virtual Screening: Applications, Critical Aspects, and Challenges*. Frontiers in Pharmacology, 9.

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., & Gérard-Marchant, P. (2020). *Array Programming with NumPy.* Nature, 585(7825), 357–362.

Hillhouse, T. M., & Porter, J. H. (2015). *A brief history of the development of antidepressant drugs: From monoamines to glutamate*. Experimental and Clinical Psychopharmacology, 23(1), 1–21.

Jankovics, H., Kovacs, B., Saftics, A., Gerecsei, T., Tóth, É., Szekacs, I., Vonderviszt, F., & Horvath, R. (2020). *Grating-coupled interferometry reveals binding kinetics and affinities of Ni ions to genetically engineered protein layers*. Scientific Reports, 10(1).

Jarmoskaite, I., AlSadhan, I., Vaidyanathan, P. P., & Herschlag, D. (2020). *How to measure and evaluate binding affinities*. ELife, 9.

Karlee Jenna Hall, Karen Van Ooteghem, & McIlroy, W. E. (2023). *Emotional state as a modulator of autonomic and somatic nervous system activity in postural control: a review*. Frontiers in Neurology, 14

Kong, W., Wang, W., & An, J. (2019). *Prediction of 5-hydroxytryptamine Transporter Inhibitor based on Machine Learning*. ArXiv.org. Available at: https://arxiv.org/abs/1910.14360

Laban, T. S., & Saadabadi, A. (2023). *Monoamine Oxidase Inhibitors (MAOI)*. Nih.gov; StatPearls Publishing.

Lemaître, G., Nogueira, F., & Aridas, C. K. (2017). *Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning*. Journal of Machine Learning Research, 18(17), 1–5.

Lodish, H.F et al. (2016). *Molecular Cell Biology*. 8th ed. New York: W.H. Freeman-Macmillan Learning.

Longone, P. (2011). *Neurosteroids as neuromodulators in the treatment of anxiety disorders*. Frontiers in Endocrinology, *2*.

Lundberg, S.M. and Lee, S.-I. (2017). *A Unified Approach to Interpreting Model Predictions*. Neural Information Processing Systems. Available at: https://papers.nips.cc/paper_files/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html.

Ly, K. S., Letavic, M. A., Keith, J. M., Miller, J. M., Stocking, E. M., Barbier, A. J., Bonaventure, P., Lord, B., Jiang, X., Boggs, J. D., Dvorak, L., Miller, K. L., Nepomuceno, D., Wilson, S. J., & Carruthers, N. I. (2008*). Synthesis and biological activity of piperazine and diazepane amides that are histamine H3 antagonists and serotonin reuptake inhibitors*. Bioorganic & Medicinal Chemistry Letters, *18*(1), 39–43.

Mattson, R. J., Catt, J. D., Denhart, D. J., Deskus, J. A., Ditta, J. L., Higgins, M. A., Marcin, L. R., Sloan, C. P., Beno, B. R., Gao, Q., Cunningham, M. A., Mattson, G. K., Molski, T. F., Taber, M. T., & Lodge, N. J. (2005*). Conformationally Restricted Homotryptamines. 2. Indole Cyclopropylmethylamines as Selective Serotonin Reuptake Inhibitors*. Journal of Medicinal Chemistry, *48*(19), 6023–6034.

Mayo Clinic. (2024). *Selective serotonin reuptake inhibitors (SSRIs)*. Mayo Clinic. Available at: https://www.mayoclinic.org/diseases-conditions/depression/in-depth/ssris/art-20044825

McKinney, W. (2010). *Data Structures for Statistical Computing in Python*. Proceedings of the 9th Python in Science Conference, 445.

Miller, A. H., & Raison, C. L. (2015). *The role of inflammation in depression: from evolutionary imperative to modern treatment target*. Nature Reviews Immunology, 16(1), 22–34.

Miranda, M., Morici, J. F., Zanoni, M. B., & Bekinschtein, P. (2019). *Brain-Derived Neurotrophic Factor: A Key Molecule for Memory in the Healthy and the Pathological Brain*. Frontiers in Cellular Neuroscience, 13(363).

MMFF94 Force Field (mmff94) — *Open Babel 3.0.1 documentation*. (n.d.). Open-Babel.readthedocs.io. Available at https://openbabel.readthedocs.io/en/latest/Forcefields/mmff94.html

Mohanty, M., & Mohanty, P. S. (2023). *Molecular docking in organic, inorganic, and hybrid systems: a tutorial review*. https://doi.org/10.1007/s00706-023-03076-1

Mohd Mursal, Ahmad, M., Hussain, S., & Mohemmed Faraz Khan. (2024). *Navigating the Computational Seas: A Comprehensive Overview of Molecular Docking Software in Drug Discovery*. IntechOpen EBooks.

Moraczewski, J., & Aedma, K. K. (2022). *Tricyclic Antidepressants.* PubMed; StatPearls Publishing.

National Institute of Neurological Disorders and Stroke. (2023). *Brain Basics: Know Your Brain*. National Institute of Neurological Disorders and Stroke. Www.ninds.nih.gov. Available at: https://www.ninds.nih.gov/health-information/public-education/brain-basics/brain-basics-know-your-brain (IMAGE 6)

National Library of Medicine. (2020). *Commonly prescribed antidepressants and how they work*. NIH MedlinePlus Magazine. Available at: https://magazine.medlineplus.gov/article/commonly-prescribed-antidepressants-and-how-they-work

Nencetti, S., Mazzoni, M. R., Ortore, G., Lapucci, A., Giuntini, J., Orlandini, E., Banti, I., Nuti, E., Lucacchini, A., Giannaccini, G., & Rossello, A. (2010). *Synthesis, molecular docking and binding studies of selective serotonin transporter inhibitors*. European Journal of Medicinal Chemistry, 46(3), 825–834.

Neurotorium. (n.d.). *Serotonin and Noradrenaline Re-Uptake Inhibitors (SNRIs)*. Available at: https://neurotorium.org/image/serotonin-and-noradrenaline-re-uptake-inhibitors-snris-2/. (IMAGE 3)

Neurotorium. (n.d.). *Tricyclic Antidepressants (TCAs)*. Available at: https://neurotorium.org/image/tricyclic-antidepressants-tcas/. (IMAGE 4)

Ouazana-Vedrines, C., Lesuffleur, T., Cuerq, A., Fagot-Campagna, A., Rachas, A., Gastaldi-Ménager, C., Hoertel, N., Limosin, F., Lemogne, C., & Tuppin, P. (2022). *Outcomes associated with antidepressant treatment according to the number of prescriptions and treatment changes: 5-year follow-up of a nation-wide cohort study*. Frontiers in Psychiatry, 13.

Owens, J. M., Knight, D. L., & Nemeroff, C. B. (2002). *Second generation SSRIS: human monoamine transporter binding profile of escitalopram and R-fluoxetine*. L'Encephale, 28(4), 350–355.

O'Boyle, N. M., Banck, M., James, C. A., Morley, C., Vandermeersch, T., & Hutchison, G. R. (2011). *Open Babel: An open chemical toolbox*. Journal of Cheminformatics, 3(1).

Palacios-Ortega, J., Rivera-de-Torre, E., Gavilanes, J. G., J. Peter Slotte, Álvaro Martínez-del-Pozo, & García-Linares, S. (2021). *Biophysical approaches to study actinoporin-lipid interactions*. Methods in Enzymology on CD-ROM/Methods in Enzymology, 307–339.

Panalytical, M. (2019). *Binding Affinity*. Dissociation Constant. Malvernpanalytical.com.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). *Scikit-learn: Machine Learning in Python*. Journal of Machine Learning Research, 12(85), 2825–2830. Available at: https://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html

Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., & Ferrin, T. E. (2004). *UCSF Chimera--a visualization system for exploratory research and analysis*. Journal of Computational Chemistry, 25(13), 1605–1612.

Physiopedia. (n.d.). *Serotonin Syndrome*. Available at: https://www.physiopedia.com/Serotonin_Syndrome. (IMAGE 1)

Plenge, P., Abramyan, A. M., Sørensen, G., Mørk, A., Weikop, P., Gether, U., Bang-Andersen, B., Shi, L., & Loland, C. J. (2020). *The mechanism of a high-affinity allosteric inhibitor of the serotonin transporter*. Nature Communications, 11(1), 1491.

Python. (2009). *re — Regular expression operations — Python 3.7.2 documentation*. Python.org. Available at: https://docs.python.org/3/library/re.html

Rokach, L. (2009). *Ensemble-based classifiers*. Artificial Intelligence Review, 33(1-2), 1–39.

