



UNIVERSITY OF WEST ATTICA

FACULTY OF ENGINEERING

DEPARTMENT OF BIOMEDICAL ENGINEERING

**Development of machine learning
models to predict the activity of
chemical compounds against the
obesity-associated melanocortin-4
receptor**

GIAKOUMOPOULOU STAVROULA

Student ID: 19388019

Supervisor

Dr. Dionisios Cavouras, Emeritus Professor

Athens 01/10/2024



ΠΑΝΕΠΙΣΤΗΜΙΟ ΔΥΤΙΚΗΣ ΑΤΤΙΚΗΣ

ΣΧΟΛΗ ΜΗΧΑΝΙΚΩΝ

ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΒΙΟΪΑΤΡΙΚΗΣ

**Ανάπτυξη μοντέλων μηχανικής μάθησης
για την πρόβλεψη δραστηρότητας
χημικών ενώσεων έναντι του
σχετιζόμενου με την παχυσαρκία
υποδοχέα της μελανοκορτίνης 4**

ΓΙΑΚΟΥΜΟΠΟΥΛΟΥ ΣΤΑΥΡΟΥΛΑ

Αριθμός Μητρώου: 19388019

Επιβλέπων Καθηγητής

Δρ. Διονύσιος Κάβουρας, Ομότιμος Καθηγητής

Αθήνα 01/10/2024

Εξεταστική Επιτροπή

Supervisor

Διονύσιος Κάβουρας

Μίνως Ματσούκας

Ευτυχία Κρίτση

Καθηγητής Ομότιμος

[ΨΗΦΙΑΚΗ ΥΠΟΓΡΑΦΗ]

Επίκουρος Καθηγητής

[ΨΗΦΙΑΚΗ ΥΠΟΓΡΑΦΗ]

Επίκουρος Καθηγήτρια

[ΨΗΦΙΑΚΗ ΥΠΟΓΡΑΦΗ]

AUTHOR'S STATEMENT

The undersigned **Giakoumopoulou Stavroula** of **Charalampos**, with registration number **19388019** student of the Department of Biomedical Engineering, Faculty of Engineering of the University of West Attica, I hereby declare responsibly that:

“I am the author of this thesis and any help I have had in its preparation is fully acknowledged and referred to in the thesis. In addition, any sources from which I have made use of data, ideas or words, either directly or paraphrased are cited in their entirety, with full reference to the authors, publisher or journal, including any sources that may have been used from the internet. I, also, certify that this paper has been written by me exclusively and is the intellectual property of both myself and the Foundation.

Violation of the above academic responsibility constitutes substantial grounds for the revocation of my diploma”.

Date

01/10/2024

The declarant



ABSTRACT

Monogenic obesity caused by mutations in the Melanocortin-4 Receptor (MC4R) gene remains a significant health challenge, despite numerous efforts to find effective treatments. The MC4R is a promising target for drug development due to its role in energy homeostasis and adipose tissue formation. This thesis explores the hybridization of Machine Learning and Molecular Modeling techniques to identify potential ligands that may act as agonists against the obesity-associated MC4R.

This study aimed to develop a predictive model using a dataset of 1,906 chemical compounds and 208 RDKit molecular descriptors to classify molecules based on their activity against MC4R. Additionally, 2,000 natural compounds were evaluated using three molecular docking software platforms to identify potential ligands for human MC4R (hMC4R) based on interaction patterns and binding affinities. The machine learning model was used to predict ligand activity, and their properties were further analyzed using two ADMET tools.

The final model demonstrated strong efficiency of activity prediction, achieving 94.28% accuracy and an AUC of 0.98. The molecular docking experiments identified six natural compounds as potential hMC4R ligands, with most sharing the same chemical scaffold. However, the ADMET analysis did not yield accurate or reliable results, limiting the ability to fully assess the safety profiles of the identified compounds. Despite this limitation, the findings suggest that the flavone scaffold could serve as a template for designing novel agonists.

Keywords: Final project, Obesity, hMC4R, Machine Learning, Molecular Modeling, Molecular Docking, ADMET prediction, Ligands, Drug Design.

ACKNOWLEDGMENTS

First and foremost, I would like to thank my friends and family for their constant support throughout this long journey of writing this thesis. To my parents, thank you for always believing in me and encouraging me to move forward. Your sacrifices have made this possible.

I would like to express my gratitude to my supervising professor, Dr. Dionisios Cavouras, for his guidance in Machine Learning and support throughout the later years of my degree.

In addition, I would like to acknowledge Dr. Eftichia Kritsi for her patience, for always answering my questions, her valuable lessons and guidance in Computational Chemistry. You have been an immense help and I can't thank you enough!

Special thanks to Dr. Minos Matsoukas and PhD candidate Vasilios Panagiotopoulos for providing the data we worked with and for their constructive comments during my internship.

Last but not least, I extend my gratitude to Dr. Paraschos Christodoulou for his assistance with our final results.

Thank you all again for being a part of this journey!

Table of Contents

List of Figures	8
List of Tables	10
List of Abbreviations	11
Introduction.....	14
Chapter 1: Theoretical Background	16
1.1 The Disease of Obesity	16
1.2 Introduction to the Biological Target.....	18
1.2.1 Protein receptors	18
1.2.2 The melanocortin system	19
1.2.3 The melanocortin-4 receptor.....	21
1.2.4 Contemporary treatments.....	22
1.3 Fundamentals of Machine Learning	23
1.3.1 Supervised learning and classifiers	26
1.3.2 Data pre-processing	32
1.3.3 Feature selection and dimensionality reduction.....	33
1.3.4 Evaluation methods.....	36
1.4 Essential Tools for Statistical Analysis.....	38
1.5 Basic Concepts of Computational Chemistry	40
1.5.1 Chemical databases	41
1.5.2 Molecular representation	41
1.5.3 Molecular descriptors.....	44
1.5.4 Virtual screening.....	46
1.5.5 ADMET properties	53
1.6 Survey of Related Research.....	55
Chapter 2: Materials and Methods	57
2.1 Computational Tools.....	57
2.2 Machine Learning Model Development	57
2.2.1 Dataset.....	57
2.2.2 Practical implementation	59
2.3 Molecular Docking Experiments	61
2.3.1 Crystal structure preparation.....	61
2.3.2 Software applications.....	62
2.3.3 Validation process.....	64
2.4 Machine Learning and Docking Hybridization	65
2.5 ADMET Prediction.....	65
Chapter 3: Results	67
3.1 Machine Learning Results	67
3.1.1 Optimal feature combination	67
3.1.2 Model validation	68
3.2 Statistical Analysis Results	70
3.3 Molecular Docking Results.....	73
3.3.1 Protein-ligand binding pose prediction using PLIP	77
3.3.2 Protein-ligand binding pose prediction using Maestro	79
3.4 MetaboAnalyst Results	83
3.5 ADMET Results.....	84
Chapter 4: Discussion	88
4.1 Future Steps	90
Bibliography	91

List of Figures

Fig. 1.1: <i>Energy balance illustration</i>	17
Fig. 1.2: <i>Schematic representation of a GPCR.</i>	19
Fig. 1.3: <i>The chemical structure of endogenous ligand α-MSH.</i>	20
Fig. 1.4: <i>Theoretical model of the human MC4R with AgRP</i>	21
Fig. 1.5: <i>The structure of Setmelanotide</i>	23
Fig. 1.6: <i>Overfitting and underfitting illustrated</i>	24
Fig. 1.7: <i>Illustrations of the 4 main categories of Machine Learning</i>	25
Fig. 1.8: <i>Illustration of the KNN classifier</i>	27
Fig. 1.9: <i>Illustration of the Naïve Bayes classifier</i>	28
Fig. 1.10: <i>Illustration of the Logistic Regression classifier</i>	29
Fig. 1.11: <i>Illustration of the Decision Tree classifier</i>	30
Fig. 1.12: <i>Illustration of the Random Forest classifier</i>	30
Fig. 1.13: <i>Illustration of the SVM classifier</i>	31
Fig. 1.14: <i>Schematic representation of Perceptron</i>	32
Fig. 1.15: <i>ROC curve</i>	39
Fig. 1.16: <i>Horizontal box plot</i>	40
Fig. 1.17: <i>Graphical representation of ibuprophen (1D through 4D)</i>	42
Fig. 1.18: <i>Representative examples of SMILES</i>	43
Fig. 1.19: <i>Simplified fingerprint generation</i>	46
Fig. 1.20: <i>The datasets are collected from external sources</i>	49
Fig. 1.21: <i>Geometric representations of pharmacophoric features</i>	51
Fig. 1.22: <i>Example of Molecular Docking using CSFIR by Schrödinger’s computational platform (Maestro Glide)</i>	52
Fig. 1.23: <i>Sequence of methods</i>	53
Fig. 1.24: <i>Schematic representation of the various drug-distribution pathways</i>	54
Fig. 2.1: <i>Illustration of the ChEMBL database</i>	57
Fig. 2.2: <i>Illustration of the PDB</i>	61
Fig. 2.3: <i>6W25: 3D crystal structure of MC4R in complex with SHU9119 (left) and chemical structure of SHU9119 (right)</i>	62
Fig. 3.1: <i>ROC curve of the optimal feature combination, using RandomForestClassifier over 10 epochs (AUC=0.98)</i>	69
Fig. 3.2: <i>MaxAbsEstateIndex boxplots (left) and ROC curve (AUC=0.89) between classes (right)</i>	71
Fig. 3.3: <i>PEOE_VSA8 boxplots (left) and ROC curve (AUC=0.88) between classes (right)</i>	71
Fig. 3.4: <i>Kappa2 boxplots (left) and ROC curve (AUC=0.91) between classes (right)</i>	72
Fig. 3.5: <i>BCUT2D_MRLOW boxplots (left) and ROC curve (AUC=0.91) between classes (right)</i>	72
Fig. 3.6: <i>SHU9119 in the binding site of MC4R (PDB: 6W25), as visualized using Maestro</i>	73
Fig. 3.7: <i>Validation. Optimal superimposition of the 3D crystal configurations with the docked structure of SHU9119 using Glide-SP (left) and Glide-XP (right).</i>	75
Fig. 3.8: <i>ZINC000000487423</i>	76
Fig. 3.9: <i>ZINC000004349406</i>	76
Fig. 3.10: <i>ZINC000169724085</i>	77
Fig. 3.11: <i>ZINC000299817569</i>	77

Fig. 3.12: ZINC000095913431	77
Fig. 3.13: ZINC000095913799	77
Fig. 3.14: Basic structure and numbering system of flavonoids (left) and flavones (right)	77
Fig. 3.15: Representative 3D binding pose of ZINC000000487423 at MC4R binding site (PDB: 6W25) using PLIP	78
Fig. 3.16: Representative 3D binding pose of ZINC000004349406 at MC4R binding site (PDB: 6W25) using PLIP	21
Fig. 3.17: Representative 3D binding pose of ZINC000169724085 at MC4R binding site (PDB: 6W25) using PLIP	78
Fig. 3.18: Representative 3D binding pose of ZINC000299817569 at MC4R binding site (PDB: 6W25) using PLIP	78
Fig. 3.19: Representative 3D binding pose of ZINC000095913431 at MC4R binding site (PDB: 6W25) using PLIP	79
Fig. 3.20: Representative 3D binding pose of ZINC000095913799 at MC4R binding site (PDB: 6W25) using PLIP	79
Fig. 3.21: Representative 2D Ligand Interaction Diagram (left) and 3D binding pose (right) of ZINC000000487423 at MC4R binding site (PDB: 6W25) using Maestro ...	80
Fig. 3.22: Representative 2D Ligand Interaction Diagram (left) and 3D binding pose (right) of ZINC000004349406 at MC4R binding site (PDB: 6W25) using Maestro ...	80
Fig. 3.23: Representative 2D Ligand Interaction Diagram (left) and 3D binding pose (right) of ZINC000169724085 at MC4R binding site (PDB: 6W25) using Maestro ...	81
Fig. 3.24: Representative 2D Ligand Interaction Diagram (left) and 3D binding pose (right) of ZINC000299817569 at MC4R binding site (PDB: 6W25) using Maestro ...	81
Fig. 3.25: Representative 2D Ligand Interaction Diagram (left) and 3D binding pose (right) of ZINC000095913431 at MC4R binding site (PDB: 6W25) using Maestro ...	82
Fig. 3.26: Representative 2D Ligand Interaction Diagram (left) and 3D binding pose (right) of ZINC000095913799 at MC4R binding site (PDB: 6W25) using Maestro ...	82
Fig. 3.27: Predicted class probabilities (left) and box plot of predictive accuracy (right) using MetaboAnalyst.	83
Fig. 3.28: ROC curve for 100 cross-validations using MetaboAnalyst	84

List of Tables

Table 1.1	<i>Classification of patients according to the BMI</i>	16
Table 1.2	<i>Examples of genes related to monogenic obesity development</i>	18
Table 1.3	<i>The five receptors of the melanocortin system</i>	20
Table 1.4	<i>Variance and Bias</i>	24
Table 1.5	<i>Binary and Multiclass classification</i>	26
Table 1.6	<i>Truth Table of a binary classification problem</i>	37
Table 1.7	<i>Canonical SMILES symbolism of chemical bonds</i>	43
Table 1.8	<i>Examples of common MDs</i>	44
Table 1.9	<i>Paradigms of FPs types</i>	44
Table 1.10	<i>Similarity search based on the Tanimoto coefficient</i>	47
Table 2.1	<i>Representative examples of RDKit molecular descriptors</i>	58
Table 2.2	<i>List of employed classifiers</i>	60
Table 3.1	<i>Optimal feature combination</i>	67
Table 3.2	<i>Average performance metrics from K-fold cross-validation for the RandomForestClassifier</i>	68
Table 3.3	<i>Average performance metrics over 10 epochs for the RandomForestClassifier</i>	69
Table 3.4	<i>Average performance metrics (mean and standard deviation) over 10 epochs for all classifiers</i>	70
Table 3.5	<i>MC4R-SHU9119 (PDB: 6W25) interaction pattern</i>	73
Table 3.6	<i>Docking Scores (kcal/mol) of the final selection of chemical compounds with MC4R (PDB: 6W25)</i>	74
Table 3.7	<i>Interaction patterns of the final selection of chemical compounds with MC4R (PDB: 6W25)</i>	75
Table 3.8	<i>Prediction of final natural compounds</i>	83
Table 3.9	<i>Setmelanotide ADMET prediction</i>	85
Table 3.10	<i>ZINC000169724085 ADMET prediction</i>	85
Table 3.11	<i>ZINC000299817569 ADMET prediction</i>	86
Table 3.12	<i>ZINC000095913431 ADMET prediction</i>	86

List of Abbreviations

ABBREVIATION	DEFINITION
7TM	7 Transmembrane
ACC	Acetyl-CoA Carboxylase
ACTH	Adrenocorticotropic Hormone
Ada.	Ada Boost
ADMET	Absorption, Distribution, Metabolism, Excretion, and Toxicity
AgRP	Agouti-Related Protein
AgRP/NY	Agouti-Related Peptide/Neuropeptide Y
ANN	Artificial Neural Network
AR	Aromatic
ASCII	American Standard Code for Information Interchange
AUC	Area Under the Curve
Bayes.	Bayesian
BBB	Blood Brain Barrier
BCUT	Burden-Cas-University of Texas
BMI	Body Mass Index
cAMP	Cyclic Adenosine Monophosphate
CART	Classification and Regression Tree
CNS	Central Nervous System
CSV	Comma Separated Values
DILI	Drug Induced Liver Injury
EBI	European Bioinformatics Institute
EMBL	European Molecular Biology Laboratory
FDA	Food and Drug Administration
FP	Fingerprint
GBT	Gradient Boosting
Glide	Grid-based Ligand Docking with Energetics