Rotamers. (2025). Ucsf.edu. https://www.cgl.ucsf.edu/chimera/docs/ContributedRotamers. (2025). Ucsf.edu. https://www.cgl.ucsf.edu/chimera/docs/ContributedSoftware/rotamers/framerot.html

Rout, A. R. (2023). *Advantages and Disadvantages of Logistic Regression*. GeeksforGeeks. Available at: https://www.geeksforgeeks.org/advantages-and-disadvantages-of-logistic-regression/

Sanacora, G., Treccani, G., & Popoli, M. (2012). *Towards a glutamate hypothesis of depression*. Neuropharmacology, 62(1), 63–77.

Sansone, R. A., & Sansone, L. A. (2014). *Serotonin Norepinephrine Reuptake Inhibitors: A Pharmacological Comparison*. Innovations in Clinical Neuroscience, 11(3-4), 37.

Sasidharan, A. (2021). *Support Vector Machine Algorithm*. GeeksforGeeks. Available at: https://www.geeksforgeeks.org/support-vector-machine-algorithm/

Schmitz, W. D., Denhart, D. J., Brenner, A. B., Ditta, J. L., Mattson, R. J., Mattson, G. K., Molski, T. F., & Macor, J. E. (2005). *Homotryptamines as potent and selective serotonin reuptake inhibitors (SSRIs)*. Bioorganic & Medicinal Chemistry Letters, 15(6), 1619–1621.

Scikit-learn. (2009). *3.2.4.3.5. sklearn.ensemble.GradientBoostingClassifier — scikit-learn 0.20.3 documentation*. Scikit-Learn.org. Available at: https://scikitlearn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingClassifier.html

scikit-learn. (2014). *sklearn.linear_model.LogisticRegression — scikit-learn 0.21.2 documentation*. Scikit-Learn.org. Available at: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

scikit-learn. (2019). *sklearn.metrics.f1_score — scikit-learn 0.21.2 documentation*. Scikit-Learn.org. Available at: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html

scikit-learn. (2019). *sklearn.svm.SVC — scikit-learn 0.22 documentation*. Scikit-Learn.org. Available at: https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html

Scikit-Learn. (2025). *sklearn.ensemble.RandomForestClassifier — scikit-learn 0.20.3 Documentation*. Scikit-Learn.org. Available at: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

scikit-learn. (n.d.). *sklearn.ensemble.VotingClassifier — scikit-learn 0.24.0 Documentation*. Scikit-Learn.org. Available at: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.VotingClassifier.html

Sharma, S., & Dang, S. (2022). *Molecular Docking Analysis of Natural Compounds Against Serotonin Transporter (SERT)*. Current Trends in Biotechnology and Pharmacy, 15(6), 83–89.

Society for Neuroscience. (2018). *A PRIMER ON THE BRAIN AND NERVOUS SYSTEM*. A Companion Publication to BrainFacts.org. Available at: https://www.brainfacts.org/-/media/Brainfacts2/BrainFacts-Book/Brain-Facts-PDF-with-links.pdf

Staff, C. (2024). *What Are the Advantages and Disadvantages of Random Forest?* Coursera. Available at: https://www.coursera.org/articles/advantages-and-disadvantages-of-random-forest

Surbhi S. & Shweta D. (2021). *Molecular Docking Analysis of Natural Compounds Against Serotonin Transporter (SERT).* Current Trends in Biotechnology and PharmacyVol. 15 (6) 83 - 89, 2021.

Takeuchi, K., Kohn, T. J., Honigschmidt, N. A., Rocco, V. P., Spinazze, P. G., Hemrick-Luecke, S. K., Thompson, L. K., Evans, D. C., Rasmussen, K., Koger, D., Lodge, D., Martin, L. J., Shaw, J., Threlkeld, P. G., & Wong, D. T. (2006). *Advances toward new antidepressants beyond SSRIs: 1-aryloxy-3-piperidinylpropan-2-ols with dual 5-HT1A receptor antagonism/SSRI activities. Part 5*. Bioorganic & Medicinal Chemistry Letters, 16(9), 2347–2351.

Tan, H.-E. (2023). *The microbiota-gut-brain axis in stress and depression*. Frontiers in Neuroscience, 17.

Taylor, C. P., Traynelis, S. F., Siffert, J., Pope, L. E., & Matsumoto, R. R. (2016). *Pharmacology of dextromethorphan: Relevance to dextromethorphan/quinidine (Nuedexta®) clinical use*. Pharmacology & Therapeutics, 164, 170–182.

Torres, P. H. M., Sodero, A. C. R., Jofily, P., & Silva-Jr, F. P. (2019). *Key Topics in Molecular Docking for Drug Design*. International Journal of Molecular Sciences, 20(18), 4574.

Trott, O., & Olson, A. J. (2009). *AutoDock Vina: Improving the Speed and Accuracy of Docking with a New Scoring function, Efficient optimization, and Multithreading*. Journal of Computational Chemistry, 31(2).

UCSF ChimeraX Home Page. (n.d.). Www.cgl.ucsf.edu. Available at: https://www.cgl.ucsf.edu/chimerax/

Varma, S., & Simon, R. (2006). *Bias in Error Estimation When Using Cross-Validation for Model Selection.* BMC Bioinformatics, *7*(1), 91.

Varoquaux, G., Raamana, P. R., Engemann, D. A., Hoyos-Idrobo, A., Schwartz, Y., & Thirion, B. (2017). *Assessing and tuning brain decoders: Cross-validation, caveats, and guidelines*. NeuroImage, 145, 166–179.

Vaskovic, J. (2023). *Action potential*. Kenhub. Available at: https://www.kenhub.com/en/library/physiology/action-potential

Wang, H. (2024). *Prediction of protein–ligand binding affinity via deep learning models*. Briefings in Bioinformatics, 25(2).

Zhang, S., Golbraikh, A., & Tropsha, A. (2006). *Development of Quantitative Structure−Binding Affinity Relationship Models Based on Novel Geometrical Chemical Descriptors of the Protein−Ligand Interfaces*. Journal of Medicinal Chemistry, 49(9), 2713–2724.

APPENDICES

A. Top 5 poses for each ligand with Chimera AutoDock Vina and Python Script

Top 5 Poses for AMLODIPINE:

Pose	Binding Affinity	RMSD Lower Bound	RMSD Upper Bound	Composite Score
1	-7.524	0.000	0.000	1.000000
2	-7.443	2.419	4.830	0.767533
4	-7.085	2.572	4.870	0.648967
5	-7.037	2.376	4.695	0.645159
3	-7.131	2.765	5.866	0.631250

Top 5 Poses for BROMOCRIPTINE:

Pose	Binding Affinity	RMSD Lower Bound	RMSD Upper Bound	Composite Score
1	-9.807	0.000	0.000	1.000000
2	-9.604	1.571	2.584	0.719442
5	-9.063	2.814	4.500	0.386319
4	-9.144	3.127	4.909	0.372074
3	-9.459	4.704	6.990	0.270925

Top 5 Poses for CLOZAPINE:

Pose	Binding Affinity	RMSD Lower Bound	RMSD Upper Bound	Composite Score
1	-9.499	0.000	0.000	1.000000
5	-8.760	2.037	2.399	0.683476
3	-9.247	3.581	7.734	0.637632
2	-9.319	8.765	11.900	0.396030
6	-8.700	6.624	11.320	0.312432

Top 5 Poses for Conformer3D_COMPOUND_CID_11310988:

Pose	Binding Affinity	RMSD Lower Bound	RMSD Upper Bound	Composite Score
1	-7.662	0.000	0.000	1.000000
2	-7.400	2.023	6.669	0.689447
3	-6.990	1.702	2.570	0.646039
9	-6.754	3.035	5.279	0.456393
7	-6.879	3.548	7.338	0.437421

Top 5 Poses for Top 5 Poses for Conformer3D_COMPOUND_CID_11447499:

Pose	Binding Affinity	RMSD Lower Bound	RMSD Upper Bound	Composite Score
1	-7.360	0.000	0.000	1.000000
2	-7.330	1.151	1.340	0.920166
3	-7.172	1.162	1.456	0.820772
4	-7.120	2.419	3.363	0.712491
5	-6.977	1.630	2.663	0.662581

Top 5 Poses for Conformer3D_COMPOUND_CID_11535974:

Pose	Binding Affinity	RMSD Lower Bound	RMSD Upper Bound	Composite Score
1	-8.398	0.000	0.000	1.000000
2	-7.732	1.976	2.545	0.744553
5	-7.444	1.899	2.622	0.672886
4	-7.512	1.486	6.160	0.631180
3	-7.563	2.381	6.427	0.620233

Top 5 Poses for Conformer3D_COMPOUND_CID_11608403:

Pose	Binding Affinity	RMSD Lower Bound	RMSD Upper Bound	Composite Score

1	-8.726	0.000	0.000	1.000000
4	-7.813	1.667	1.957	0.708850
2	-7.874	1.768	6.613	0.620253
3	-7.824	2.142	6.314	0.604478
7	-7.483	2.276	3.092	0.594335

Top 5 Poses for Conformer3D_COMPOUND_CID_11623136:

Pose	Binding Affinity	RMSD Lower Bound	RMSD Upper Bound	Composite Score
1	-7.547	0.000	0.000	1.000000
2	-7.367	3.346	4.339	0.726730
5	-7.163	1.324	1.651	0.722580
3	-7.252	1.321	6.486	0.672397
9	-7.029	3.872	8.468	0.444365

Top 5 Poses for Conformer3D_COMPOUND_CID_11658763:

Pose	Binding Affinity	RMSD Lower Bound	RMSD Upper Bound	Composite Score
1	-7.677	0.000	0.000	1.000000
3	-7.248	6.010	7.688	0.528387
9	-6.765	2.783	4.896	0.503150
8	-6.814	4.017	6.380	0.456314
10	-6.724	3.861	6.524	0.425735

Top 5 Poses for Conformer3D_COMPOUND_CID_11673089:

Pose	Binding Affinity	RMSD Lower Bound	RMSD Upper Bound	Composite Score
1	-7.890	0.000	0.000	1.000000
2	-7.693	1.852	2.716	0.718743
3	-7.541	2.146	3.401	0.618048
4	-7.381	2.000	2.913	0.583127

8	-7.209	2.700	4.349	0.423784

Top 5 Poses for Conformer3D_COMPOUND_CID_11694324:

Pose	Binding Affinity	RMSD Lower Bound	RMSD Upper Bound	Composite Score
1	-7.829	0.000	0.000	1.000000
2	-7.741	2.593	3.280	0.852055
3	-7.628	2.573	3.318	0.815793
4	-7.349	2.777	6.542	0.660779
5	-7.348	3.664	6.331	0.644944

Top 5 Poses for Conformer3D_COMPOUND_CID_16006089:

Pose	Binding Affinity	RMSD Lower Bound	RMSD Upper Bound	Composite Score
1	-10.430	0.000	0.000	1.000000
2	-10.230	1.453	2.141	0.760300
3	-10.180	4.004	6.768	0.403548
5	-9.696	2.384	8.885	0.287613
4	-10.070	4.821	7.863	0.273415

Top 5 Poses for Conformer3D_COMPOUND_CID_24855949:

Pose	Binding Affinity	RMSD Lower Bound	RMSD Upper Bound	Composite Score
1	-8.986	0.000	0.000	1.000000
3	-8.716	2.393	8.324	0.480704
10	-8.193	2.321	3.499	0.429517
9	-8.231	2.737	4.512	0.389789
4	-8.649	3.246	9.113	0.382380

Top 5 Poses for Conformer3D_COMPOUND_CID_24855953:

Pose	Binding	RMSD Lower	RMSD Upper	Composite

	Affinity	Bound	Bound	Score
1	-9.651	0.000	0.000	1.000000
3	-8.667	1.731	2.234	0.650647
7	-8.308	2.860	4.408	0.468666
4	-8.664	4.310	7.563	0.411505
8	-8.260	4.629	7.313	0.313507

Top 5 Poses for Conformer3D_COMPOUND_CID_24855981:

Pose	Binding Affinity	RMSD Lower Bound	RMSD Upper Bound	Composite Score
1	-10.530	0.000	0.000	1.000000
3	-10.120	1.800	2.337	0.693584
4	-10.100	2.464	9.066	0.456693
5	-9.925	2.340	9.459	0.388982
2	-10.390	5.274	11.450	0.347788

Top 5 Poses for Conformer3D_COMPOUND_CID_24856012:

Pose	Binding Affinity	RMSD Lower Bound	RMSD Upper Bound	Composite Score
1	-9.634	0.000	0.000	1.000000
4	-9.228	2.274	2.831	0.649327
9	-8.750	1.687	2.317	0.531373
5	-9.000	3.098	4.613	0.471443
3	-9.257	3.796	8.683	0.399562

Top 5 Poses for Conformer3D_COMPOUND_CID_24856046:

Pose	Binding Affinity	RMSD Lower Bound	RMSD Upper Bound	Composite Score
1	-10.080	0.000	0.000	1.000000
3	-9.578	1.547	1.809	0.730895
8	-9.028	2.648	3.135	0.480455

5	-9.367	3.705	5.412	0.474801
6	-9.347	3.455	5.865	0.468723

Top 5 Poses for Conformer3D_COMPOUND_CID_24856107:

Pose	Binding Affinity	RMSD Lower Bound	RMSD Upper Bound	Composite Score
1	-10.600	0.000	0.000	1.000000
3	-9.937	4.285	7.633	0.435235
4	-9.682	4.185	7.606	0.374716
2	-10.340	6.246	11.360	0.353193
7	-9.471	4.561	7.177	0.313788

Top 5 Poses for Conformer3D_COMPOUND_CID_24947569:

Pose	Binding Affinity	RMSD Lower Bound	RMSD Upper Bound	Composite Score
1	-9.335	0.0000	0.000	1.000000
2	-9.263	0.8117	1.056	0.916583
4	-8.765	1.8450	2.846	0.697736
9	-8.156	2.6330	4.209	0.473574
5	-8.713	4.3130	7.265	0.455903

Top 5 Poses for Conformer3D_COMPOUND_CID_24947939:

Pose	Binding Affinity	RMSD Lower Bound	RMSD Upper Bound	Composite Score
1	-10.190	0.000	0.000	1.000000
3	-9.794	1.410	1.813	0.816973
2	-9.934	2.001	3.001	0.799177
4	-9.640	2.561	3.763	0.694763
5	-9.597	5.489	7.796	0.483775

Top 5 Poses for Conformer3D_COMPOUND_CID_24964158:

Pose	Binding Affinity	RMSD Lower Bound	RMSD Upper Bound	Composite Score
1	-9.353	0.000	0.000	1.000000
2	-8.989	1.530	1.872	0.723190
3	-8.970	1.825	2.201	0.687595
4	-8.668	1.799	2.729	0.573182
9	-8.390	1.963	2.717	0.472402

Top 5 Poses for Conformer3D_COMPOUND_CID_44351345:

Pose	Binding Affinity	RMSD Lower Bound	RMSD Upper Bound	Composite Score
1	-11.080	0.000	0.000	1.000000
3	-10.540	1.040	1.840	0.769295
9	-10.070	1.978	3.182	0.573985
7	-10.240	2.782	3.870	0.563039
10	-9.755	2.599	3.458	0.460074

Top 5 Poses for Conformer3D_COMPOUND_CID_44390396:

Pose	Binding Affinity	RMSD Lower Bound	RMSD Upper Bound	Composite Score
1	-7.644	0.0000	0.000	1.000000
2	-7.374	0.9077	1.115	0.854380
3	-7.318	2.1340	2.858	0.740583
4	-7.096	1.6210	2.142	0.716419
6	-6.831	2.6120	6.355	0.479039

Top 5 Poses for Conformer3D_COMPOUND_CID_44456154:

Pose	Binding Affinity	RMSD Lower Bound	RMSD Upper Bound	Composite Score
1	-10.030	0.000	0.000	1.000000
2	-9.996	1.329	1.873	0.848732

3	-9.768	1.550	1.972	0.773039
4	-9.766	2.067	2.617	0.720116
6	-9.250	2.752	3.816	0.503481

Top 5 Poses for DANTROLENE_COMPOUND_CID_6914273:

Pose	Binding Affinity	RMSD Lower Bound	RMSD Upper Bound	Composite Score
1	-9.284	0.0000	0.000	1.000000
2	-9.258	1.1960	1.781	0.872258
3	-9.239	1.7070	2.073	0.828532
4	-8.961	0.9312	2.016	0.791366
6	-8.468	2.0760	2.717	0.556085

Top 5 Poses for Diphenhydramine_COMPOUND_CID_3100:

Pose	Binding Affinity	RMSD Lower Bound	RMSD Upper Bound	Composite Score
1	-7.836	0.0000	0.000	1.000000
2	-7.442	0.3864	4.309	0.617308
3	-7.325	2.1910	5.086	0.419748
7	-7.171	2.5440	3.172	0.410435
9	-7.105	2.0870	3.273	0.405202