GPCR	G-Protein Coupled Receptor
GUI	Graphical User Interface
H	Hydrophobic
HB	Hydrogen Bond
HBA	Hydrogen Bond Acceptor
HBD	Hydrogen Bond Donor
HI	Hydrophobic Interactions
HIA	Human Intestinal Absorption
hMC4R	Human Melanocortin-4 Receptor
IDE	Integrated Development Environment
II	Inductive Interactions
InChI	IUPAC International Chemical Identifier
KNN	K-Nearest Neighbor
LBP	Ligand Binding Pocket
LDA	Linear Discriminant Analysis
LogReg.	Logistic Regression
MC4R	Melacortin-4 Receptor
MD	Molecular Descriptor
MP	Multilayer Perceptron
MSH	Melanocyte Stimulating Hormone
MT-II	Melanotan-II
n-D	n-Dimensional
NMR	Nuclear Magnetic Resonance
NPI	Non-Polar Interactions
PCA	Principal Component Analysis
PDB	Protein Data Bank
PDBQT	Protein Data Bank, Partial Charge (Q) and Atom Type (T)
PEOE	Partial Equalization of Orbital Electronegativity
Percep.	Perceptron

PI	Positive Ionizable
pi-pi	π - π interactions
PLIP	Protein-Ligand Interaction Profiler
POMC	Pro-opiomelanocortin
POMC/CART	Pro-opiomelanocortin/cocaine-amphetamine related transcript
QSAR	Quantitative Structure-Activity Relationship
RCSB PDB	Research Collaboratory for Structural Bioinformatics Protein Data Bank
RF	Random Forest
RFE	Recursive Feature Elimination
RMSD	Root Mean Squared Deviation
ROC	Receiver Operating Characteristic
SAR	Structure-Activity Relationship
SB	Salt Bridge
SDF	Structure Data File
SMILES	Simplified Molecular Input Line Entry System
SVM	Support Vector Machine
UCSF	University of California, San Francisco
USEPA	US Environmental Research Laboratory
VS	Virtual Screening
VSA	Van der Waals Surface Area
WHO	World Health Organization
XGB	Extreme Gradient Boosting
XVOL	Exclusion Volume

Introduction

In the present world, where technological advancements play a pivotal role in shaping daily life, it is possible to address challenges that were once considered impossible. The adaptation and utilization of these advancements in biomedical research led to discoveries aimed primarily at enhancing the quality of human life. At the core of this research area are Machine Learning techniques. More specifically, the use of Machine Learning techniques to study the properties, characteristics, and behavior of chemical compounds significantly impacts the discovery of innovative drugs due to their enhanced computational power and accuracy. Furthermore, merging Machine Learning techniques with Computational Chemistry tools accelerates and guides the rational search of novel pharmaceutical substances.

Obesity is among the diseases that pose a public health challenge and demands urgent attention and intervention. Undeniably, obesity has risen to become one of the most widespread chronic health issues globally, posing an increased risk of morbidity. Moreover, the percentage of the population affected by this condition has steadily increased over the years. It is responsible for the development of life-threatening conditions, including diabetes mellitus, arterial hypertension, and myocardial infarction. The factors contributing to the disease and the forms in which it occurs vary. Of particular interest is obesity resulting from inherited disorders and gene mutations. A typical example is monogenic obesity, which is regarded as a rare and severe form of the condition triggered by a mutation in a specific gene.

Despite the rarity of the phenomenon, the mutated melanocortin-4 receptor (MC4R) is responsible for a higher proportion of obesity in adults and children due to genetic factors. As a consequence, researchers are focusing on mutations in the gene encoding MC4R and subsequently, on the functionality of the receptor.

These reasons render MC4R an attractive potential target for developing and clinically testing drug therapies for hereditary obesity. Current studies focus on identifying agonists, which are molecules that bind to the receptor and trigger its activation, resulting in a biological response akin to that of the natural signaling molecule.

In this thesis, virtual screening techniques are employed alongside the development of Machine Learning models to predict the activity of chemical compounds against the obesity-associated MC4R. In particular, through the utilization of suitable data processing methods, Machine Learning, and Statistical Analysis techniques, this study aims to identify the molecular descriptors (biomarkers) of chemical compounds that distinctly differentiate active compounds from non-active ones. Subsequently, using a crystal structure of MC4R with the confirmed ligand SHU9119, potential active compounds against the receptor will be examined through the application of Molecular Docking. Lastly, the outcomes of both procedures will be correlated and further assessed based on the Absorption, Distribution, Metabolism, Excretion, and Toxicity (ADMET) criteria.

Chapter 1 outlines the theoretical concepts and methodologies employed in this study. Specifically, it explores the obesity epidemic, analyzing the biological target and its relevance to human health. Fundamental concepts of Machine Learning and Statistical Analysis are also explained. Additionally, the chapter elucidates Computational Chemistry tools, including virtual screening and ADMET prediction.

Chapter 2 expounds on the materials and the procedures used in this study to

analyze the available data and obtain results. It outlines the selection and processing of the dataset to prepare it for Machine Learning algorithms and subsequent Statistical Analysis. Additionally, it details the process of molecular docking of potential ligands to MC4R and how these results are correlated with the developed machine-learning models for molecule qualification. Last but not least, it describes the further assessment of their metabolic properties.

Chapter 3 presents the results derived from the applied computational procedures. In more detail, diagrams illustrating the results of machine learning and statistical analysis on the optimal model are provided. Molecular docking results are presented in summary tables and through representative poses of the chemical compounds within the receptor's binding site. Selected ADMET properties are also highlighted.

Chapter 4 provides a brief discussion of our findings and suggests potential future steps to enhance our research.

Limitations and challenges in this thesis primarily stemmed from the materials used. The large dataset size led to time-consuming computations during the development of machine learning models. Similarly, one of the available molecular docking software platforms required significant computational power, resulting in a slow docking process.

Chapter 1: Theoretical Background

1.1 The Disease of Obesity

In recent decades, research conducted in the domain of health has demonstrated the exponential growth of obesity across a wide range of the human population worldwide. According to the World Health Organization (WHO), since 1948, obesity has been defined as a chronic and complex disease characterized by excess deposits of adipose tissue in the body, subsequently affecting an individual's health. It has been demonstrated to have a direct correlation with a higher probability of developing diabetes mellitus, cardiovascular disease including myocardial infarction and arterial hypertension, deterioration of both bone and reproductive health, and various types of cancer.

A practical measure to estimate body weight based on clinical observation is the Body Mass Index (BMI). This indicator was established as a criterion for evaluating the risk and likelihood of diseases associated with increased accumulation of adipose tissue in the body, as previously stated. BMI is calculated using the body mass measured in kilograms (kg) and the height of the patient calculated in meters (m). Understandably, it does not take into consideration parameters such as age or gender. As a consequence, it is expected to overestimate or underestimate the presence of obesity in some cases. However, the long-term practice of this method has resulted in a plethora of available data thus permitting comparisons and conclusions based on age and gender. To this day, it continues to be a commonly utilized technique for categorizing obese patients by experts, due to its simplicity and lack of invasiveness.

Table 1.1 presents the classification of patients according to the BMI value [1].

$$BMI = \frac{m \text{ (kg)}}{h^2 \text{ (m}^2\text{)}}, \quad (1.1)$$

where m is the mass of the body and h is the height of the individual.

Table 1.1 Classification of patients according to the BMI.

Classification	BMI (kg/m ²)	Risk
Underweight	<18.5	Low
Normal weight	18.5 – 24.9	Normal
Overweight	25.0 – 29.9	Increased
Obesity (Stage I)	30.0 – 34.9	Average
Obesity (Stage II)	35.0 – 39.9	High
Obesity (Stage III)	>40	Extremely high

Following statistical studies implemented by WHO in the year 2022, it is estimated that 43% of adults in the world are identified as overweight, while 16% are

classified as obese, demonstrating a more than doubling of rates for a period of 3 decades. Remarkable rates are also evident in children, in comparison to previous years, with rates of approximately 20% [2]. In European countries, around 60% of adults fall into the overweight and obese category, whereas the corresponding figure for children is 8%. In addition, it is estimated that excess weight beyond the normal range is responsible for over 1.2 million deaths each year in Europe [3].

The rapid progression of the probability of occurrence of the disease in the general population has led to the necessity of investigating the causes of the phenomenon. Scientific research to date reinforces the notion that obesity is a multifactorial disease as environmental, behavioral, and biological causes contribute to the establishment of a positive energy balance (Figure 1.1). Both environmental and behavioral causes are associated with changes in energy balance. More specifically, when energy intake exceeds expenditure over time, the body accumulates fat, leading to weight gain. Biological factors hinder the attainment of energy balance, further increasing its complexity. For instance, biological factors may include neuroendocrine disorders or alterations in the gut microbiome. Nonetheless, researchers are particularly interested in the impact of genetic mutations on the onset and course of this condition [4].

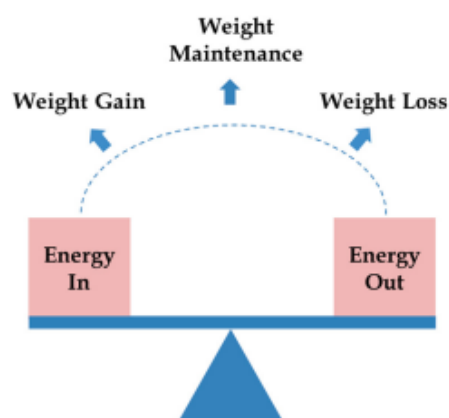


Fig. 1.1: Energy balance illustration [4].

Further investigation on the influence of the genetic code on metabolic disorders has led to the discovery of over 250 obesity-related genes and gene markers, documenting the existence of a genetic predisposition [5]. Therefore, obesity from a genetic perspective is divided into three subgroups: monogenic, polygenic, and syndromic obesity. Monogenic obesity stems from mutations in a single gene, polygenic obesity results from a combination of mutations of multiple genes, and syndromic obesity is associated with inherited disorders, like intellectual disability [6].

Monogenic obesity is classified as the rarest and most severe form of the disease, due to the reduced feeling of satiety and increased appetite. The first evidence demonstrating the association between weight change and a defective gene was presented in 1997 by Montague *et al.* [7]. More precisely, the researchers reported an association between mutations in the gene responsible for encoding leptin, identifying it as a crucial regulator of energy homeostasis. So far, about 20 individual, gene-related disorders leading to autosomal obesity have been documented.

The table below (Table 1.2) summarizes the prevalently acknowledged genes associated with the development of monogenic obesity. It has been observed that these gene mutations are located in the leptin – melanocortin pathway of the Central Nervous System (CNS) and are critical in the regulation of energy homeostasis [4], [6], [8].

Table 1.2 *Examples of genes related to monogenic obesity development.*

Gene	Encoded Protein	Main Function	Classification	Statistics
<i>LEP</i>	Leptin	Protein hormone produced by adipocytes that regulates energy intake	Severe type, occurs in the early days of life	Less than 100 patients, worldwide
<i>LEPR</i>	Receptor of leptin	Binds leptin and activates the synthesis of POMC	Severe type, occurs in the early days of life	About 2-3% of severe, early obesity cases
<i>POMC</i>	Pro-opiomelanocortin	A precursor polypeptide of melanocortins	Severe type, occurs in the early months of life	Less than 10 patients, worldwide
<i>MC4R</i>	Receptor of melanocortin 4	Binds α -MSH, it is expressed in the hypothalamus and activates anorexigenic signals	Severe type, occurs during childhood	About 2-3% of obesity cases in children and adults
<i>PCSK1</i>	Proprotein Convertase Subtilisin/Kexin Type 1	Participates in the biosynthesis of insulin	Severe type, occurs during childhood	Less than 20 patients worldwide

1.2 Introduction to the Biological Target

1.2.1 Protein receptors

Proteins are biochemical compounds that are among the most versatile macromolecules in living systems. They play vital roles in numerous biological processes, providing mechanical support, catalyzing biochemical reactions, and facilitating movement. The diverse functions of proteins arise from both the variety and the specialized roles of their constituent building blocks, known as amino acids. Amino acids are interconnected by peptide bonds, forming polypeptide chains and define the complex, three-dimensional (3D) structure of proteins. They consist of a central carbon atom, an amino group, a carboxyl group, a hydrogen atom and a side group (R), which is different for each amino acid [9], [10].

Cell boundaries are defined by intrinsically impermeable barriers known as membranes, which prevent molecules produced inside the cell from escaping and foreign molecules from entering. Despite this barrier, transport systems exist, making it possible to regulate the movement of specific molecules, allowing certain substances to enter and others to be expelled. These systems rely on membrane proteins, such as pumps and channels. However, the cell membrane remains impermeable to larger, polar signaling molecules, like hormones. To receive signals from the external environment, different types of specialized, integrated membrane-bound proteins called receptors interact with these signaling molecules and transmit the information into the cell's interior.

A receptor typically consists of extracellular and intracellular structural domains. The signaling molecule, known as a ligand, is recognized by a specific binding site on the receptor's extracellular side. This interaction leads to the formation of the receptor-ligand complex, resulting in both structural changes in the receptor and the intracellular region. These alterations induce changes in the concentration of small molecules, called second messengers, which relay the information to produce a response from the cell. One critical second messenger is cyclic adenosine monophosphate (cAMP). Before its termination, the signaling process modifies the function of molecules that directly regulate metabolic pathways, gene expression, and even the transmission of nerve impulses.

One of the most significant categories of protein receptors is the seven-transmembrane (7TM) receptors, which penetrate the membrane's lipid bilayer seven times (Figure 1.2). Upon binding with a ligand, these receptors change their stereochemistry, subsequently activating G proteins (G from guanylonucleotide), hence the term G protein-coupled receptors (GPCRs). When activated, G proteins stimulate the activity of adenylate cyclase, which increases the concentration of cAMP. Mutations in the genes encoding these receptors are associated with multiple diseases. As a result, the majority of drugs target the receptors to modify their function [9].

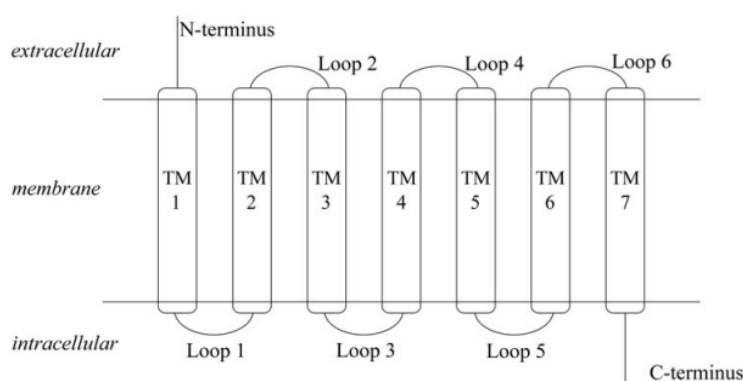


Fig. 1.2: Schematic representation of a GPCR. There is a N-terminal extracellular region (the end with a free amino group) and three extracellular loops interacting with a ligand molecule along with a C-terminal intracellular region (the end with a free carboxyl group) and three intracellular loops interacting with a G protein [11-12].

1.2.2 The melanocortin system

The melanocortin system is defined as an extremely important and complex component of human physiology.

It consists of melanocortins, a group of peptide hormones derived from the post-translational cleavage of the gene product of pro-opiomelanocortin (POMC) [13]. The most prevalent melanocortins are melanocyte-stimulating hormones (α -MSH, β -MSH, γ -MSH) and adrenocorticotrophic hormone (ACTH). They actively participate in various physiological functions within the human body, including skin pigmentation, immune function, and appetite regulation. A common feature of melanocortins is the amino acid sequence His-Phe-Arg-Trp, a pharmacophore essential for the biological activity of these peptides [14].