Top 5 Poses for FLUTAMIDE_COMPOUND_CID_3397:

Pose	Binding Affinity	RMSD Lower Bound	RMSD Upper Bound	Composite Score
1	-8.199	0.000	0.000	1.000000
2	-7.813	2.066	2.731	0.795774
3	-7.714	2.256	2.935	0.759441
8	-7.385	2.562	3.294	0.651121
5	-7.546	4.445	6.828	0.594228

Top 5 Poses for INDATRALINE_COMPOUND_CID_3703:

Pose	Binding Affinity	RMSD Lower Bound	RMSD Upper Bound	Composite Score
1	-9.796	0.000	0.000	1.000000
2	-9.672	1.876	2.503	0.878129
3	-9.571	2.129	2.925	0.837126
4	-9.202	1.630	2.471	0.756060
7	-8.862	3.680	5.652	0.558608

Top 5 Poses for IPRINDOLE_COMPOUND_CID_21722:

Pose	Binding Affinity	RMSD Lower Bound	RMSD Upper Bound	Composite Score
1	-8.240	0.000	0.000	1.000000
2	-8.158	1.770	5.066	0.793670
3	-8.048	1.212	4.172	0.785478
5	-7.975	1.501	4.250	0.744781
4	-7.994	1.759	5.004	0.727217

Top 5 Poses for LOSARTAN_COMPOUND_CID_3961:

Pose	Binding Affinity	RMSD Lower Bound	RMSD Upper Bound	Composite Score
1	-9.809	0.000	0.000	1.000000
3	-9.304	3.045	3.701	0.386599
7	-8.898	2.066	2.538	0.385310
6	-8.938	2.122	3.184	0.366317
2	-9.398	2.622	7.062	0.313633

Top 5 Poses for METOCLOPRAMIDE_COMPOUND_CID_4168:

Pose	Binding Affinity	RMSD Lower Bound	RMSD Upper Bound	Composite Score
1	-6.760	0.000	0.000	1.000000

4	-6.547	2.696	6.338	0.468835
3	-6.575	3.698	6.893	0.399112
9	-6.237	2.899	4.056	0.382474
7	-6.305	2.711	6.090	0.349849

Top 5 Poses for MIRTAZAPINE_Conformer3D_COMPOUND_CID_4205:

Pose	Binding Affinity	RMSD Lower Bound	RMSD Upper Bound	Composite Score
1	-8.735	0.000	0.000	1.000000
6	-8.518	1.080	1.355	0.720238
3	-8.582	1.262	4.741	0.571578
4	-8.555	2.717	4.450	0.439059
8	-8.464	2.262	4.808	0.414710

Top 5 Poses for NALTREXONE_COMPOUND_CID_5360515:

Pose	Binding Affinity	RMSD Lower Bound	RMSD Upper Bound	Composite Score
1	-8.912	0.000	0.000	1.000000
2	-8.845	2.444	4.857	0.780188
3	-8.742	2.731	3.601	0.764124
4	-8.546	2.473	5.466	0.648800
10	-8.046	2.006	2.939	0.534010

Top 5 Poses for NEFAZODONE_COMPOUND_CID_4449:

Pose	Binding Affinity	RMSD Lower Bound	RMSD Upper Bound	Composite Score
1	-10.540	0.000	0.000	1.000000
2	-10.110	3.838	5.306	0.448241
8	-9.937	4.320	4.954	0.358037
4	-10.050	5.087	6.050	0.334267
9	-9.920	4.521	5.642	0.318890

Top 5 Poses for PIROXICAM_COMPOUND_CID_54676228:

Pose	Binding Affinity	RMSD Lower Bound	RMSD Upper Bound	Composite Score
1	-9.076	0.000	0.000	1.000000
3	-8.741	1.920	3.100	0.631585
2	-8.990	2.905	7.734	0.520041
4	-8.717	3.127	4.496	0.499194
5	-8.693	4.062	5.180	0.407860

Top 5 Poses for QUETIAPINE_COMPOUND_CID_5002:

Pose	Binding Affinity	RMSD Lower Bound	RMSD Upper Bound	Composite Score
1	-8.679	0.000	0.000	1.000000
5	-8.124	3.295	5.796	0.629053
2	-8.259	4.557	7.809	0.594568
8	-7.818	3.254	3.907	0.577849
3	-8.152	4.049	7.881	0.574715

Top 5 Poses for REBOXETINE_COMPOUND_CID_127151:

Pose	Binding Affinity	RMSD Lower Bound	RMSD Upper Bound	Composite Score
1	-8.538	0.000	0.000	1.000000
7	-7.817	1.748	2.763	0.594357
3	-7.971	2.678	4.156	0.587815
6	-7.907	2.218	4.662	0.566756
4	-7.955	3.580	5.652	0.514157

Top 5 Poses for TRIMIPRAMINE_COMPOUND_CID_5584:

Pose	Binding Affinity	RMSD Lower Bound	RMSD Upper Bound	Composite Score

1	-8.627	0.000	0.000	1.000000
2	-8.579	2.660	7.273	0.665815
6	-8.188	1.553	4.610	0.630041
5	-8.252	2.237	5.047	0.617087
3	-8.260	2.502	6.115	0.578712

Top 5 Poses for ZOLPIDEM_COMPOUND_CID_5732:

Pose	**Binding Affinity**	**RMSD Lower Bound**	**RMSD Upper Bound**	**Composite Score**
1	-8.958	0.000	0.000	1.000000
2	-8.878	1.768	4.159	0.803351
4	-8.655	2.395	4.685	0.677699
7	-8.412	1.668	2.404	0.658587
5	-8.582	2.020	4.799	0.655180

Top 5 Poses for ATOMOXETINE_COMPOUND_CID_54841:

Pose	**Binding Affinity**	**RMSD Lower Bound**	**RMSD Upper Bound**	**Composite Score**
1	-8.230	0.0000	0.000	1.000000
3	-7.909	0.7422	2.055	0.798195
2	-8.095	1.5020	5.638	0.658978
6	-7.112	1.4230	2.632	0.541864
4	-7.318	1.7440	5.612	0.456037

Top 5 Poses for ACETAMINOPHEN_COMPOUND_CID_1983:

Pose	**Binding Affinity**	**RMSD Lower Bound**	**RMSD Upper Bound**	**Composite Score**
1	-6.078	0.000	0.000	1.000000
2	-5.971	1.478	4.769	0.829390
3	-5.867	2.069	2.623	0.801272
4	-5.834	2.381	2.804	0.774520

5	-5.774	2.948	3.944	0.711360

Top 5 Poses for AMITRIPTYLINE_COMPOUND_CID_2160:

Pose	Binding Affinity	RMSD Lower Bound	RMSD Upper Bound	Composite Score
1	-8.991	0.000	0.000	1.000000
2	-8.401	2.073	5.554	0.656883
5	-8.310	2.416	4.389	0.646271
4	-8.382	2.763	5.805	0.628623
7	-8.228	4.435	7.240	0.510925

Top 5 Poses for ASPIRIN_COMPOUND_CID_2244:

Pose	Binding Affinity	RMSD Lower Bound	RMSD Upper Bound	Composite Score
1	-6.570	0.000	0.000	1.000000
2	-6.526	1.760	3.484	0.799232
3	-6.458	2.023	2.544	0.788851
5	-6.099	2.834	4.294	0.539706
6	-6.061	2.615	5.110	0.504824

Top 5 Poses for BUBROPION_COMPOUND_CID_444:

Pose	Binding Affinity	RMSD Lower Bound	RMSD Upper Bound	Composite Score
1	-7.263	0.000	0.000	1.000000
2	-7.189	1.985	2.630	0.772865
3	-7.091	1.808	2.775	0.730354
4	-6.868	1.722	3.040	0.620837
7	-6.617	1.954	2.483	0.513507

Top 5 Poses for CIMETIDINE_COMPOUND_CID_2756:

Pose	Binding	RMSD Lower	RMSD Upper	Composite

	Affinity	Bound	Bound	Score
1	-6.130	0.0000	0.000	1.000000
2	-6.074	0.7883	2.159	0.833194
9	-5.772	1.8890	3.353	0.519221
3	-5.966	1.6440	7.098	0.517508
5	-5.838	3.1090	4.244	0.461997

Top 5 Poses for CITALOPRAM_CID_2771:

Pose	Binding Affinity	RMSD Lower Bound	RMSD Upper Bound	Composite Score
1	-8.991	0.000	0.000	1.000000
3	-8.240	2.085	2.753	0.703838
2	-8.298	3.848	6.322	0.601099
4	-8.226	4.213	6.023	0.580426
6	-8.103	4.871	6.692	0.519678

Top 5 Poses for CLOMIPRAMINE_COMPOUND_CID_2801:

Pose	Binding Affinity	RMSD Lower Bound	RMSD Upper Bound	Composite Score
1	-8.493	0.000	0.000	1.000000
3	-7.906	1.659	2.123	0.639394
2	-8.115	2.771	5.887	0.493080
6	-7.832	2.256	5.391	0.455850
10	-7.408	2.933	5.362	0.286330

Top 5 Poses for CLOPHENIRAMINE_COMPOUND_CID_2725:

Pose	Binding Affinity	RMSD Lower Bound	RMSD Upper Bound	Composite Score
1	-7.871	0.000	0.000	1.000000
3	-7.393	1.875	3.060	0.627946
5	-7.302	1.817	3.251	0.596842

2	-7.651	2.997	5.248	0.552960
6	-7.258	2.536	5.356	0.459077

Top 5 Poses for DESIPRAMINE_COMPOUND_CID_2995:

Pose	Binding Affinity	RMSD Lower Bound	RMSD Upper Bound	Composite Score
1	-8.359	0.000	0.000	1.000000
2	-8.051	2.001	5.329	0.688746
4	-7.718	2.197	5.362	0.563910
7	-7.532	1.587	5.271	0.521417
3	-7.860	3.401	7.546	0.517803

Top 5 Poses for DESVENLAFAXINE_COMPOUND_CID_125017:

Pose	Binding Affinity	RMSD Lower Bound	RMSD Upper Bound	Composite Score
1	-7.352	0.000	0.000	1.000000
2	-7.286	1.697	3.983	0.683070
3	-7.249	1.459	3.986	0.670105
4	-7.210	2.100	5.052	0.559244
6	-7.110	2.207	3.737	0.538199

Top 5 Poses for DEXTROMETHORPHAN_COMPOUND_CID_5360696:

Pose	Binding Affinity	RMSD Lower Bound	RMSD Upper Bound	Composite Score
1	-8.997	0.000	0.000	1.000000
2	-8.990	2.222	4.155	0.856418
3	-8.376	1.961	4.138	0.709919
4	-8.282	2.174	4.732	0.668760
6	-7.809	2.731	5.342	0.524128

Top 5 Poses for DOXEPIN_COMPOUND_CID_667477:

Pose	Binding Affinity	RMSD Lower Bound	RMSD Upper Bound	Composite Score
1	-8.607	0.000	0.000	1.000000
2	-8.162	4.657	7.271	0.555206
5	-8.060	3.920	6.612	0.547975
4	-8.072	4.559	6.596	0.536998
3	-8.129	4.724	7.447	0.536928

Top 5 Poses for DULOXETINE_COMPOUND_CID_60835:

Pose	Binding Affinity	RMSD Lower Bound	RMSD Upper Bound	Composite Score
1	-8.528	0.000	0.000	1.000000
2	-8.029	2.340	5.868	0.550223
4	-8.000	2.571	6.180	0.521642
3	-8.015	3.569	6.488	0.486426
7	-7.885	3.591	6.149	0.438085

Top 5 Poses for ESCITALOPRAM_COMPOUND_CID_146570:

Pose	Binding Affinity	RMSD Lower Bound	RMSD Upper Bound	Composite Score
1	-9.175	0.000	0.000	1.000000
4	-8.365	2.104	3.220	0.635389
2	-8.489	3.841	6.266	0.563323
7	-8.257	4.625	6.475	0.469445
3	-8.404	5.685	7.559	0.463302

Top 5 Poses for FEXOFENADINE_COMPOUND_CID_3348:

Pose	Binding Affinity	RMSD Lower Bound	RMSD Upper Bound	Composite Score
1	-10.93	0.000	0.000	1.000000
2	-10.89	2.061	3.445	0.764486

5	-10.72	1.425	3.080	0.744072
4	-10.75	2.366	4.592	0.657978
8	-10.30	1.889	2.306	0.561988

Top 5 Poses for FLUOXETINE_CID_3386:

Pose	Binding Affinity	RMSD Lower Bound	RMSD Upper Bound	Composite Score
1	-9.268	0.000	0.000	1.000000
2	-8.629	1.446	2.284	0.724027
4	-8.103	1.931	3.231	0.562049
6	-7.901	2.080	3.555	0.502949
3	-8.480	4.260	6.644	0.411342

Top 5 Poses for FLUVOXAMINE_CID_5324346:

Pose	Binding Affinity	RMSD Lower Bound	RMSD Upper Bound	Composite Score
1	-8.007	0.0000	0.000	1.000000
3	-7.916	0.9886	1.507	0.913120
2	-7.939	1.3330	1.798	0.904758
4	-7.906	1.6230	2.329	0.875393
5	-7.477	1.8660	2.460	0.734461

Top 5 Poses for IBUPROFEN_COMPOUND_CID_3672:

Pose	Binding Affinity	RMSD Lower Bound	RMSD Upper Bound	Composite Score
1	-7.505	0.000	0.000	1.000000
3	-7.075	1.226	2.057	0.780508
2	-7.453	3.690	6.342	0.775007
7	-6.756	1.724	2.413	0.649922
9	-6.612	2.578	3.486	0.559104

Top 5 Poses for IMIPRAMINE_COMPOUND_CID_3696:

Pose	Binding Affinity	RMSD Lower Bound	RMSD Upper Bound	Composite Score
1	-8.448	0.000	0.000	1.000000
3	-8.003	1.733	5.548	0.640254
2	-8.055	3.154	7.505	0.564812
4	-7.804	1.908	5.919	0.551163
7	-7.692	2.996	5.764	0.477680

Top 5 Poses for LEVOMILNACIPRAN_COMPOUND_CID_6917779:

Pose	Binding Affinity	RMSD Lower Bound	RMSD Upper Bound	Composite Score
1	-7.854	0.000	0.000	1.000000
2	-7.107	1.371	2.015	0.600660
5	-6.996	1.527	1.723	0.569127
6	-6.890	1.846	2.482	0.483956
3	-7.020	2.094	3.198	0.474232

Top 5 Poses for MAPROTILINE_COMPOUND_CID_4011:

Pose	Binding Affinity	RMSD Lower Bound	RMSD Upper Bound	Composite Score
1	-8.967	0.000	0.000	1.000000
2	-8.928	2.487	5.165	0.802928
3	-8.825	2.822	5.363	0.756733
6	-8.255	2.015	3.472	0.642726
5	-8.348	2.407	4.969	0.627659

Top 5 Poses for MAZINDOL_COMPOUND_CID_4020:

Pose	Binding Affinity	RMSD Lower Bound	RMSD Upper Bound	Composite Score
1	-8.709	0.000	0.000	1.000000

2	-8.686	2.778	5.187	0.605710
5	-8.411	2.696	5.034	0.522707
7	-8.110	3.483	6.165	0.323434
3	-8.522	5.168	7.462	0.304373

Top 5 Poses for METFORMIN_COMPOUND_CID_4091:

Pose	Binding Affinity	RMSD Lower Bound	RMSD Upper Bound	Composite Score
1	-5.364	0.000	0.000	1.000000
2	-5.154	1.412	1.518	0.894515
4	-4.808	1.397	1.667	0.788814
3	-4.875	1.880	2.665	0.787595
5	-4.749	1.383	1.394	0.775192

Top 5 Poses for MILNACIPRAN_COMPOUND_CID_65833:

Pose	Binding Affinity	RMSD Lower Bound	RMSD Upper Bound	Composite Score
1	-7.218	0.000	0.000	1.000000
2	-7.021	1.815	2.008	0.682278
4	-6.846	1.607	2.305	0.586511
8	-6.749	1.378	1.975	0.561457
3	-6.876	1.942	3.097	0.547866

Top 5 Poses for NAPROXEN_COMPOUND_CID_156391:

Pose	Binding Affinity	RMSD Lower Bound	RMSD Upper Bound	Composite Score
1	-8.782	0.000	0.000	1.000000
4	-8.154	1.199	2.269	0.770645
2	-8.327	1.269	6.448	0.710655
3	-8.182	1.426	6.641	0.669012
7	-7.503	2.935	6.837	0.466617

Top 5 Poses for NORTRIPTYLINE_COMPOUND_CID_4543:

Pose	Binding Affinity	RMSD Lower Bound	RMSD Upper Bound	Composite Score
1	-8.992	0.000	0.000	1.000000
2	-8.745	2.312	5.458	0.725699
3	-8.669	2.274	4.474	0.717858
6	-8.322	1.710	4.743	0.589444
4	-8.645	5.686	7.326	0.559641