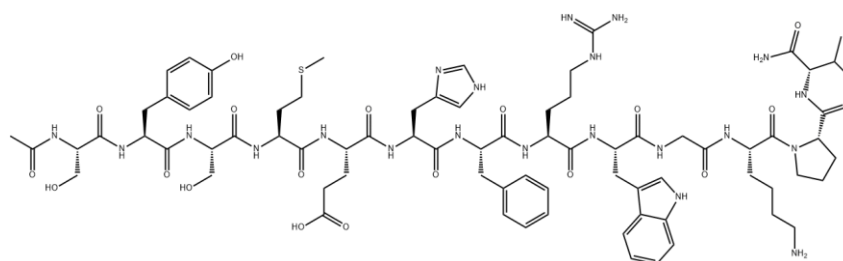


Fig. 1.3: The chemical structure of endogenous ligand α -MSH.

The protein receptors within the melanocortin system are part of the larger family of GPCRs. These receptors originate from five distinct genes that encode them, each bearing a corresponding name (MC1R through MC5R). Research conducted so far indicates both their different tissue distribution and functionality. Melanocortin receptors are activated by one or more melanocortin peptides (α -MSH, β -MSH, γ -MSH, ACTH). These signaling molecules lead to an increase in cAMP concentration, regulating the functions associated with each receptor of this system [13].

In contrast to melanocortins, which act as receptor agonists, the melanocortin system features two endogenous antagonists: the agouti and agouti-related protein (AgRP) peptides. They are the only inhibitory peptides identified among the 7TM family, distinguished by their selectivity [14].

Table 1.3 briefly presents the five subtypes of melanocortin receptors. It is worth noting that all of the melanocortin receptors are associated with cAMP generation via the stimulatory G protein G_s and adenylate cyclase.

Table 1.3 The five receptors of the melanocortin system.

Receptor	Expression	Main Function	Agonist	Antagonist
MC1R	Melanocytes, skin glands, hair follicle, testis, pituitary etc	Skin and hair pigmentation, inflammation	α -MSH, ACTH	Agouti
MC2R	Adrenal cortex, skin, adipocytes	Steroidogenesis	ACTH	Agouti
MC3R	Brain, placenta, testis, heart, gut	Energy homeostasis	α -MSH, β -MSH, γ -MSH, ACTH	Agouti, AgRP

Table 1.3 (Continued).

<i>MC4R</i>	Brain, adipose tissue	Energy homeostasis, sexual behavior	α -MSH, β -MSH, ACTH	Agouti, AgRP
<i>MC5R</i>	Adrenal gland, kidney, adipose tissue, lymph node, lung, testis, uterus etc	Exocrine function (sebaceous gland secretion)	α -MSH, ACTH	-

This thesis investigates the melanocortin-4 receptor (MC4R).

1.2.3 The melanocortin-4 receptor

MC4R gene, which is located on chromosome 18q21.3, encodes the melanocortin-4 receptor (MC4R), member of the melanocortin protein receptor family [15]. This receptor (Figure 1.4) is a GPCR consisting of 332 amino acids. It plays a central role in regulating appetite, maintaining energy balance, and influencing the formation of adipose tissue in the body, as indicated by pharmacological and genetic studies in both animal and human models [16].

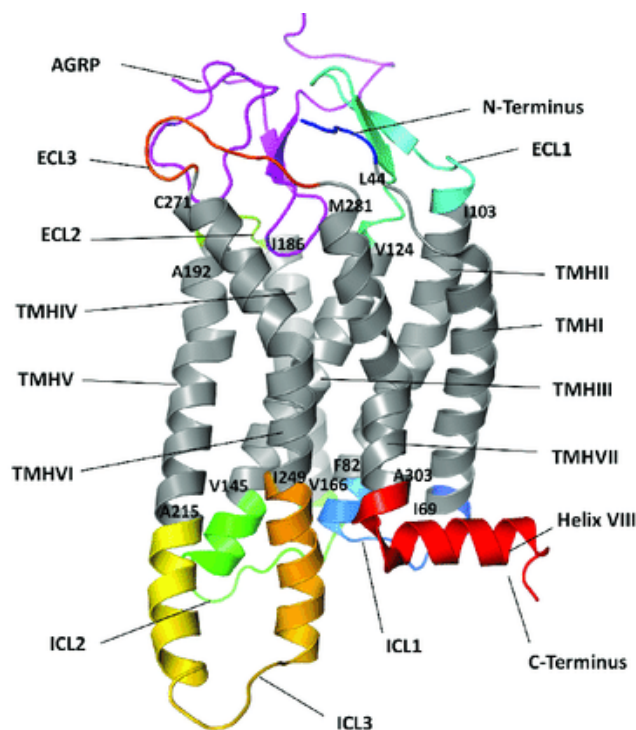


Fig. 1.4: Theoretical model of the human MC4R with AgRP. As evident, N-terminus is represented in dark blue, C-terminus in red, and the seven transmembrane helices (TMH) in gray. The extracellular (ECL) and intracellular (ICL) loops are represented by different colors. The boundary amino acids of TMH are labeled according to the protein sequence alignment (adapted from [17], [18]).

It is mainly expressed in the CNS and more specifically, in the paraventricular nucleus of the hypothalamus of the brain. Receptor activity is regulated by two functionally antagonistic populations of neurons associated with eating behavior:

- i. The anorexigenic pro-opiomelanocortin/cocaine-amphetamine related transcript (*POMC/CART*) neurons.
- ii. The orexigenic agouti-related peptide/neuropeptide Y (*AgRP/NPY*) neurons [19].

Agonists, such as α -MSH, activate the receptor and regulate the feeling of satiety. In contrast, the endogenously produced AgRP antagonist inhibits the receptor's activity and leads to an increase in levels of appetite.

The consequences of inhibiting MC4R function, such as hyperphagia, hyperinsulinemia, and hyperglycemia, were demonstrated in 1997, by Huszar *et al.* [20], in a study conducted on animal models of mice. Over the years, research on the human MC4R and its mutations has elucidated the receptor's role in adipose tissue formation and energy balance regulation. So far, over 100 different mutations of MC4R gene have been documented in large samples of obesity-affected populations, making it the most prevalent form of the disease stemming from genetic disorders [21].

In conclusion, understanding the role of MC4R in the regulation of both energy balance and metabolism, in general, is significant in domains of research related to the discovery of targeted, therapeutic interventions against monogenic obesity.

1.2.4 Contemporary treatments

Over three decades ago, research efforts aimed at discovering pharmacotherapy for this type of monogenic obesity, centering on the utilization of regulatory hormones, specifically MSH. However, these hormones, as discerned in Table 1.3, lack specificity concerning the melanocortin receptor subtype. In addition, they demonstrated adverse side effects in clinical trials, such as skin hyperpigmentation and Addison's disease. Consequently, this proposal was ultimately dismissed. This indication created an urgent need to identify molecules exhibiting similar activity to endogenously produced agonists, characterized by a higher degree of selectivity. The molecules considered as potential agonists are divided into three categories:

- i. Linear peptide agonists.
- ii. Cyclic peptide agonists.
- iii. Non-peptide agonists.

It is noted that linear peptides, due to their structure, demonstrate a significant resemblance to melanocortin peptide hormones, but cyclic peptides exhibit stronger interactions with the receptor. In addition, non-peptide molecules possess high levels of selectivity.

Some potential MC4R ligands, such as melanotan II, despite their promising results during computational studies, failed in clinical trials [16].

In the year 2020, the first pharmaceutical treatment was approved by the Food and Drug Administration (FDA), targeting obesity originating from genetic

predisposition [22]. Setmelanotide (also known as RM-49 or commercially as *Imcivree*) may be prescribed for individuals with POMC, PCSK1, and LEPR deficiencies, who are seeking weight management solutions. This treatment is suitable for children (aged 6 and above) and adults, who carry these mutations in the leptin-melanocortin pathway. It is a cyclic peptide, showing 20 times higher selectivity towards MC4R, compared to the natural ligand α -MSH. In addition, as noted by Hammad *et al.* [21], Setmelanotide presents a higher binding affinity to MC4R, in contrast with the endogenous form of α -MSH. Unlike previous clinically tested molecules, this cyclic peptide does not exhibit cardiovascular side effects such as tachycardia or increased blood pressure, highlighting its safety profile for therapeutic use [22].

Figure 1.5 illustrates the chemical structure of Setmelanotide.

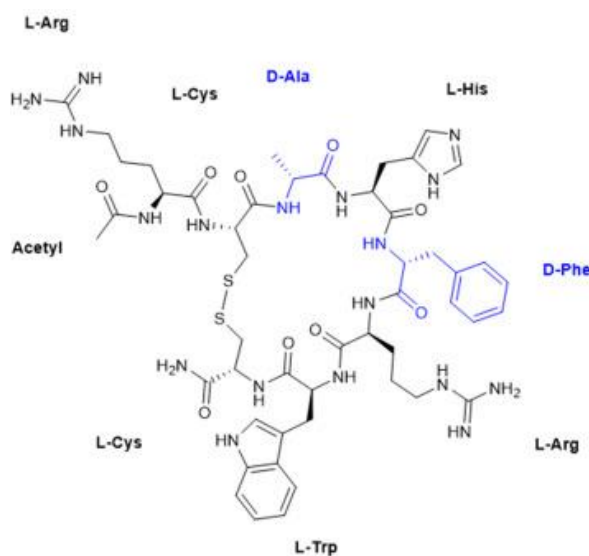


Fig. 1.5: The structure of Setmelanotide [23].

1.3 Fundamentals of Machine Learning

The term "Machine Learning" encompasses the scientific field dedicated to the various ways in which computers are trained, learn, and adapt based on data. It is a fusion of computer science and mathematics, arising from the computational challenges of constructing predictive models when analyzing vast quantities of data. Machine learning is based heavily on statistical concepts, including probability, but excels in capabilities, such as decision-making [24]. The process of designing and developing a Machine Learning model consists of 3 stages: training, testing and final evaluation.

In general, when referring to "models" in the context of Machine Learning, a specification of a mathematical (or probabilistic) relationship that exists between different variables is implied [25].

A dataset is a collection of observations, known as data, organized in a way that facilitates processing, manipulation, and analysis. Typically, data within a dataset originate from authoritative sources of scientific literature. Datasets, which are utilized for research purposes, are retrieved from databases. A database is defined as a digital repository for the storage, organization, and efficient management of a large amount of information.

The information within a dataset that supplies the algorithms is referred to as features. The quality and the number of features utilized to train a Machine Learning model are directly related to its performance and the accuracy of its predictive capabilities [24]. Before constructing a reliable model, it is necessary to consider these parameters to avoid certain risks. For instance, overfitting is a frequent challenge in Machine Learning. A model performs well on the data used its training but struggles to generalize, leading to poor performance on new data. This phenomenon is characterized by high variance in the model, stemming from a large number of parameters, rendering the model overly complex. This means that the model identifies specific inputs rather than the factors that contribute to predicting the desired output. Likewise, a model may present underfitting, i.e. high bias. This indicates that it is not complex enough to capture the pattern of the training data, so it performs poorly on new data [25], [26].

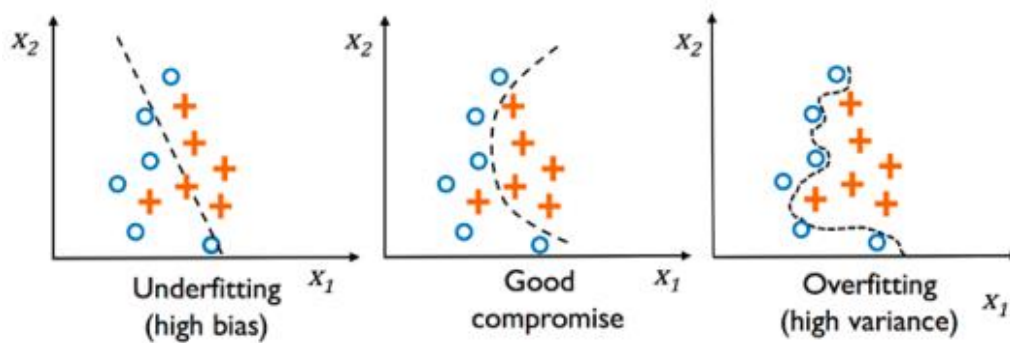


Fig. 1.6: *Overfitting and underfitting illustrated by the comparison of a linear decision boundary to non-linear decision boundaries of higher complexity [26].*

Table 1.4 provides the definitions of the terms “Variance” and “Bias” [26].

Table 1.4 *Variance and Bias.*

Terminology in Machine Learning	
<i>Variance</i>	A consistency (or variability) measure of a model's prediction for the classification of a particular example, after retraining.
<i>Bias</i>	A systematic error measure, unrelated to randomness, i.e. how much the predictions deviate from the correct values.

In general, there are 4 main categories of Machine Learning. These are Supervised Learning, Unsupervised Learning, Semi-supervised Learning, and Reinforcement Learning [27].

- **Supervised Learning**

The purpose of Supervised Machine Learning is to train a model using a set of examples as data inputs, which are categorized into predefined classes. More specifically, labeled training data are utilized to create a model capable of predicting

reliable outcomes for future, unidentified (unlabeled) data [26], [28]. This category consists of two main methods: classification and regression. Classification involves assigning data to discrete classes based on available characteristics. On the other hand, regression focuses on predicting continuous values, rather than discrete ones, according to a given set of characteristics [24], [29].

- **Unsupervised Learning**

In contrast to Supervised Learning, where data labels are known before model training, Unsupervised Learning creates predictive models and operates without pre-existing data labels or information about data structure. Different algorithms are applied to identify similarities and differences, clustering data based on naturally occurring patterns. In other words, clustering algorithms are exploratory data analysis techniques, employed to categorize objects into groups according to their degree of similarity. Since it is a data-driven process, it does not require any intervention from a human observer [24].

- **Semi-supervised Learning**

This category is a combination of the two previous types of machine learning. It operates using both labeled and unlabeled data. The main goal is to create an algorithm characterized with higher performance than the one achieved through the application of Supervised Learning alone [29]. This type of model training is used when some missing outputs can be approximated using available training data [24].

- **Reinforcement Learning**

The aim in Reinforcement Learning is to develop a system, known as agent, which automatically improves its performance by interacting with a specific environment through trial and error. This type of learning is based on rewards and penalties, aiming to increase these rewards or minimize risks [24]. Since information about the current state of the environment typically includes the reward signal, this method can be considered a field related to Supervised Learning [26].

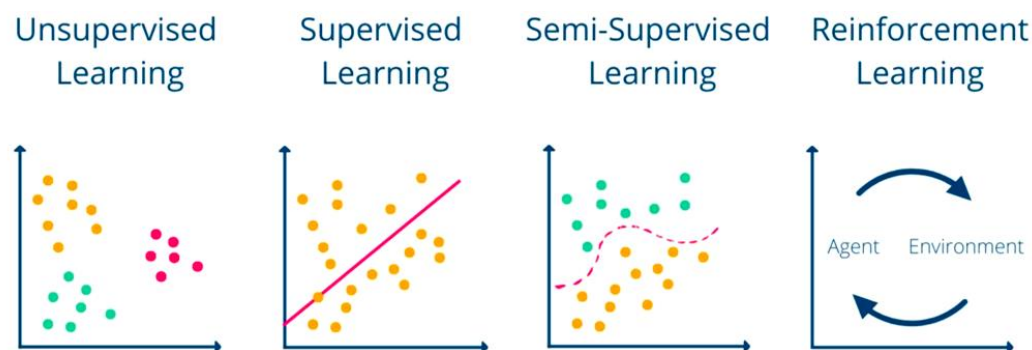


Fig. 1.7: Illustrations of the 4 main categories of Machine Learning [30].