Top 5 Poses for PAROXETINE_COMPOUND_CID_43815:

Pose	Binding Affinity	RMSD Lower Bound	RMSD Upper Bound	Composite Score
1	-10.090	0.000	0.000	1.000000
2	-8.475	2.841	3.936	0.433060
3	-8.439	2.912	3.813	0.427706
5	-8.390	3.148	4.010	0.400246
6	-8.350	3.472	6.180	0.298234

Top 5 Poses for PROBENECID_COMPOUND_CID_4911:

Pose	Binding Affinity	RMSD Lower Bound	RMSD Upper Bound	Composite Score
1	-7.594	0.000	0.000	1.000000
2	-7.467	1.768	2.446	0.856924
3	-7.266	1.263	1.896	0.826020
5	-6.926	1.395	2.771	0.704417
6	-6.675	4.718	8.204	0.413108

Top 5 Poses for PROTRIPTYLINE_COMPOUND_CID_4976:

Pose	Binding Affinity	RMSD Lower Bound	RMSD Upper Bound	Composite Score

1	-8.882	0.0000	0.000	1.000000
2	-8.619	1.4640	2.166	0.813885
5	-8.245	1.7120	2.207	0.673190
4	-8.265	0.8695	4.643	0.648642
3	-8.269	1.8420	4.963	0.614176

Top 5 Poses for RANITIDINE_COMPOUND_CID_3001055:

Pose	Binding Affinity	RMSD Lower Bound	RMSD Upper Bound	Composite Score
1	-6.398	0.000	0.000	1.000000
5	-6.141	1.303	1.523	0.657612
8	-5.924	1.688	2.056	0.465515
2	-6.356	3.575	4.603	0.445504
10	-5.855	1.526	2.138	0.431936

Top 5 Poses for SERTRALINE_CID_68617:

Pose	Binding Affinity	RMSD Lower Bound	RMSD Upper Bound	Composite Score
1	-9.461	0.000	0.000	1.000000
2	-9.285	1.770	2.416	0.807273
4	-8.520	1.794	2.484	0.625660
5	-8.430	3.799	6.049	0.406762
3	-8.794	5.552	7.009	0.386596

Top 5 Poses for VENLAFAXINE_COMPOUND_CID_5656:

Pose	Binding Affinity	RMSD Lower Bound	RMSD Upper Bound	Composite Score
1	-7.585	0.000	0.000	1.000000
2	-7.532	1.498	2.307	0.802880
3	-7.466	1.597	3.720	0.713850
4	-7.459	2.580	4.957	0.607789

7	-7.138	1.927	2.764	0.576117

Top 5 Poses for VILAZODONE_COMPOUND_CID_6918314:

Pose	Binding Affinity	RMSD Lower Bound	RMSD Upper Bound	Composite Score
1	-10.530	0.000	0.000	1.000000
5	-10.160	1.620	2.298	0.688417
8	-10.100	3.857	5.300	0.483140
10	-9.849	2.624	5.249	0.417318
2	-10.340	4.501	13.070	0.378059

Top 5 Poses for VORTIOXETINE_COMPOUND_CID_9966051:

Pose	Binding Affinity	RMSD Lower Bound	RMSD Upper Bound	Composite Score
1	-8.832	0.000	0.000	1.000000
5	-8.074	1.377	2.376	0.553724
2	-8.313	2.288	5.737	0.412751
3	-8.203	2.006	5.516	0.408739
10	-7.775	1.463	3.965	0.382683

B. Setup and Execution of Multiple Ligand Docking Using AutoDock Vina in Ubuntu

This code below outlines the step-by-step procedure used to perform multiple ligand docking using AutoDock Vina in a Linux-based (Ubuntu) environment. It includes installation of required packages, preparation of receptor and ligand files, execution of batch docking via a Perl automation script and finally extraction of the 10 poses for each ligand.

For installation of packages and libraries:

sudo apt update

sudo apt upgrade

sudo apt install gcc

sudo apt install cmake

sudo apt install build-essential

sudo apt install libfftw3-dev # or

sudo apt-get install -y libfftw3-dev

For installation of Open Babel and AutoDock Vina:

sudo apt install openbabel

sudo apt install autodock-vina

For minimizing the ligands:

obminimize -ff MMFF94 -n 1000 *.sdf

Conversion to .pdbqt Format:

obabel -isdf *.sdf -opdbqt -O *.pdbqt

Creation ligand list:

ls *.pdbqt > ligand.txt

Run the docking simulations:

perl Vina_linux.pl
ligand.txt

Get final results:
tail -n11 *.log > results.txt

Scripts and supporting files (such as Vina_linux.pl and conf.txt) were sourced from the following GitHub repository:

 https://github.com/DweipayanG/Multiple_Ligand_Docking_Vinna

C. Summary of Residue Interactions and Molecular Descriptors

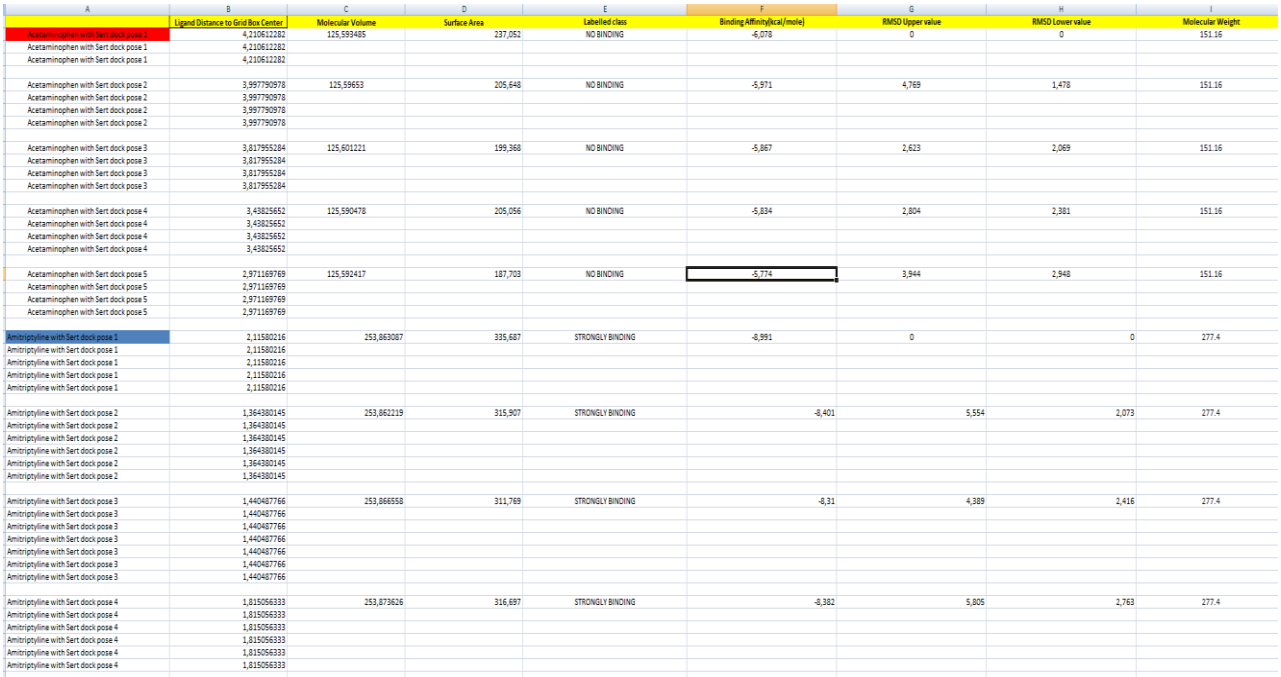
The following snapshots represent key portions of the dataset used for the whole machine learning pipeline:

- Residue Interaction Snapshot (from the *"Residue-Analysis"* sheet): This table contains all the interacted residues from protein with the 5 best poses from each ligand. Each cell is a tuple (e.g., 0,1,0,1) denoting the number of hydrogen bonds, hydrophobic interactions, van der Waals contacts and other interactions respectively. Rows correspond to ligand-protein docking poses, while columns represent interacting residues. The respective chapter in the present thesis is 3.4.2 (Image 56).
- Molecular Descriptors Snapshot (from the *"Molecular Descriptors"* sheet): This includes all molecular descriptors described in chapter 3.4.1 for each ligand pose (Image 57).

These sheets were then merged into a single dataset that served as input for the classification pipeline and feature selection process detailed in the main methodology.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC
1	hydrogen,hydrophobic,van der waal,other type	TYR 95	ALA 96	VAL 97	ASP 98	LEU 99	GLY 100	ASN 101	TRP 103	ARG 104	TYR 107	ILE 108	ALA 110	GLN 111	ASN 112	GLY 113	GLY 114	ALA 169	ILE 172	ALA 173	TYR 175	TYR 176	ASN 177	ILE 179	PHE 263	PHE 311	LYS 314	PRO 315	ASN 3
2	Acetaminophen with Sert dock pose 1	0,0,1,0			0,0,1,0													0,0,1,0	0,1,0,0	0,1,0,0		0,1,1,0							
3	Acetaminophen with Sert dock pose 2	0,0,1,0																0,0,1,0	0,1,0,0	1,0,0,0		0,1,0,0	0,0,1,0						
4	Acetaminophen with Sert dock pose 3	0,0,1,0																	0,1,0,0	0,0,1,0		0,1,0,0	0,0,1,0						
5	Acetaminophen with Sert dock pose 4	0,0,1,0																	0,1,0,0			0,1,0,0	0,0,1,0						
6	Acetaminophen with Sert dock pose 5	1,0,0,0																	0,1,0,0			0,0,1,0							
7	Amitriptyline with Sert dock pose 1	0,0,1,0			0,0,1,0													0,0,1,0	0,3,0,0	0,0,1,0		0,1,0,0							
8	Amitriptyline with Sert dock pose 2	0,0,1,0			0,0,1,0														0,2,0,0		0,0,1,0	0,1,0,0							
9	Amitriptyline with Sert dock pose 3	0,2,0,0	0,0,1,0		1,0,0,0		0,0,1,0												0,2,0,0		0,0,1,0	0,1,0,0							
10	Amitriptyline with Sert dock pose 4	0,1,0,0	0,0,1,0		0,0,0,1														0,1,0,0	0,0,1,0	0,0,1,0	0,1,0,0	0,0,1,0						
11	Amitriptyline with Sert dock pose 5					0,0,1,0	0,0,1,0		0,0,1,0	0,0,1,0									0,0,1,0		0,0,1,0	0,0,1,0							
12	Aspirin with Sert dock pose 1	0,1,0,0	1,0,0,0		0,0,1,0														0,0,1,0			0,0,1,0							
13	Aspirin with Sert dock pose 2	0,0,1,0	0,0,1,0		0,0,1,0														0,0,1,0			0,1,0,0							
14	Aspirin with Sert dock pose 3	0,0,1,0			0,0,1,0														0,1,0,0			0,0,1,0							
15	Aspirin with Sert dock pose 4	1,1,0,0			0,0,1,0														0,1,0,0			1,0,0,0							
16	Aspirin with Sert dock pose 5	0,0,1,0			0,0,1,0														0,1,0,0		0,0,1,0	0,0,1,0							
17	Bupropion with Sert dock pose 1	0,0,1,0																0,1,0,0	0,2,0,0			0,0,1,0							
18	Bupropion with Sert dock pose 2	0,0,1,0			0,0,1,0													0,1,0,0	0,2,0,0			0,0,1,0							
19	Bupropion with Sert dock pose 3	0,0,1,0			0,0,1,0													0,0,1,0	0,1,0,0	0,1,0,0		0,2,0,0	0,0,1,0						
20	Bupropion with Sert dock pose 4	0,0,1,0																	0,2,0,0		0,0,1,0								
21	Bupropion with Sert dock pose 5	0,1,0,0			0,0,1,0														0,1,0,0		0,0,1,0	0,0,1,0							
22	Cimetidine with Sert dock pose 1	0,0,0,1	0,0,1,0		0,0,0,1													1,1,0,0	0,2,0,0	0,2,0,0		0,1,0,0	0,0,1,0						
23	Cimetidine with Sert dock pose 2	1,1,0,1	0,0,1,0		0,0,0,1													1,1,0,0	0,2,0,0	0,2,0,0		0,1,0,0	0,0,1,0						
24	Cimetidine with Sert dock pose 3	0,0,1,0			0,0,0,1													0,1,0,0	0,2,0,0	0,1,0,0	0,0,1,0	0,0,1,0	0,0,1,0						
25	Cimetidine with Sert dock pose 4	0,0,1,0			0,0,1,0													0,0,1,0	0,0,1,0	0,0,1,0		0,0,0,1	1,0,0,0						
26	Cimetidine with Sert dock pose 5	0,0,1,0			1,0,0,0	0,0,1,0													0,1,0,0		0,0,1,0	0,1,0,0							
27	Citalopram with Sert dock pose 1	3,0,0,0	0,0,1,0		1,0,0,0													0,0,1,0	0,2,0,0	1,0,0,0	0,0,1,0	0,1,0,0	0,0,1,0						
28	Citalopram with Sert dock pose 2	0,0,1,0			1,0,0,0		0,0,1,0												0,2,0,0			0,1,0,0							
29	Citalopram with Sert dock pose 3	0,0,1,0			0,0,1,0					1,0,0,0									0,0,1,0		0,0,1,0	0,0,1,0							
30	Citalopram with Sert dock pose 4	0,0,1,0			1,0,0,0	0,0,1,0	0,0,1,0		0,0,1,0										0,0,1,0		0,0,1,0	0,1,0,0		0,0,1,0					
31	Citalopram with Sert dock pose 5	0,0,1,0			0,0,1,0	0,0,1,0	0,0,1,0		0,0,1,0	2,0,0,0									0,0,1,0		1,0,0,0	0,0,1,0		0,0,1,0					
32	Clomipramine with Sert dock pose 1	0,0,1,0			0,0,1,0		0,0,1,0												0,3,0,0			0,2,0,0							
33	Clomipramine with Sert dock pose 2	1,0,0,0			1,0,0,0														0,2,0,0			0,2,0,0							
34	Clomipramine with Sert dock pose 3	0,0,1,0			1,0,0,0		0,0,1,0			0,1,0,0									0,0,1,0			0,1,0,0							
35	Clomipramine with Sert dock pose 4	0,0,1,0			0,0,1,0	0,1,0,0	0,0,1,0												0,1,0,0		0,0,1,0	0,2,0,0							
36	Clomipramine with Sert dock pose 5	0,1,0,0			0,0,1,0	0,0,1,0	0,0,1,0		0,0,1,0	0,0,1,0									0,0,1,0		0,0,1,0	0,2,0,0		0,0,1,0					
37	Clorpheniramine with Sert dock pose 1	0,0,1,0			0,0,1,0														0,2,0,0	0,1,0,0	0,0,1,0	0,2,0,0							
38	Clorpheniramine with Sert dock pose 2	1,1,0,0	1,0,0,0		0,0,1,0														0,3,0,0	0,0,1,0		0,1,0,0							
39	Clorpheniramine with Sert dock pose 3	0,1,0,0	0,0,1,0		0,0,0,1														0,1,0,0		0,0,1,0	0,1,0,0							
40	Clorpheniramine with Sert dock pose 4	2,1,0,0	0,0,1,0		1,0,0,0														0,3,0,0			0,1,0,0							
41	Clorpheniramine with Sert dock pose 5	0,0,1,0			0,0,1,0														0,1,0,0		0,1,0,0	0,3,0,0							

Image 56: Residue-Interaction Snapshot from my personal excel file.