In conclusion, the appropriate selection of a Machine Learning technique depends on factors such as the type of available data, the field of application, and the objectives pursued each time. These ensure the development of an efficient, high-performance model capable of supporting new data.

1.3.1 Supervised learning and classifiers

Machine Learning in Data Science is evolving into an exceptionally important research tool. It finds extensive application in a multitude of scientific domains and industries for data analysis. For such purposes, this thesis utilizes Supervised Learning methodologies.

A classifier is characterized as a decision-making algorithm, or a mathematical model, designed to sort input data into one of the C_i ($i=1, 2, 3\dots n$) categories. As previously noted, inputs are regarded as features, commonly presented in the form of a vector [31], [32].

The most common classification problems are listed in Table 1.5 [29].

Table 1.5 *Binary and Multiclass classification.*

Types of Classification	
<i>Binary</i>	Two classes of data available, such as "true-false", "normal-malignant", "active-non-active".
<i>Multiclass</i>	Multiple classes of data available ($n>2$).

It is worth noting that Supervised Learning algorithms are divided into two subcategories: parametric and non-parametric models. In parametric models, parameters are calculated from the training data to create a function. As a result, this function classifies the unknown data, without requiring the assistance of a training set. Such algorithms include the Perceptron, Logistic Regression, Linear Support Vector Machine (SVM), and Linear Discriminant Analysis (LDA). On the contrary, non-parametric models lack a fixed set of parameters, are flexible, and constantly adapt to the complexity of the data. Examples include K-Nearest Neighbors (KNN), SVM with nonlinear kernels, Decision Tree/Random Forest [26].

Supervised Learning classifiers that appear frequently in scientific literature are presented below:

- **K-Nearest Neighbors (KNN)**

The Nearest Neighbors classifier is one of the simplest prediction models available. Quite deliberately, they disregard much of the information, as the prediction for each new record depends on a few points closest to it. The KNN is often characterized as a "lazy learner" since it does not train a predictive model. Instead, it stores the training data and makes predictions based on the set of k -similar instances ($k=1, 2\dots n$, where k is a predetermined number). A measure commonly used is the Euclidean distance. Each new observation is assigned to a class according to the majority vote of k -nearest neighbors. In this approach, the classifier is continuously adjusted during data collection, while its accuracy is dependent on the dataset's quality. Admittedly, it is an algorithm of computational complexity, which increases linearly with the amount of training data. [24], [26], [29].

$$d_{euclidean} = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}, \quad (1.2)$$

where x_i and y_i are the two observations between of which the distance is calculated [34].

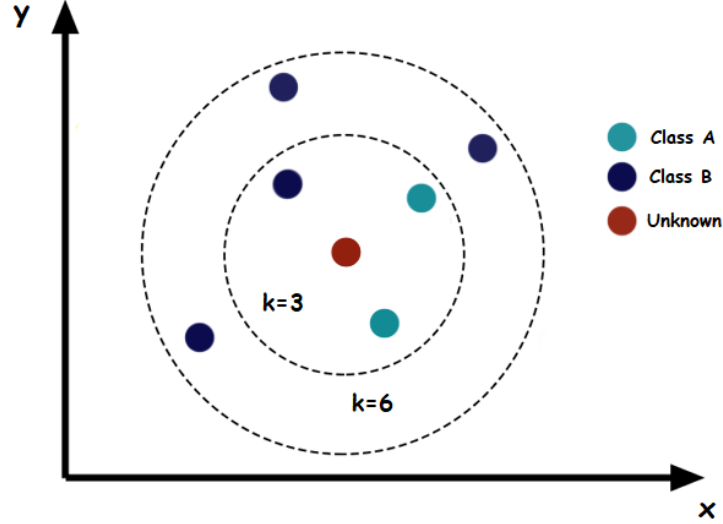


Fig. 1.8: Illustration of the KNN classifier by Author.

▪ Naïve Bayes

Bayesian classifiers are a set of algorithms based on Bayes' theorem, i.e. *a practice for calculating the conditional probability based on prior knowledge and the naive assumption that each feature is independent of the other*. The formula for Bayes' theorem is presented below:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}, \quad (1.3)$$

where $P(A|B)$ corresponds to the probability of an event A occurring when event B has occurred, given new information $P(B|A)$, and a prior belief in the probability of the event $P(A)$. $P(B)$ corresponds to the probability of event B [34].

Specifically, Bayes' theorem establishes a relationship between a class variable y and a dependent feature vector x_1 through x_j [35]. In the context of Machine Learning, Naïve Bayes classifier is based on:

$$P(y|x_1 \dots x_j) = \frac{P(x_1 \dots x_j|y)P(y)}{P(x_1 \dots x_j)}, \quad (1.4)$$

where $P(y|x_1 \dots x_j)$ is called the *posterior* probability of class y (observation), given its values for the j features, $x_1 \dots x_j$, $P(x_1 \dots x_j|y)$ is the *likelihood* of an observation's values for features $x_1 \dots x_j$, given their class y , $P(y)$ is the *prior* belief for the probability of class y and lastly, $P(x_1 \dots x_j)$ is the *marginal probability* [34].

The user selects the statistical distribution of the probability for each feature in the data. In most cases, the Gaussian Naïve Bayes classifier, which follows the normal distribution, is preferred.

$$P(x_i|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right), \quad (1.5)$$

where σ_y^2 and μ_y are the variance and mean values of feature x_i for class y [34], [35].

It is worth noting, that Bayesian classification has the ability to perform quite well even with limited amounts of training data.

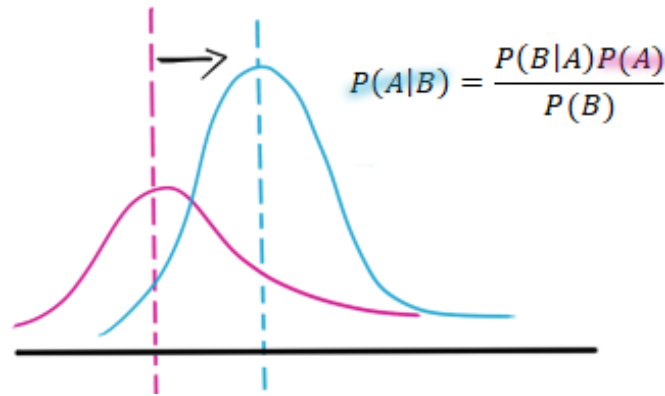


Fig. 1.9: Illustration of the Naïve Bayes classifier by Author (inspired by Destin Gong [33]).

▪ Logistic Regression

Logistic Regression is a binary, predictive classification model with a high level of performance on linearly separable classes. More specifically, a linear model (e.g. $\alpha_0 + \alpha_1 x$) is included within the logistic function $\frac{1}{1+e^{-z}}$, also known as the sigmoid function.

$$P(y_i = 1|X) = \frac{1}{1 + e^{-(a_0 + a_1 x)}}, \quad (1.6)$$

where $P(y_i = 1|X)$ is the probability of the i -th observation, y_i is class 1, X is the training data and a_0, a_1 are parameters to be learned.

Thereby, the output is constrained between 0 and 1. If the probability $P(y_i = 1|X)$ is greater than 0.5, class 1 is predicted; otherwise, class 0 is predicted. In

essence, the new inputs are assigned to one of the two data classes after comparing the probability to a predefined threshold [34].

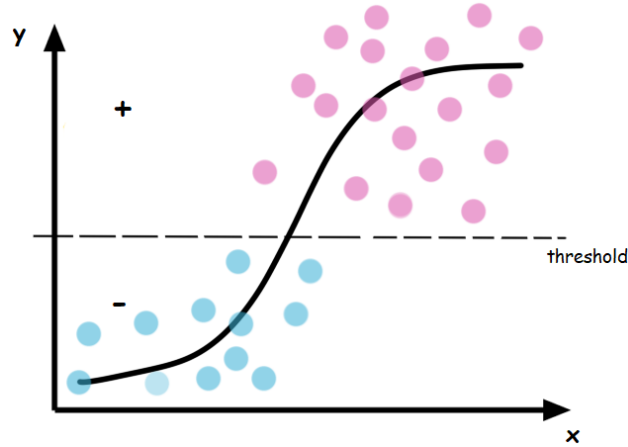


Fig. 1.10: Illustration of the Logistic Regression classifier by Author (inspired by Destin Gong [33]).

▪ Decision Trees

A Decision Tree utilizes a tree structure to represent several possible decision paths and associated outcomes for each one. The decision rules are linked in a chain-like manner. The initial rule is positioned at the top of the tree, with subsequent rules branching off below it. Each node in the Decision Tree represents a query and each branch represents a possible outcome, leading to new nodes. A branch without a decision rule at the end is called a leaf and represents a class label [24]. When a leaf node is accessed by a data sample, the label of the corresponding node will be assigned to the sample. To elaborate, each new instance is sorted by checking the feature specified by a particular node, beginning from the root of the tree and continuing to the branch corresponding to the feature value. The commonly used criteria for classification are the Gini coefficient and Entropy [29].

$$Gini(E) = 1 - \sum_{i=1}^c p_i^2 \quad (1.7)$$

where $Gini(E)$ is the Gini coefficient, c is the total number of events and p_i is the probability of an event occurring for each i event.

$$H(x) = - \sum_{i=1}^n p(x_i) \log_2 p(x_i) \quad (1.8)$$

where $H(x)$ is the entropy of the random variable x , n is the total number of events and $p(x_i)$ is the probability of event x_i occurring.

One of the most well-known algorithms that fall into the category of decision trees is the *Classification and Regression Tree (CART)*, a binary tree in which each root represents a single input and a split point on that variable. The leaf nodes contain an output, which is used to make predictions. In general, decision trees are mainly used in simpler, straightforward problems, and on small data sets, representing both linear and non-linear relationships efficiently. However, these classifiers exhibit problems of overfitting [24].

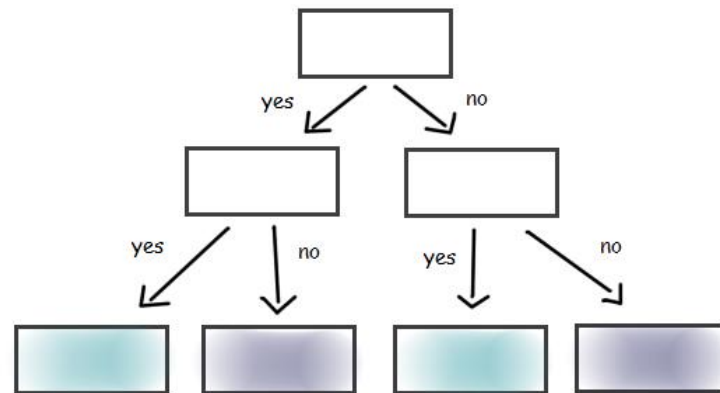


Fig. 1.11: Illustration of the Decision Tree classifier by Author (inspired by Destin Gong [33]).

- **Random Forest**

The Random Forest classifier is an extension of the previous algorithm. It is a technique that aims to improve accuracy by combining multiple models [27]. This method uses the “parallel ensembling” technique, which fits multiple decision tree classifiers, as shown in Figure 1.12, on different subsets of the data. The majority vote is utilized for classification. This reduces the issue of overfitting while remaining unaffected by noise [24], [29].

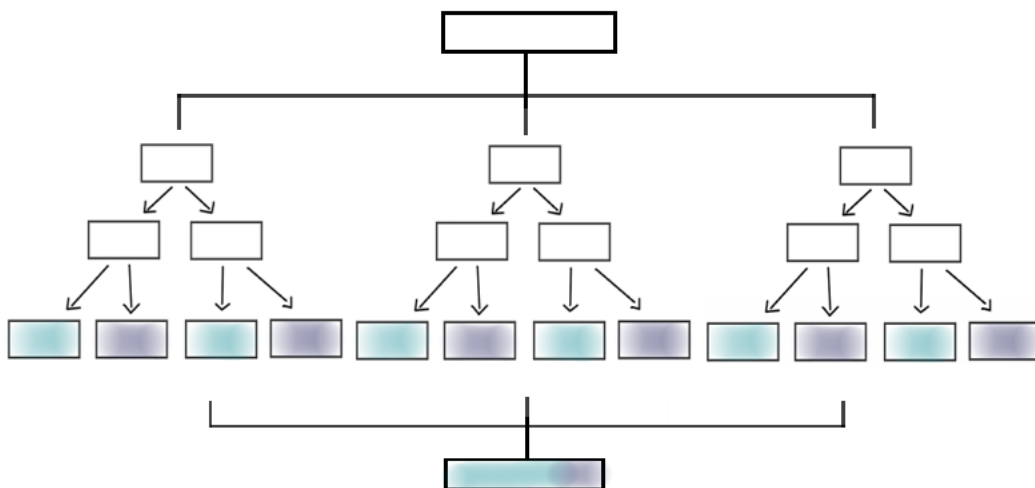


Fig. 1.12: Illustration of the Random Forest classifier by Author (inspired by Destin Gong [33]).

- **Support Vector Machine (SVM)**

Classification using the SVM algorithm is achieved based on the position of the data in relation to a boundary line between two classes. Initially, the data are plotted as points in an n-dimensional space, where feature values correspond to coordinates. The border represents the hyperplane, which maximizes the distance between points from different classes. SVM is known as a non-probabilistic, binary algorithm since it separates the data into two classes mainly used on small datasets. Otherwise, model training becomes complex and time-consuming. The model uses a subset of the training data to make the classification more efficient [24].

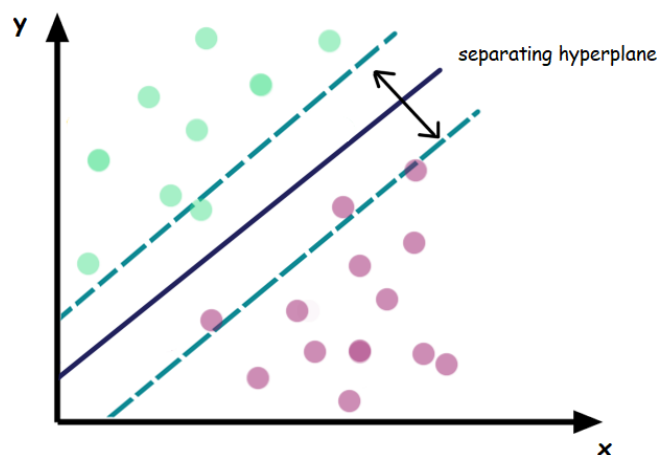


Fig. 1.13: Illustration of the SVM classifier by Author (inspired by Destin Gong [33]).

- **Artificial Neural Networks (ANN)**

An Artificial Neural Network (ANN) is a prediction model inspired by the functioning of the human brain, commonly used as a classifier. The simplest neural network is called *Perceptron*, consisting of a single neuron with n binary inputs and one output node. It computes a weighted sum of the inputs. Perceptron is triggered if the sum is greater than or equal to 0. The equation used for separation in Perceptron is:

$$d(x) = (\sum_{i=1}^n w_i x_i) + w_{n+1}, \quad (1.9)$$

where w represents the weights and $w_{n+1} = 1$.

The unknown pattern x is classified into one of two categories depending on whether the value of the function is closest to 1 or -1. In Figure 1.14, if $d(x) > 0$, the pattern is classified as category 1, and if $d(x) < 0$, it is classified as category 2. When $d(x) = 0$, the pattern isn't classified into any category [36].