	Ligand Distance to Grid Box Center	Molecular Volume	Surface Area	Labelled class	Binding Affinity(kcal/mole)	RMSD Upper value	RMSD Lower value	Molecular Weight
Acetaminophen with Sert dock pose 1	4,210612282	125,593485	237,052	NO BINDING	-6,078	0	0	151.16
Acetaminophen with Sert dock pose 1	4,210612282							
Acetaminophen with Sert dock pose 1	4,210612282							
Acetaminophen with Sert dock pose 2	3,997790978	125,59653	205,648	NO BINDING	-5,971	4,769	1,478	151.16
Acetaminophen with Sert dock pose 2	3,997790978							
Acetaminophen with Sert dock pose 2	3,997790978							
Acetaminophen with Sert dock pose 2	3,997790978							
Acetaminophen with Sert dock pose 3	3,817955284	125,601221	199,368	NO BINDING	-5,867	2,623	2,069	151.16
Acetaminophen with Sert dock pose 3	3,817955284							
Acetaminophen with Sert dock pose 3	3,817955284							
Acetaminophen with Sert dock pose 3	3,817955284							
Acetaminophen with Sert dock pose 4	3,43825652	125,590478	205,056	NO BINDING	-5,834	2,804	2,381	151.16
Acetaminophen with Sert dock pose 4	3,43825652							
Acetaminophen with Sert dock pose 4	3,43825652							
Acetaminophen with Sert dock pose 4	3,43825652							
Acetaminophen with Sert dock pose 5	2,971169769	125,592417	187,703	NO BINDING	-5,774	3,944	2,948	151.16
Acetaminophen with Sert dock pose 5	2,971169769							
Acetaminophen with Sert dock pose 5	2,971169769							
Acetaminophen with Sert dock pose 5	2,971169769							
Amitriptyline with Sert dock pose 1	2,11580216	253,863087	335,687	STRONGLY BINDING	-8,991	0	0	277.4
Amitriptyline with Sert dock pose 1	2,11580216							
Amitriptyline with Sert dock pose 1	2,11580216							
Amitriptyline with Sert dock pose 1	2,11580216							
Amitriptyline with Sert dock pose 1	2,11580216							
Amitriptyline with Sert dock pose 2	1,364380145	253,862219	315,907	STRONGLY BINDING	-8,401	5,554	2,073	277.4
Amitriptyline with Sert dock pose 2	1,364380145							
Amitriptyline with Sert dock pose 2	1,364380145							
Amitriptyline with Sert dock pose 2	1,364380145							
Amitriptyline with Sert dock pose 2	1,364380145							
Amitriptyline with Sert dock pose 2	1,364380145							
Amitriptyline with Sert dock pose 3	1,440487766	253,866558	311,769	STRONGLY BINDING	-8,31	4,389	2,416	277.4
Amitriptyline with Sert dock pose 3	1,440487766							
Amitriptyline with Sert dock pose 3	1,440487766							
Amitriptyline with Sert dock pose 3	1,440487766							
Amitriptyline with Sert dock pose 3	1,440487766							
Amitriptyline with Sert dock pose 3	1,440487766							
Amitriptyline with Sert dock pose 3	1,440487766							
Amitriptyline with Sert dock pose 4	1,815056333	253,873626	316,697	STRONGLY BINDING	-8,382	5,805	2,763	277.4
Amitriptyline with Sert dock pose 4	1,815056333							
Amitriptyline with Sert dock pose 4	1,815056333							
Amitriptyline with Sert dock pose 4	1,815056333							
Amitriptyline with Sert dock pose 4	1,815056333							
Amitriptyline with Sert dock pose 4	1,815056333							

Image 57: Molecular Descriptors Snapshot from my personal excel file.

D. Python Code for Nested Cross-Validation and Feature Selection

The following code implements the full pipeline mentioned in Chapter 3.5.4 and 3.5.5 for multiclassification analysis using XGBoost with nested cross-validation. It includes:

- Preprocessing and standardization (although not necessary in XGBoost)
- Class weight balancing
- Feature selection with the aid of RFE
- Hyperparameter optimization using GridSearchCV
- Outer-loop evaluation with 5-fold stratified CV
- Evaluation Metrics as described in 3.5.5 chapter
- Robust feature identification across folds that are later used in Explainability section

Similar procedure was followed for all remaining classifiers.

CODE:

```
warnings.filterwarnings('ignore')

X = aggregated_df.drop(columns=["Labelled_class"])
y = aggregated_df["Labelled_class"]
classes = np.unique(y)
n_classes = len(classes)


class_weights = compute_class_weight(class_weight='balanced', classes=classes, y=y)
```

```
class_weight_dict = dict(zip(classes, class_weights))


def perform_rfe_and_hyperparam_cv(X_train, y_train):
    scaler = StandardScaler()
    X_train_scaled = pd.DataFrame(scaler.fit_transform(X_train), columns=X_train.columns)


    sample_weights = y_train.map(class_weight_dict)


    xgb_temp = XGBClassifier(n_estimators=100, random_state=42, eval_metric='mlogloss', use_label_encoder=False)
    xgb_temp.fit(X_train_scaled, y_train, sample_weight=sample_weights)

    importances = pd.Series(xgb_temp.feature_importances_, index=X_train.columns)
    top_features = importances.sort_values(ascending=False).head(30).index
    X_top = X_train_scaled[top_features]

    param_grid = {
        'max_depth': [4],
        'learning_rate': [ 0.05,0.1],
        'reg_alpha': [1,2],
        'reg_lambda': [4, 5],
        'gamma': [0.5, 0.8],
        'subsample': [0.7],
        'colsample_bytree': [0.7]
    }

    inner_cv = StratifiedKFold(n_splits=3, shuffle=True, random_state=42)

    best_score = -np.inf
    best_features = None
    best_params = None

    for n_features in range(15, 31, 5):
        rfe = RFE(estimator=xgb_temp, n_features_to_select=n_features, step=5)
        rfe.fit(X_top, y_train)

        selected_cols = X_top.columns[rfe.support_]
        X_rfe = X_top[selected_cols]

        sample_weights_rfe = y_train.map(class_weight_dict)
        xgb_model = XGBClassifier(n_estimators=100, random_state=42, eval_metric='mlogloss', use_label_encoder=False)
        gs = GridSearchCV(xgb_model, param_grid, cv=inner_cv, scoring='f1_macro', n_jobs=-1)
        gs.fit(X_rfe, y_train, sample_weight=sample_weights_rfe)

        if gs.best_score_ > best_score:
            best_score = gs.best_score_
            best_features = selected_cols
            best_params = gs.best_params_

    return list(best_features), best_params


outer_cv = StratifiedKFold(n_splits=5, shuffle=True, random_state=42)
all_y_test, all_y_pred, all_y_proba = [], [], []
fold_accuracies, fold_best_feature_lists, fold_best_params = [], [], []

for train_index, test_index in outer_cv.split(X, y):
    X_train, X_test = X.iloc[train_index], X.iloc[test_index]
    y_train, y_test = y.iloc[train_index], y.iloc[test_index]

    best_features, best_params = perform_rfe_and_hyperparam_cv(X_train, y_train)
    fold_best_feature_lists.append(best_features)
    fold_best_params.append(best_params)

    scaler = StandardScaler()
    X_train_scaled = scaler.fit_transform(X_train[best_features])
    X_test_scaled = scaler.transform(X_test[best_features])

    sample_weights_final = y_train.map(class_weight_dict)
    final_model = XGBClassifier(
        n_estimators=100,
        random_state=42,
        use_label_encoder=False,
        num_class=n_classes,
```

```
        objective='multi:softprob',
        eval_metric='mlogloss',
        **best_params
    )
    final_model.fit(X_train_scaled, y_train, sample_weight=sample_weights_final)

    y_pred = final_model.predict(X_test_scaled)
    y_prob = final_model.predict_proba(X_test_scaled)

    all_y_test.append(y_test)
    all_y_pred.append(y_pred)
    all_y_proba.append(y_prob)

    test_acc = accuracy_score(y_test, y_pred)
    fold_accuracies.append(test_acc)
    print(f"✅Fold training accuracy: {accuracy_score(y_train, final_model.predict(X_train_scaled)):.3f}, test accuracy: {test_acc:.3f}")


all_y_test = pd.concat(all_y_test)
all_y_pred = np.concatenate(all_y_pred)
all_y_proba = np.vstack(all_y_proba)

print(f"Nested CV Accuracy: {np.mean(fold_accuracies):.3f}")
print(f"Macro Precision: {precision_score(all_y_test, all_y_pred, average='macro'):.3f}")
print(f"Macro Recall:    {recall_score(all_y_test, all_y_pred, average='macro'):.3f}")
print(f"Macro F1:        {f1_score(all_y_test, all_y_pred, average='macro'):.3f}")
print("\nConfusion Matrix:\n", confusion_matrix(all_y_test, all_y_pred))
print("\nClassification Report:\n", classification_report(all_y_test, all_y_pred))


plt.figure(figsize=(8, 6))
y_test_bin = label_binarize(all_y_test, classes=classes)
for i in range(n_classes):
    fpr, tpr, _ = roc_curve(y_test_bin[:, i], all_y_proba[:, i])
    plt.plot(fpr, tpr, label=f"Class {i} (AUC = {auc(fpr, tpr):.2f})")
plt.plot([0, 1], [0, 1], "k--")
plt.legend(); plt.xlabel("FPR"); plt.ylabel("TPR"); plt.title("XGBoost ROC (One-vs-Rest)")
plt.show()


feature_counter = Counter([feat for feature_list in fold_best_feature_lists for feat in feature_list])
print("\nFeature selection frequency across folds:", feature_counter)

robust_features = [feat for feat, count in feature_counter.items() if count >= 3]
print("\nRobust features (selected in >=3 folds):", robust_features)

best_fold_index = np.argmax(fold_accuracies)
print(f"Best Outer Fold Index: {best_fold_index}, Accuracy: {fold_accuracies[best_fold_index]:.3f}")
print("Best Hyperparameters for that fold:", fold_best_params[best_fold_index])
```