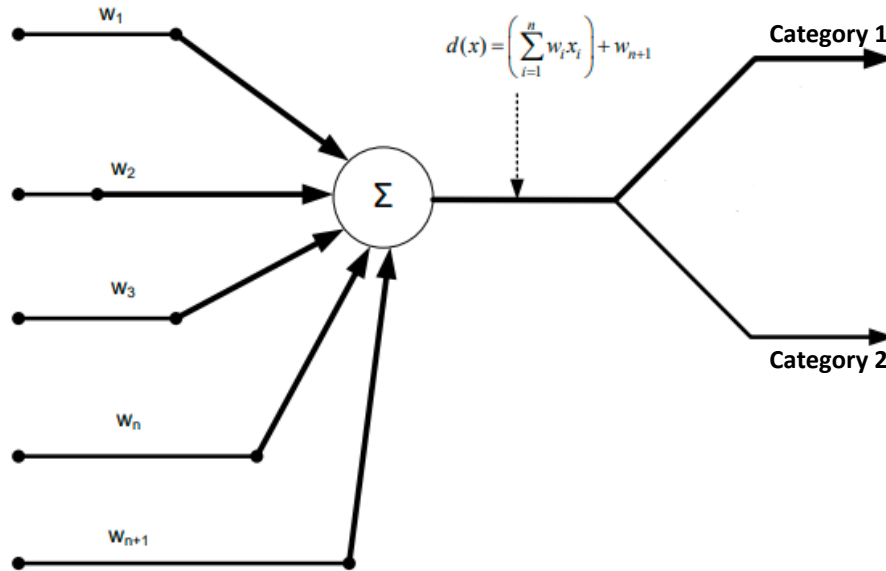


Fig. 1.14: Schematic representation of Perceptron (adapted from [36]).

This simple model is mainly limited to linearly separable matters. It is well known that the structure of the brain is inconceivably complex. A prediction model that approximates how it functions is the *feedforward artificial neural network - multilayer perceptron (MP)*, which has distinct layers of neurons, connected in a specific sequence. This structure comprises an input layer, one or more hidden layers, and the output layer. The input layer is responsible for forwarding the data input. Each hidden layer consists of neurons that receive outputs from the preceding layer, process them making calculations, and transmit the results to the subsequent layer. The output layer produces the final outputs. Each neuron, except for the input layer, has a weight corresponding to each of its inputs and a bias. Usually, the bias equals 1 and is added to the weights vector. If the goal is the binary classification of data, a sigmoid function output layer could be used, scaling the output between 0 and 1, reflecting the probability of prediction [25].

1.3.2 Data pre-processing

Data preprocessing refers to the process performed on raw data to make it suitable for input to other processing procedures, such as those of Machine Learning. Once obtained, there is a possibility that the data may contain some errors. These errors include unnecessary information and noise. Therefore, data pre-processing influences the performance of a model. In general, there are no specific rules for the selection of a pre-processing method, but there is a dependence on the type of data. [37].

In Machine Learning, normalization, which involves adjusting the scaling of data, is a common preprocessing task. This is necessary because the classifiers mentioned earlier assume that all features are scaled the same. In most cases, scales are selected either between 0 and 1 or between -1 and 1. There are a variety of techniques for data normalization.

In this thesis, some of the most frequently used methods found in scientific literature are presented below:

- **Min-Max Scaling**

Min-max Scaling is the most basic normalization methodology. The minimum and maximum values of a feature are used to change the data scale within a certain range. Typically, it is preferred in Neural Networks.

$$x'_i = \frac{x_i - \min(x)}{\max(x) - \min(x)}, \quad (1.10)$$

where x is the feature vector, x_i is an individual element of the feature x , and x'_i is the adjusted element.

- **Standardization**

Standardization is an alternative to the previous method so that the characteristics follow a standard normal distribution. The data are adjusted to have a mean equal to 0 and a standard deviation equal to 1.

$$x'_i = \frac{x_i - \bar{x}}{\sigma}, \quad (1.11)$$

where x_i is an individual element of the feature x , \bar{x} is the mean and σ is the standard deviation.

- **Euclidian Norm**

In normalization techniques, such as those described above, data scaling is applied to the features. However, scale change can also be applied along individual observations. In the Euclidean Norm, the values of individual observations are adjusted so that the sum of their lengths equals 1. This approach is employed in scenarios where multiple equivalent features exist.

$$\|x\|_2 = \sqrt{x_1^2 + x_2^2 + \dots + x_n^2}, \quad (1.12)$$

where x corresponds to an individual observation and x_n is the value of the observation for the n -th feature [34].

1.3.3 Feature selection and dimensionality reduction

In most cases, when handling high-dimensional datasets, it is necessary to prioritize the most significant features. Performing feature selection and reduction procedures before training a Machine Learning model is crucial. This helps eliminate misleading or redundant features that carry similar semantic interpretations. These approaches aim to reduce the dimensions of the dataset, retaining only high-quality features that provide useful information. As a result, this methodology prevents

overfitting and reduces computational complexity, shortening the training time of a model, and enhancing its performance.

Feature selection methodologies are distinguished into three categories: *filters*, *wrappers*, and *embedded/hybrid* methods. Filter methodologies select the most important features for model development according to their statistical properties. In contrast, wrapper methodologies use trial and error to find the subset of features that yields the highest predictive ability. Finally, embedded methodologies select a subset of features with high significance as an extension of the training process to learning algorithms. However, it should be noted, that filter methods are characterized by low computational cost but with inefficient reliability in classification as compared to wrapper methods, which are better suitable for high dimensional data sets. Embedded/hybrid methods have been recently developed and utilize the advantages of both filter and wrapper approaches. A hybrid approach uses an independent test and a performance evaluation function of the features subset [34], [38].

A few indicative examples of such techniques are described below:

- **Variance Thresholding**

Variance Thresholding is one of the simplest feature reduction techniques. According to this practice, low-variance features are deemed less useful to be included in the development of an efficient model, compared to those of high-variance. The selection of the variance threshold is subjective and determined by the user.

$$\operatorname{operatorname{Var}}(x) = \frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2, \quad (1.13)$$

where $\operatorname{operatorname{Var}}(x)$ equals to the variance of a feature, x is the vector of a feature, x_i is an individual value of the feature and μ is its mean value.

Thus, features whose variance value does not meet the threshold value are removed from the subsequent calculations [34].

- **Significance Test Ranking**

The application of a Test of Significance between classes is widely used to assess the ability to identify patterns of individual features. The criterion for sorting the features is defined as the presence of a statistically significant difference between the classes of the feature in question [39]. Any Test of Significance starts with a null hypothesis equal to 0, representing the theory that is proposed as an assertion. In this process, the null hypothesis posits the absence of a statistically significant difference between classes. Conversely, the alternative hypothesis is proposed to verify the test and is deemed true when the null hypothesis is rejected. Depending on the data distribution, whether normal or not, an appropriate test is selected (e.g. t-test, chi-squared tests, ANOVA, Mann Whitney Wilcoxon U-test, etc.). The probability, denoted by the variable p (p -value), is then calculated and compared with a predefined threshold. A small p , typically below the threshold, indicates opposition to the null hypothesis, suggesting that the feature is relevant and should be

retained [40]. Tests of Significance assess each feature separately, disregarding possible relationships of dependence.

- **Correlation Ranking**

This approach is commonly utilized in cases where there is a suspicion of a high correlation among available features in a high-dimensional data set. Generally, a correlation matrix is created to summarize and quantify linear relationships between variables. If two features are highly correlated, the information they provide is considered identical and one of them is discarded [34]. This matrix displays the correlation coefficient (e.g. Pearson's r), depicting the linear pairwise dependence of characteristics. The coefficient ranges from -1 to 1, where a value of 1 indicates a perfect agreement between a pair of features, a value of 0 denotes that there is no dependence, and a value of -1 signifies that there is complete disagreement, with one 'ranking' being the inverse of the other [26].

- **Mixed Criterion**

The Mixed Criterion technique is a combination of the Test of Significance and Correlation Ranking methods described previously. This process is performed to ensure that features selected in correlation analysis exhibit a relatively high degree of statistically significant difference between the available classes. Typically, combining these methods further reduces the number of features contributing to the final predictive model, thereby ensuring that they contain significant information.

- **Principal Component Analysis (PCA)**

When examining a set of available data and addressing specific problems to be resolved each time, it may be desirable to maintain the variance, while reducing the number of features [34]. Principal Component Analysis (PCA) is an unsupervised dimensionality reduction technique of the original data matrix, whereby patterns in the data are detected and identified according to the correlation between features [26]. More specifically, this procedure aims to identify the principal components (*new variables*) in a high-dimensional space, retaining the highest variance. Subsequently, the data are projected into a new subspace with dimensions equal to or smaller than the original. PCA is a process assuming linear relationships between variables and is highly influenced by the scale of the data. Even when the input features exhibit a degree of correlation, the principal components are uncorrelated with each other [26], [38].

- **Recursive Feature Elimination (RFE) Wrapper**

Recursive Feature Elimination (RFE) is one of the most widely utilized feature selection techniques. It is an automated method used to determine the importance of available features [34]. The basic idea is to train a model iteratively, using classification algorithms such as the SVM classifier, involving parameters or otherwise, weights. The main requirement is the normalization of the data. Initially, the model is trained using all available features. During RFE, according to the model fit, weights are assigned to the features. The features with the lowest parameter values are then removed, and the process is repeated for the remaining ones [41]. This

method is combined with cross-validation evaluation to determine the optimal number of features required for designing a robust and representative model [34]. Cross-validation is described in more detail in the next subsection.

1.3.4 Evaluation methods

Evaluation methodologies in machine learning are crucial for assessing the performance, reliability, and generalizability of predictive models. These methodologies encompass a wide range of techniques for estimating classifiers' predictive capabilities and reviewing their suitability for specific applications. The term "*accuracy*", which represents the percentage of correctly classified data, is extensively employed. It's essential that a model not only fits well with training data but also performs sufficiently on predictions involving new, unknown data. Some commonly used evaluation methodologies for these purposes are presented below:

- **Self-Consistency**

In Self-Consistency, the classifier is trained once and evaluated using all available data. Each data point is classified into one of the available classes, providing a preliminary assessment of the data, primarily focusing on the separability of classes rather than the accuracy parameter.

- **Hold-Out**

The Hold-Out approach is one of the most fundamental techniques for evaluating a classification algorithm. It entails splitting the data set into two subsets: a subset of data for training the model and a subset of data for testing the model's performance. The proportion of this random division depends on the size of the original data set. Usually, this ratio is selected equal to 70-30 or 80-20. Model training utilizes only the training set, while performance evaluation is calculated using the testing set. Despite its directness, this method is highly dependent and sensitive to the random way in which the data are divided into the two subsets.

- **Cross-Validation**

Cross-Validation involves dividing the available data into several subsets or folds. The model is trained on a portion of the total subsets and tested on the remaining ones. This process is repeated multiple times using different subsets for training and validation of the classifier. The results of each iteration are averaged to estimate the overall performance of the model. Some special cases can be distinguished:

- i. **Leave-One-Out.** As previously explained for Cross-Validation techniques, this approach is iterative. In the Leave-One-Out method, the process is repeated as many times as there are data points in the dataset [42]. In each performed iteration, the classifier is trained using all but one of the available data points [31]. Subsequently, the model's performance is estimated using the data point excluded from the training process. It is a method with a low level of bias, since each validation is performed on a single data point. At the same time, there is no randomness in the way the data are separated, so the

evaluation is characterized as stable. Nevertheless, it is a time-consuming and computationally expensive process, resulting in its primarily application on smaller, balanced datasets.

- ii. ***K-fold Cross-Validation.*** In K-fold Cross-Validation, the dataset is split into k equal folds or subsets. In each iteration, $k - 1$ folds are used to train the model, while the remaining fold is used for model evaluation. This process is repeated k times, with each fold serving as the validation set once.
- iii. ***External Cross-Validation.*** Through this approach, the data are randomly divided into three parts (for example, 40%, 30%, and 30%) and follow two types of validation: internal and external. In the internal, the system is designed with 40% of the data and evaluated from the first 30% of the data. This process is performed repeatedly for different combinations of features until the combination that returns the highest classification accuracy is obtained. The best performing system is then used in the so-called external validation to classify the remaining data, in this case, the remaining 30%. The total procedure is repeated at least 10 times. The generalization performance of the classifier is estimated by averaging the performance in each iteration. It is a methodology that requires a vast amount of available data to separate them into corresponding groups.

- **Bootstrap**

The Bootstrap method is primarily preferred for smaller datasets. It is similar to hold-out, except that the data sample for model training can be selected randomly, but in such a way that the same patterns appear more than once (re-substitution) [42].

The performance estimation of a model is primarily calculated using the *Truth Table*, also known as the confusion matrix. Table 1.6 provides an example of such a table for a binary classification problem with two categories, K1 and K2:

Table 1.6 *Truth Table of a binary classification problem.*

		TRUTH TABLE		
		Computational classification		
Experimental classification			K₁	K₂
	K₁	TP	FN	
	K₂	FP	TN	

TP is the number of correctly classified cases (true positive values) and TN is the number of incorrectly classified cases (true negative values). Similarly, FP and FN are false predictions.

The *accuracy* of the classification, as a measure of quality, is calculated as:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \times 100 \quad (1.14)$$

Significant evaluation parameters include *sensitivity* (or *recall*), which measures the model's ability to correctly identify true positive cases (e.g., abnormal cases), and *specificity*, which measures the model's ability to correctly identify true negative cases (e.g., normal cases).

$$sensitivity = \frac{TP}{TP + FN} \quad (1.15)$$

$$specificity = \frac{TN}{TN + FP} \quad (1.16)$$

Furthermore, *precision* is a performance metric which denotes the proportion of predicted positive cases that are in fact positive [43].

$$precision = \frac{TP}{TP + FP} \quad (1.17)$$

Finally, the *F1-score*, also known as the F-measure, is a score that represents the harmonic mean of the recall and precision. The relative contribution of recall and precision to the F1-score are considered equal. Like precision, the F1-score ranges from 0 to 1, where 1 indicates perfect performance [44].

$$F1 - score = \frac{2 \times TP}{2 \times TP + FP + FN} \quad (1.18)$$

1.4 Essential Tools for Statistical Analysis

Machine learning is widely adopted as a research tool in healthcare-related disciplines. Admittedly, it relies heavily on statistical concepts. *Statistical Analysis* provides the theoretical foundation and a variety of tools upon which a large part of Machine Learning techniques is based, rendering these two disciplines interrelated. Statistical concepts are pivotal for interpreting, visualizing, and validating, and by extension, supporting the results produced by Machine Learning models. In summary, the merging of disciplines makes it a powerful tool that contributes to decision-making.

This subsection describes tools within the scope of Statistical Analysis that are utilized in the context of this thesis.

- **Tests of Significance**

Significance testing, aside from serving as a feature reduction technique, is an example of a common methodology used for evaluating model results. Its operational

principles were discussed in subsection 1.3.3 and it represents one of the simplest methods for comparing and assessing relationships between variables. Typically, the data encountered in biomedical applications are hardly normally distributed or their distribution is unknown. Hence, the *Mann-Whitney Wilcoxon U-test* is employed as a Test of Significance. This non-parametric statistical test requires no assumptions about the data distribution and applies to two unpaired classes of data. The data examined each time should correspond to two random, independent samples. These samples are combined, and calculations are performed, assuming their origin to be negligible. The statistical test yields the U parameter, which determines the probability p using tables or specialized software. As noted, this value is compared to a predefined threshold. Depending on whether it is lower or higher than this threshold, the null hypothesis is either rejected or retained, respectively.

- **Receiver Operating Characteristic (ROC) curves**

Receiver Operating Characteristic (ROC) curves are a schematic representation used to further evaluate the performance of binary classification. This method owes its origins to radar systems, where it was developed as a technique to separate signal and noise [45]. In essence, it's a graph used to visualize the model's ability to separate data into two classes for specific attributes. A ROC curve plots the coordinates of the points using "sensitivity" as the y-axis and "1-specificity", which is the percentage of false positives, as the x-axis, as shown in Figure 1.15.

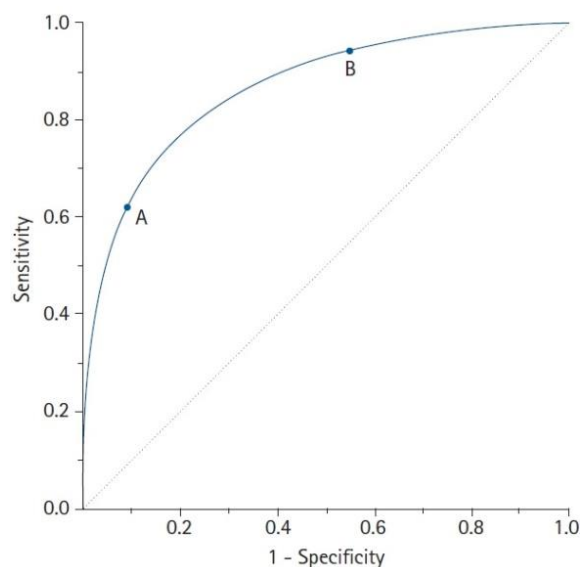


Fig. 1.15: ROC curve [46].

The *Area Under the Curve (AUC)* is a measure of *accuracy* defined as the area enclosed by the graph of the ROC curve. An ideal case of a ROC curve corresponds to an AUC equal to 1. The more the curve shifts up and to the left, the more the accuracy of a test increases, as the sensitivity tends towards 1 and the percentage of false positive results tends towards 0. In general, for a model's performance to be deemed acceptable, the AUC value should typically exceed 0.8 [46].

- **Box plots**

Another approach of graphical representation that is broadly utilized in interpreting and evaluating the results of a model is box plots. It is a simple descriptive statistics method defined for the schematic representation of numerical data from a series of observations. Typically, in biomedical applications, it is a useful approach to compare the distribution of data between classes on a particular feature. A box plot can be represented either vertically or horizontally. It is considered essential to present its constituent parts for its proper interpretation.

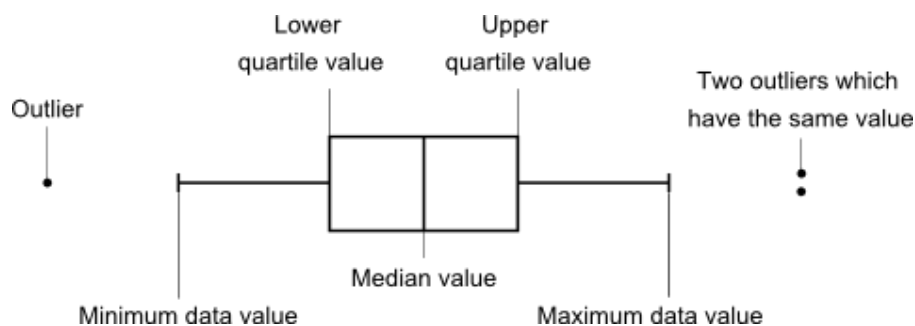


Fig. 1.16: Horizontal box plot [47].

As evident from Figure 1.16, the box is structured by the lower quartile (Q1, 25th percentile) and the upper quartile (Q3, 75th percentile), which constitute a percentage equal to 25% of the data respectively. The vertical line, which divides the box into two parts, represents the median and indicates that 50% of the data appears to the right and 50% to the left. The whiskers extend on either side of the box to a length equal to 1.5 times the interquartile range. Their edges correspond to a representative minimum and maximum value, while the points outside the whiskers indicate extreme data values [48].

1.5 Basic Concepts of Computational Chemistry

Undoubtedly, *Computational Chemistry* lies at the heart of the design and discovery of new drugs. The advancement of mathematical algorithms has facilitated their broader application in the field of chemistry, particularly in the development of *in silico* methodologies. This term refers to the simulation of molecular systems and the performance of necessary calculations using a computer, aiming for property prediction and further study of physicochemical changes of chemical compounds. *Molecular Modeling* derives from Theoretical Chemistry and is the root of *in silico* studies. It guides *in vitro* experimental studies and, consequently, advances research in the field. Molecular Modeling encompasses various capabilities, including the 3D representation of molecules, the calculation of molecular properties, the investigation of molecule binding to receptors, the development of Quantitative Structure-Activity Relationship models (QSAR), and the study of metabolic properties (Absorption-Distribution-Metabolism-Excretion-Toxicity, ADMET) [49]. Therefore, Computational Chemistry in Pharmaceuticals aims to improve the experimental process as much as possible to produce new and safe pharmaceutical products.

1.5.1 Chemical databases

As already mentioned, a database is an important tool that contains a plethora of stored information.

One of the most common models that database systems follow is the *relational* model, a simple approach to representing data in the form of tables or else matrixes. Generally, column headers represent the titles of different fields while rows represent the values or the records of a matrix structure [50]. In the context of Machine Learning, each column is considered an attribute, and each row is an individual observation.

In the pharmaceutical search, access to information related to small molecules, their activity, and other properties is critical. However, there is no specific requirement from institutions, such as scientific journals, for researchers to submit the results of small molecule tests to databases. Information published in articles is presented in an unstructured form, such as images, which makes it difficult to retrieve and process. At the same time, new research on disease mechanisms may alter existing knowledge. These factors have led to the development of open-access databases, which collect bibliographic data by conducting systematic reviews. This type of data fuels algorithms and *in silico* experiments for investigating biological systems [51].

Depending on the specific problem at hand, relevant data is sought in specialized libraries. Nowadays, there are many public databases dedicated to medicinal chemistry applications. For example, one of the commonly employed databases for retrieving chemical compounds used in pharmaceutical research is the *ZINC* library. In addition, *PubChem* provides information on the physicochemical properties of millions of elements, along with access to relevant scientific articles [52]. *ChEMBL* is a public database focusing on bioactive molecules covering a range of properties, targets, and organisms. Other sources focus on more specific properties, such as the binding energy of micromolecules and thus, the binding affinity of the ligand to the macromolecule. Databases, such as *PDBbind*, collect this kind of information on protein-ligand complexes from the *Protein Data Bank (PDB)*, an international repository of macromolecular structures [53].

1.5.2 Molecular representation

It is widely acknowledged that molecules are practically real entities and for their various applications, it's essential to extract information about their structure or chemical properties. Thus, symbolic representations allow the derivation of the necessary information about the molecule in question. The amount and type of information are directly linked to the complexity of the molecular representation, selected according to the problem at hand. Categories of molecular representation are commonly distinguished based on spatial dimension.

- **0-Dimensional (0D)**

The chemical formula is the simplest form of molecular representation and captures the number and presence of chemical elements. The 0D representation is independent of information related to molecular structure and connectivity between atoms.

- **1-Dimensional (1D)**

Molecules are represented by a count of their structures, such as molecular fragments, functional groups, or substituents. It does not require complete knowledge of the molecular structure.

- **2-Dimensional (2D)**

This representation takes into account the connectivity of atoms in terms of the presence and nature of chemical bonds. Usually, the molecule is perceived as a graph, whose edges are bonds and the vertices are atoms.

- **3-Dimensional (3D)**

In 3D representation, the molecule is viewed as a geometric object in space, taking into account the nature of the atoms, their connectivity, and spatial configuration. Cartesian coordinates (x - y - z) are utilized for this purpose.

- **4-Dimensional (4D)**

The fourth dimension is used to characterize and quantify the interactions of a molecule with the active site of a receptor [54], [55].

Figure 1.17 illustrates the molecular representation of ibuprofen based on spatial dimensionality.

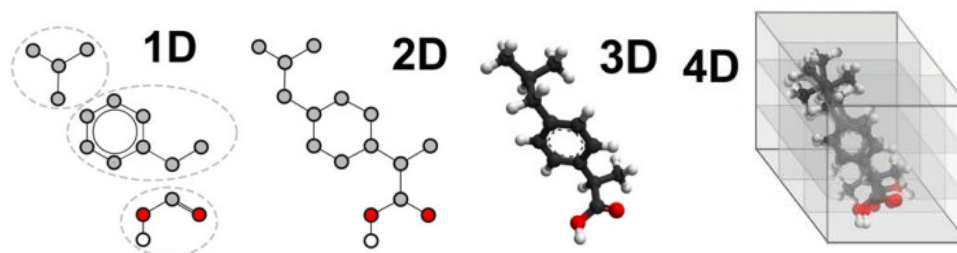


Fig. 1.17: Graphical representation of ibuprofen (1D through 4D, adapted from [54], [55]).

In Computational Chemistry, molecules need to be in a machine-readable format to undergo computational processes. As computers and their processing power have evolved, flexible methods of molecular representation have been developed. The most prevalent category is string representations. They consist of characters from the *American Standard Code for Information Interchange (ASCII)* character encoding standard, offering the advantage of being a method that is easily comprehensible by humans [56]. The *Simplified Molecular Input Line Entry System (SMILES)* format is the most popular line notation, specifically designed for computer use by chemists. It was created by David Weininger in 1986 at the US Environmental Research Laboratory (USEPA). It was further developed at Daylight Chemical Information Systems, encoding stereochemistry in an intuitive way [57]. It is a specification for describing the structure of organic chemical molecules using short ASCII strings. The rules developed for representation through SMILES apply to almost any chemical structure. To begin with, neighboring atoms are positioned adjacent to each other,

with the symbols of the atoms identical to those on the periodic table. However, it's important to note that hydrogen atoms are not explicitly represented. Branches in the molecular structure are denoted by enclosures in parentheses. In linear representations of cyclic structures, a bond is severed at each ring, and the atoms of the connecting ring are subsequently followed by the same digit in the text representation. Additionally, atoms within an aromatic ring are denoted by lowercase letters [58]. Daylight introduced a commercial product for generating canonical SMILES, but since their algorithm was proprietary, other commercial and open-source software developers created their own algorithms for generating canonical SMILES. For instance, in 2005, the *IUPAC International Chemical Identifier (InChI)* was released for the first time providing a canonical representation linking information from various databases on the same chemical compounds, since there was a need for a community standard for a canonical linear representation. In short, to achieve this, the InChI algorithm uses a normalization procedure, a canonicalization algorithm, and a layered structure to help identify isomers. InChI resolves many of the chemical ambiguities that are not addressed by SMILES, particularly concerning the stereogenic centers, tautomers, and valence issues. However, InChI is difficult to interpret by humans in most cases [57].

Table 1.7 lists the symbolism of chemical bonds as represented in SMILES [58].

A few representative examples of SMILES are illustrated in Figure 1.18.

Table 1.7 Canonical SMILES symbolism of chemical bonds.

Bonds	SMILES Symbolism
Single	-
Double	=
Triple	#
Aromatic	:




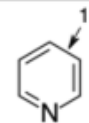
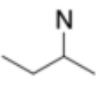
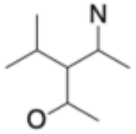
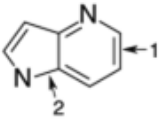
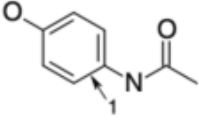
			
CCC	CC=C	CC#C	c1ccncc1
			
CCC(C)N	CC(C)C(C(C)N)C(C)O	c1cc2c(cc[nH]2)nc1	CC(=O)Nc1ccc(cc1)O

Fig. 1.18: Representative examples of SMILES [55].

1.5.3 Molecular descriptors

As time has progressed and technology has evolved in the scientific field of *in silico* and computer-aided drug discovery research, there's been a growing necessity to encapsulate the complexity of molecules and their molecular attributes in a manner that facilitates the implementation of mathematical calculations for their analysis [59]. Thus, molecular descriptors were defined, which are measurable concepts directly related to the structure and representation of molecules. More specifically, a molecular descriptor can be determined as the final result of a logical and mathematical process that converts the chemical information of a molecule into useful numbers or the result of a standard experiment encoding different aspects of its structure and activity. Applications, such as predictions of protein-ligand interactions, ligand-based virtual screening, and molecular similarity studies, are based on sets of molecular descriptors [54], [55].

As already mentioned, the manner in which molecules are represented is linked to molecular descriptors and the type of information that can be derived. This information typically encompasses physical, chemical, and topological characteristics. Molecular descriptors are computed using a variety of computational tools and libraries, including RDKit, CDK, and PyBel, among others.

Depending on the logic underlying them, they are clustered into two main categories: “Classical” molecular descriptors (MDs) and binary fingerprints (FPs).

▪ “Classical” Molecular Descriptors (MDs)

To date, hundreds of MDs have been proposed, each capturing the unique properties of molecules in various ways. They are designed to encode either a single feature or a set of features of varying complexity into a single number [54]. The majority of MDs can be classified, as it is apparent, according to the dimensionality of the molecular representation, and are indicated as integer, binary, or continuous numbers.

MDs can be subjected to scaling, selection, and feature reduction techniques and for this reason; they are exploited as inputs to algorithms for the design of Machine Learning models.

Table 1.8 provides representative examples and abbreviations of various libraries used for molecular descriptors, categorized by their dimensionality [60].

Table 1.8 *Examples of common MDs.*

Type	Common MDs' Symbolism	Definition
1D	• Weight	• Molecular weight
	• Mr	• Molar refractivity
	• logP(o/w)	• Log of the octanol/water partition coefficient
2D	• a_nN	• Number of N atoms in the molecule

Table 1.8 (Continued).

2D	• b_Double	• Number of double non-aromatic bonds
	• vsa_acc	• Approximate sum of VDW surface area of H bond acceptors
3D	• ASA+	• Solvent accessible surface area of all atoms with positive partial charge
	• Pmi	• Principal moment of inertia
	• Vol	• VDW volume

▪ **Binary Fingerprints (FPs)**

Unlike MDs, FPs are a complete, binary representation of the structural parts of a molecule. They are a more complex form of descriptors and depend on the how rows of bits are converted and encoded by the 1D and 2D molecular representation [61]. They represent the presence ("1") or absence ("0") of specific functional groups and are only meaningful when used as a whole. FPs are mainly applied to perform fast calculations involving molecular similarity or molecular diversity. Recently, they have also started to be exploited to study bioactivity patterns by calculating the frequency of molecular fragments [54]. There are many types of FPs, which cover a wide range of different subgroups and use different numbers of bits.

Typical categories of FPs are listed in Table 1.9 [61].

Table 1.9 Paradigms of FPs types.

Type	Common Approaches	Definition
<i>Substructure keys-based FPs</i>	• MACCS	• Two variants: 960 and 166 structural keys (bits), based on SMARTS patterns, covering most of chemical features
	• PubChem FP	• 881 bits, commonly used for similarity searching and neighboring
	• BCI FPs	• The standard substructure includes 1052 bits , but can be modified by the user

Table 1.9 (Continued).

<i>Topological or path-based FPs</i>	• Daylight FP	• Consisting of up to 2048 bits, it encodes all possible connectivity pathways through a given length molecule
	• Molprint2D	• Encodes the atom environments of each atom of the molecular connectivity table, which are represented by strings of various sizes
<i>Circular FPs</i>	• ECFPs	• They represent circular atom neighborhoods and produce FPs of variable length, based on the Morgan algorithm. Mostly used with a diameter of 4 (ECFP4)

Figure 1.19 provides an example of a molecule's representation using FPs.

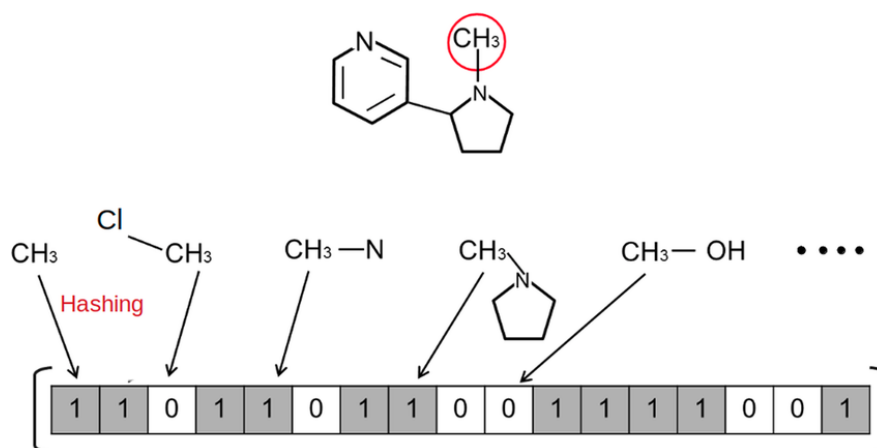


Fig. 1.19: Simplified fingerprint generation: the hashing function sets just 1 bit per pattern [62].

1.5.4 Virtual screening

According to Walters, Stahl, and Murcko in “Virtual Screening – An Overview”, Virtual Screening (VS) involves utilizing high-performance computing techniques to identify potential drug candidates for a specific target by analyzing large databases of chemical compounds [63]. It has become a fundamental aspect of the drug discovery process since its purpose is highly correlated to finding novel chemical structures and therefore, new scaffolds that bind to a macromolecular target of interest. As a result, it acts as a filter for the selection, synthesis, or purchase of compounds and promotes the rational search for potential medicinal products.

First and foremost, it's essential to understand that VS is a significantly wide domain encompassing a variety of available methods for its

purposes. Consequently, VS methodologies and techniques are divided into two general types: Ligand-based VS and Structure-based VS.

▪ **Ligand-based VS methodology**

The basic principle, which underlies ligand-based VS approaches, is the supposition that molecules sharing certain relevant and identical features with confirmed ligands should exhibit similar properties and effects concerning a specific, biological target. These features, or molecular similarities, are presented in the format of molecular descriptors, discussed in detail in section 1.5.3 of this thesis. For ligand-based VS to be conducted, one or more active compounds known to bind to the selected target, as well as an available database of molecules are mandatory. The process begins with calculating molecular descriptors. The next step involves preparing the database to meet the requirements of the most significant descriptors and the selected ligand-based method. There are several techniques distinguished, differing in complexity:

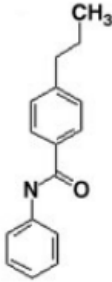
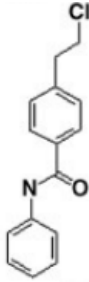
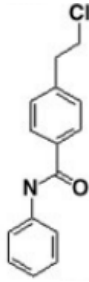
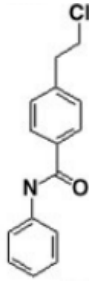
- i. **Similarity Search.** A straightforward approach is the application of a similarity search. When executing this type of search, since there is no specific way to quantify the similarity between molecules, multiple coefficients can be measured and reviewed accordingly. For instance, *Tanimoto* is the most widespread similarity coefficient dependent on fingerprint bits. Distance similarity measures, such as *Euclidean distance*, are commonly employed as well. The difference between these techniques is that the Tanimoto coefficient takes into consideration the presence of common features whereas Euclidean distance considers the dissimilarity of attributes. The practice of similarity searches results in a list of compounds, often referred to as nearest neighbors. These compounds are ranked based on the resemblance of the reference compound and the structures being searched. The list can be constrained to a specific number of nearest neighbors or compounds surpassing a certain similarity threshold. Determining this threshold lacks a general rule, particularly since there isn't a similarity value that definitively discriminates between active and inactive compounds.

Table 1.10 presents an example of a similarity search based on the Tanimoto coefficient, ranging from 0 to 1. In this case, the coefficient has been calculated according to Daylight FPs with a fixed length of 1024 bits. As evident from their chemical structures, the first two pairs of molecules differ only by the substitution of a methyl group with chlorine. The last example indicates a pair of compounds with low similarity [64].

Table 1.10 Similarity search based on the Tanimoto coefficient.

Molecule 1	Molecule 2	Tanimoto Coefficient
$\text{H}_3\text{C}-\text{CH}_3$	$\text{H}_3\text{C}-\text{Cl}$	0.286

Table 1.10 (Continued).

		0.851
		0.055

- ii. **Machine Learning Models.** A more complex approach involves the utilization of chemical compound datasets with known activities to train machine learning models. These models exceed in advantages, in contrast to the previously noted methods, and intend to identify compounds that share structural similarities with known active compounds, distinguishing active from inactive molecules. The activity of the chemical compounds in a dataset is represented through MPs or FPs. Machine learning is used for the development of prediction models and thus, to relate structural information of chemical compounds to biological activity. Since these techniques also involve information related to the inactivity of molecules, considering a specific target, *Structure-Activity Relationship (SAR)* patterns can be extracted [64], [65]. A SAR is a qualitative association between a chemical substructure and its presence in chemical compounds, exhibiting certain biological effects, also known as endpoints. The necessity to reduce cost and time, coupled with the evaluation of vast amounts of molecules resulted in the development of *Quantitative Structure-Activity Relationships (QSARs)*. A QSAR is a mathematical model that quantitatively relates a numerical measure of chemical structure, such as a physicochemical property, to a physical property or an endpoint [66]. Nowadays, QSAR modeling is divided into different types, deriving from the dimensionality of molecular descriptors (1D to n D). In this thesis, the term QSAR specifically refers to *Classical 2D-QSAR* methods, which entail comparing structural features like 2D-pharmacophores with biological activities. It is a theoretical method used in drug discovery and development that aims to identify a statistically significant correlation in two different approaches: between the chemical structure and continuous properties, such as pIC_{50} , K_i , *etc.*, or between the chemical structure and categorical, binary, biological, or toxicological properties. The first approach utilizes regression techniques, whereas the second one classification

techniques. Thus, QSARs are employed in four main areas: to prioritize existing chemicals for further testing or evaluation, to classify and label new chemicals, to assess the risk of new and existing chemicals and to fill possible data gaps. However, it should be noted that the development of predictive models faces major obstacles originating from the data quality, such as errors in chemical structure and experimental results, and thus, data curation procedures are applied. For the sake of completeness, such procedures involve the removal of organometallics, counter ions, mixtures, and inorganics, the normalization of specific chemotypes, structural cleaning, such as detection of valence violations, and standardization of tautomeric forms. Additional curation elements include averaging, aggregating, or removal of duplicates to produce a single bioactivity result [67], [68].

The workflow of QSAR-based VS is presented in Figure 1.20.

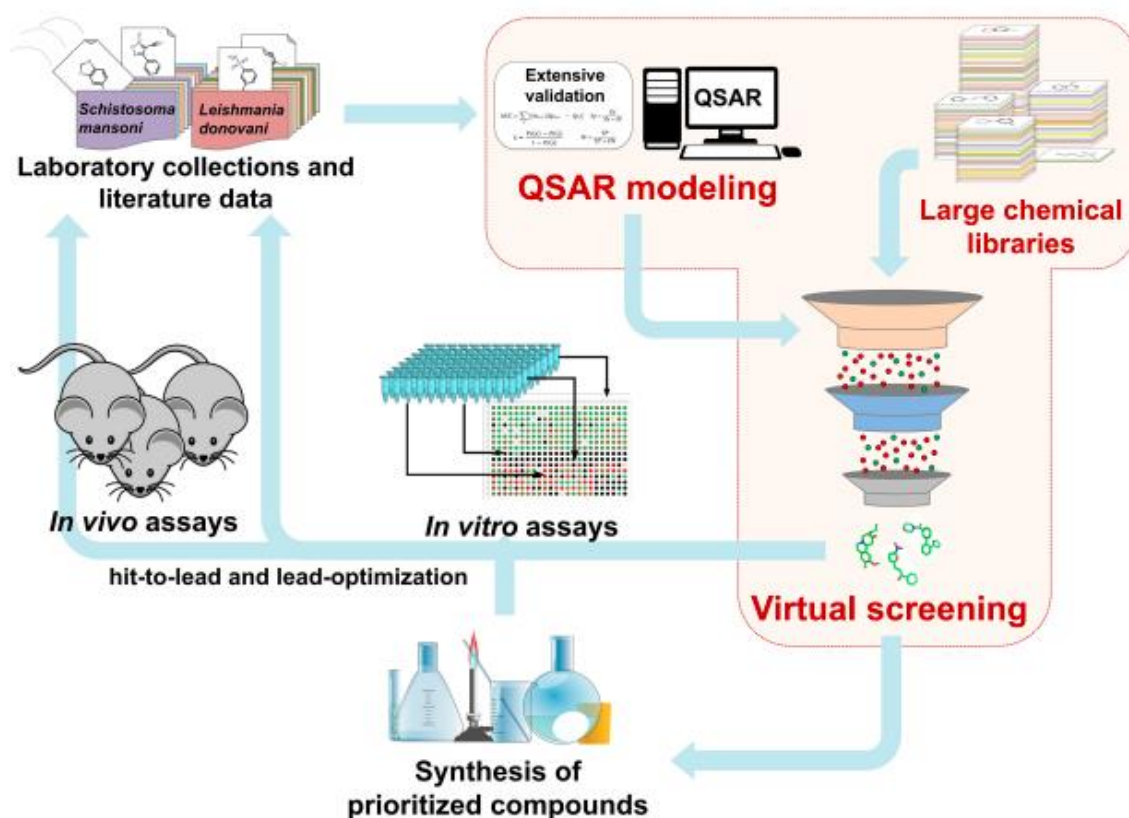


Fig. 1.20: The datasets are collected from external sources. After data curation and the development of QSAR models, compounds from chemical libraries predicted as active are prioritized for *in vitro* assays. Subsequently, *in vivo* assays are conducted [67].

A substantial downside of ligand-based VS methodologies is that they focus solely on the ligand, neglecting information associated with the target.

- **Structure-based VS methodology**

Structure-based VS methods often referred to as receptor-based or target-based methods, apply different techniques in comparison to those of ligand-based VS, focusing on the interactions between a ligand and a certain drug target. The main

requirement for these approaches is the availability of a 3D structure, obtained from X-ray crystallography or other structure elucidation methods, such as Nuclear Magnetic Resonance (NMR) spectroscopy, or even computational methods like homology modeling. There are two fundamental, complimentary structured-based VS techniques, differing in their methodology and purposes:

- i. **Active-site derived pharmacophore methods.** These types of methods take into consideration a pharmacophoric model constructed from a 3D structure of the target, to capture the underlying ligand-receptor pattern of interactions and the general active-site topology [65]. According to 1998 IUPAC, a *pharmacophore* is “the ensemble of steric and electronic features that is necessary to ensure the optimal supramolecular interactions with a specific biological target structure and to trigger, or to block its biological response”. Therefore, it is not a real molecular representation or association of functional groups, but an abstract concept that represents the possible interactions between molecules and the target of interest. 3D pharmacophore modeling is, also, based on the theory that molecules exhibiting similar chemical functionalities and spatial arrangement lead to biological activity on the same target. To demonstrate significant predictive power, it discards information that is not directly related to the binding site and classifies interactions to pharmacophoric properties, such as hydrophobic areas, positively and negatively charged groups, and hydrogen bonds (donors and acceptors). These properties are represented by geometric entities, shown in Figure 1.21. In addition, the parameter of the *Root Mean Square Deviation (RMSD)*, which quantitatively measures the average distance between atoms, is calculated and utilized for structure comparison of the biomolecules, evaluating how well a predicted ligand pose matches a reference pose. *Catalyst*, *Molecular Operating Environment (MOE)*, and *LigandScout* are examples of pharmacophore VS software platforms [69]. The formula of RMSD is presented below [70]:

$$RMSD(a, b) = \sqrt{\frac{1}{n} \sum_{i=1}^n (a_{ix} - b_{ix})^2 + (a_{iy} - b_{iy})^2 + (a_{iz} - b_{iz})^2}, \quad (1.19)$$

where a_i and b_i refer to molecule “1” and molecule “2”, respectively. The subscripts x, y, z correspond to Cartesian coordinates.



Fig. 1.21: Geometric representations of pharmacophoric features 1—hydrogen bond acceptor (HBA), 2—hydrogen bond donor (HBD), 3—negative ionizable (NI), 4—positive ionizable (PI), 5—hydrophobic (H), 6—aromatic (AR), 7—exclusion volume (XVOL) [71].

- ii. **Molecular Docking.** A fundamental structure-based VS type, which is commonly used in research, is Molecular Docking. Molecular Docking intends to predict the structure of the intermolecular complex formed between two or more constituent molecules, usually a small molecule and a protein target. This is the reason why this method is also referred to as protein-ligand docking in related literature. Docking protocols can be described as a combination of two complimentary components, since they include the computational adjustment of candidate ligands to a protein target, using search algorithms, followed by the application of a scoring function to estimate the binding probability of a small molecule to the protein [72]. To put it in other words, the search for the precise ligand conformations and orientations, namely referred to as posing, is implemented by various docking algorithms. In addition, scoring functions predict binding free energies (kJ/mol, kcal/mol), in order to evaluate and rank poses of chemical compounds, calculated by docking algorithms [65].

To comprehend this methodology, the principles underlying Molecular Docking are going to be explained. Since the receptor is a macromolecule, drug compounds bind to a specific region. When working with a protein receptor this region is called *binding pocket*. The force that impels the binding of molecules to the receptor is their stereoelectronic properties. These determine whether attractive, repulsive, steric or electrostatic forces will be exerted. In the context of ligand-protein binding, *Gibbs free energy* (ΔG) is associated with binding affinity, representing the stability of the complex and quantifies the strength of interactions between a drug molecule and a receptor. It is the sum of *electrostatic* (EI), *inductive* (II), *non-polar* (NPI) and *hydrophobic* (HI) interactions between said molecules, reduced by the term *expressing the loss of energy or entropy* (ΔG^*). At this point, it is essential to mention that the presence of hydrogen bonds, as electrostatic forces of attraction, is responsible for stabilizing the protein-ligand complex.

$$\Delta G_{total} = \Delta G_{EI} + \Delta G_{II} + \Delta G_{NPI} + \Delta G_{HI} - \Delta G^* \quad (1.20)$$

where ΔG_{total} corresponds the sum of Gibbs free energies. The subscripts indicate the energies of the referred interactions and ΔG^* the parameter of entropy.

It is, also, defined as the energy of the complex reduced by the energies of the receptor and the ligand.

$$\Delta G_{bind} = \Delta G_{complex} - (\Delta G_{ligand} + \Delta G_{receptor}) \quad (1.21)$$

where ΔG_{bind} corresponds to the binding affinity, $\Delta G_{complex}$ is equal to Gibbs energy of the complex, ΔG_{ligand} is equal to Gibbs energy of the ligand and $\Delta G_{receptor}$ the Gibbs energy of the receptor.

The several intermolecular interactions that occur when drug molecules approach the receptor are significant for the theoretical study and development of new pharmaceutical products. Examples of Molecular Docking software platforms are *Autodock Vina*, *Glide (Schrödinger Suite)* and *SwissDock* [73], [74].

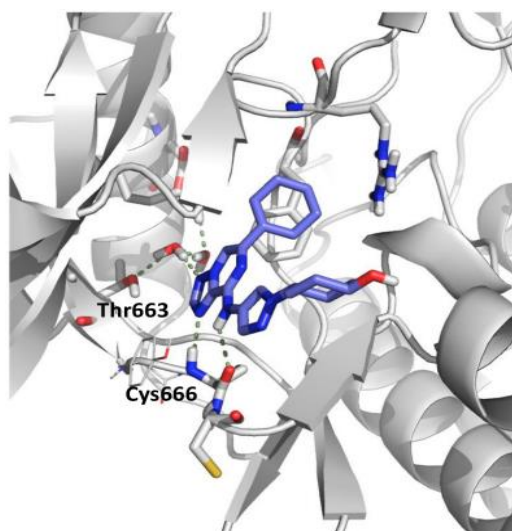


Fig. 1.22: Example of Molecular Docking using CSF1R by Schrödinger's computational platform (*Maestro Glide*) [75].

To conclude with, if both target and ligand structures are known in virtual screening, the whole information can be used in a combined approach.

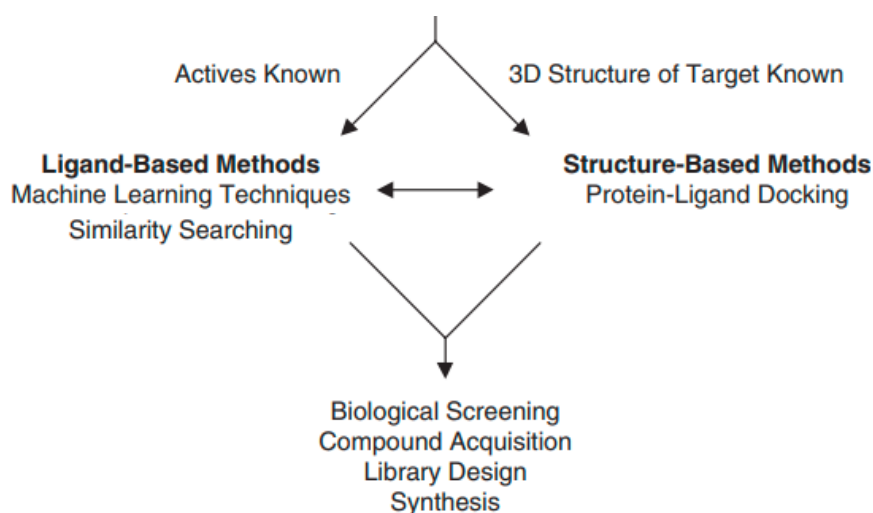


Fig. 1.23: Sequence of methods (adapted from [76]).

1.5.5 ADMET properties

For a thorough *in silico* study of a receptor's potential ligands, Molecular Modeling provides the ability to analyze their metabolic properties. The acronym ADMET stands for the study of Absorption, Distribution, Metabolism, and Excretion of bioactive substances, covering the pharmacokinetics of molecules, as well as various Toxicities. In short, using special computational techniques, the path of the potential drug within the organism is determined in the early stages of the research process. An approximate evaluation is conducted to determine if the drug molecule will bind to the biological target, how long it will stay in the bloodstream and its potential adverse effects on the body [77].

The body's absorption of a medicine varies depending on the method of administration. Generally, the absorption of active substances from orally or rectally administered medicines is called *enteral*. In contrast, absorption following intravenous, intramuscular, or any other route of administration that bypasses the gastrointestinal tract is called *parenteral*. This distinction arises from the *first-pass effect*, where a significant amount of the drug is neutralized by enzymes in the stomach, intestines, and liver cells. Calculating the percentage of an active substance that enters the bloodstream is of utmost importance. For this reason, the parameter of *bioavailability* is defined [78].

$$\text{bioavailability} = \frac{\text{the quantity of medicine entering circulation}}{\text{the quantity of medicine administered}} \quad (1.22)$$

The pharmaceutical substances do not remain in the bloodstream; instead, they move to the biological target and are distributed to tissues and cells through the blood vessels. In most cases, the majority of active substances need to pass through the cell membrane. Factors such as molecular weight determine the ability of a drug to move from the plasma to the extracellular fluid, while further transport within the cell, in the case of an intracellular target, depends on the drug's lipophilic nature. On the contrary, if the target is a protein receptor, the substance binds directly to it [79]. In ADMET prediction, the *blood-brain barrier (BBB)*, a semipermeable and extremely

selective system in the CNS, is specifically addressed [80]. In short, it is a protective membrane that encloses the capillaries of the circulatory system in the brain and protects it from the passive diffusion of unwanted, polar compounds of blood circulation.

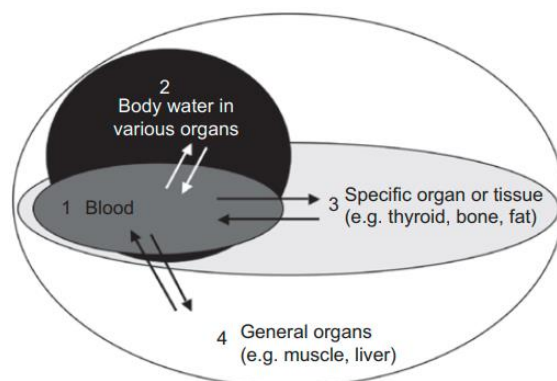


Fig. 1.24: Schematic representation of the various drug-distribution pathways [79].

The majority of small molecules used as drugs are xenobiotics, meaning they are foreign to the physiological biochemistry of the human body. Consequently, upon entering the body, they are targeted by a range of metabolic enzymes as a defense mechanism. The enzymes, in this case, modify the unfamiliar structure to facilitate its elimination from the body. The majority of molecules undergo metabolic reactions, resulting in the formation of metabolites. Depending on the case, they might either lose the activity of the original drug or become more active. Identifying the metabolites of a new drug is necessary before its approval. The reactions that occur during metabolism are classified into Phase I reactions, such as oxidation, reduction, and hydrolysis, and Phase II reactions, where new polar groups are attached to enhance the water solubility of metabolic products. The active groups produced in Phase I serve as substrates for Phase II reactions.

The excretion of pharmaceutical substances is inextricably linked to their structure and various properties. It happens either through the elimination of their original structure or the excretion of their metabolites. Excretion is achieved in different routes, such as urine, sweat, and respiration. In general, the liver and kidneys are considered the main organs of excretion. The pharmacokinetic parameter, which quantifies the ability of the body to eliminate the drug after its entrance into the systemic circulation, is called *clearance*. The final clearance of the drug is equivalent to the combined result of all the individual elimination processes it undergoes. Clearance, from a specific organ, depends on its *blood flow* and the drug's *extraction ratio*, which refers to the percentage of the drug in the blood excreted from the organ on each passage through the same organ.

$$CL_{organ} = Q * E , \quad (1.23)$$

where Q is the blood flow of the organ and E the extraction ratio.

Finally, the toxicity factor is also taken into consideration. While dosage undeniably influences the likelihood of toxicities, the active substance of a drug can interact with enzymes and receptors outside the intended biological target, leading to

side effects. In addition, there is the possibility of the production of toxic metabolites. For these reasons, it's crucial to be able to predict various side effects that may arise in the body, particularly *hepatotoxicity* [79], [80].

Predicting these properties, along with assessing the affinity of new molecules to the receptor in the early stages of designing a potential drug, determines its suitability for further experimental *in vitro* studies.

1.6 Survey of Related Research

The identification of ligands for MC4R has been an active area of research due to its critical role in regulating energy homeostasis, appetite, and body weight, as previously noted. Multiple studies have focused on characterizing both agonists and antagonists of MC4R to understand its function and therapeutic potential. This subsection reviews the methodologies and findings from recent surveys that have contributed to discovering MC4R ligands, focusing on their potential as treatments for monogenic obesity.

As evident, this thesis concentrates on a combined drug discovery strategy based on machine learning models and molecular docking tools. Researchers have previously addressed such methods to accelerate the drug-designing process and direct the subsequent *in vitro* experiments. For instance, in 2018, Zhang *et al.* [81] employed this combined approach to identify acetyl-CoA carboxylase (ACC) inhibitors. They developed machine learning models using molecular descriptors to distinguish between active and inactive chemical compounds, thus refining the screening process. Their research proposes that an initial model-based search using machine learning classifiers, followed by molecular docking, can enhance the precision in identifying potential hits for a specific target. Additionally, recent efforts have focused on developing computational tools that integrate molecular modeling and machine learning to advance rational drug design. In 2023, Xia *et al.* [82] discuss their efforts in creating such a tool, detailing the challenges and limitations encountered during their study, as well as outlining future steps for further improvements.

Regarding MC4R, numerous studies have been conducted to identify potential agonists for the treatment of obesity, utilizing the advantages of both data science and computational chemistry:

In 2002, Andersson and Lundstedt [83] introduced a hierarchical design approach for establishing QSAR models focused on MC4R. This method prioritizes the selection of a representative subset of compounds from a larger candidate set for QSAR model development. The utilized set of compounds shared a phenyl ring as a common structural core and was characterized as a suitable scaffold for investigation. Their model demonstrated strong predictive capabilities and has been used to design new chemical compounds with enhanced activity.

A few years later, in 2005, Cai *et al.* [84] conducted extensive studies to convert non-selective ligands of the melanocortin family receptors into selective analogues. They used computational techniques to explore the 3D structures of the bioactive forms of Melanotan II (MT-II) and SHU9119, by developing 3D topographical models of the ligands. Additionally, they explored the SAR of α -MSH, β -MSH and γ -MSH in various ways to use as templates for potent and stable ligands.

Numerous efforts targeting MC4R have successfully identified potent and selective agonists. In 2018, Gonçalves, Palmer and Meldal [16], provide valuable insights on recent advancements and collective efforts in this area, particularly

emphasizing the critical structural differences in molecules that lead to strong selectivity. They point out that the majority of efforts focus on peptide agonists, specifically cyclic peptides, which show the highest agonist activity. This focus on peptides is largely driven by their structural similarity to the natural hormone POMC, which plays a key role in regulating MC4R activity.

In 2019, Falls and Zhang [85] constructed a homology model of MC4R to perform docking studies. To optimize and validate their model, they selected the endogenous ligand α -MSH and the small molecule agonist THIQ. They conducted point mutation studies on four different MC4R mutations to assess the impact of these polymorphisms on the binding affinity of α -MSH and Setmelanotide. The researchers suggest that their work could serve as a valuable platform for designing future selective and potent ligands, particularly those that can simultaneously interact with the orthosteric and allosteric binding sites, addressing the lack of available crystallographic data.

Similar to this thesis, a year later Martin *et al.* [86] employed the novel co-crystal structure of MC4R-SHU9119 (PDB: 6W25) for their molecular docking experiments. Their goal was to enhance the selectivity of MC4R peptide ligands by designing a series of cyclic peptides based on this crystal complex, leading to the discovery of ligands with improved affinity for MC4R. Additionally, they conducted *in vitro* pharmacological characterization of these analogues for further evaluation.

In summary, the existing research on MC4R agonists has made significant progress in identifying potent and selective compounds as ligands. However, there are still ongoing challenges in achieving long-term stability and specificity. This study employs a combination of machine learning and molecular modeling techniques to identify potential MC4R ligands, with a specific focus on non-peptidic agonists. The following sections will provide an exploration of the methodology and related results.

Chapter 2: Materials and Methods

2.1 Computational Tools

All computational tasks were performed using a laptop equipped with an Intel(R) Core(TM) i3-4000M processor (CPU) operating at 2.40GHz, 6 GB DDR3 RAM, a 1 TB HDD for storage, and an Intel HD 4600 graphics card. The system operated on Windows 10 Home 64-bit. The hardware and software configurations ensured efficient handling of the computational requirements of this research.

2.2 Machine Learning Model Development

Visual Studio Code, an Integrated Development Environment (IDE), was used along with Python 3.9.12 for developing and running scripts for data processing and analysis. Python is a high-level programming language commonly associated with data-related tasks, due to its simplicity and range of available packages and libraries.

2.2.1 Dataset

The dataset utilized for the purposes of this thesis was obtained from ChEMBL (<https://www.ebi.ac.uk/chembl/>). As noted, it is a chemical database of bioactive molecules with pharmaceutical properties, managed and curated manually by the European Bioinformatics Institute (EBI), part of the European Molecular Biology Laboratory (EMBL). As the largest open-access source of medicinal chemistry data, it provides information on small molecules, including their chemical structure, biological activity, mechanisms of action against biological targets, and genome. This knowledge is derived from a variety of scientific literature, bioassays, and calculated properties. The latest versions of ChEMBL (Figure 2.1) include patent data and additional information from other databases. Datasets are available in multiple formats, typically exported as Comma Separated Values (CSV) files [87], [88].

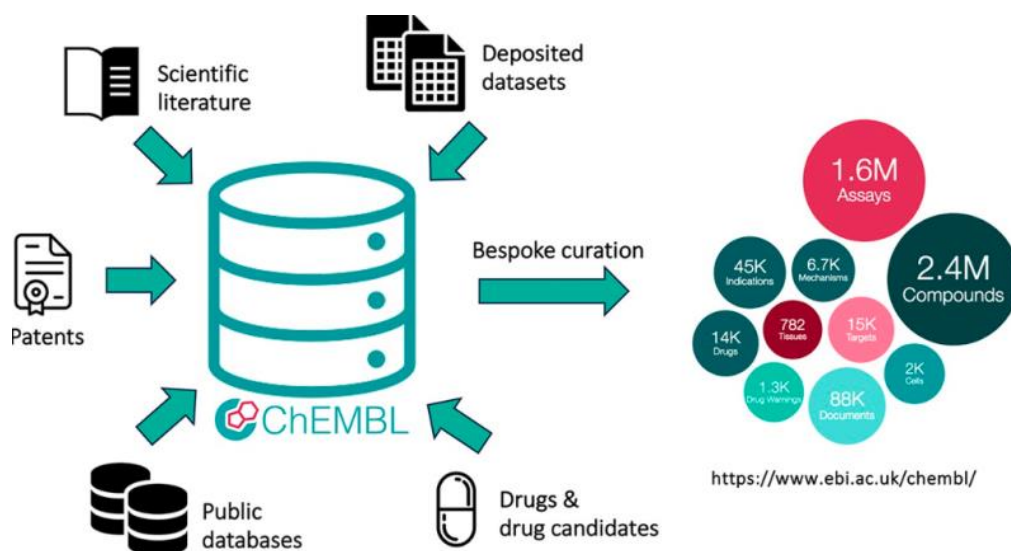


Fig. 2.1: Illustration of the ChEMBL database [87].

The dataset is a subset of activities titled as “*Bioactivity data for target CHEMBL259-IC50*”, containing 2,177 records. These records correspond to chemical compounds tested against the human MC4R and additional information, such as the *Molecule ChEMBL ID*, *Molecule Name*, *Molecular Weight*, *AlogP*, *SMILES* and *IC₅₀* associated characteristics like *Standard Relation* and *Standard Value*.

The IC₅₀ or half-maximal inhibitory concentration value, is a measure of the concentration of a drug or compound required to inhibit a particular biological or biochemical process by 50% [89]. Typically, when dealing with this type of binary classification problem a threshold corresponding to the Standard Value of the IC₅₀ is selected. This is a common practical approach in the early stages of drug discovery. In this case, a threshold of 10 μM is employed as a filter to configure the initial dataset obtained from ChEMBL, to label chemical compounds as “Active” or “Not Active” towards the target of MC4R. This general rule considers chemical compounds with IC₅₀ less than or equal to 10 μM as “Active”, indicating sufficient potency towards the target. Chemical compounds with IC₅₀ greater than 10 μM are considered “Not Active” [90].

For duplicate chemical compounds, the IC₅₀ values were reviewed. If the values differed among duplicates, the median IC₅₀ was calculated and used to replace the existing values, ensuring consistency. Additionally, chemical preprocessing was performed on the SMILES representations, where salts and counter ions were removed to standardize the compounds. This procedure improves data quality by focusing solely on the compounds, ensures that different representations of the same molecule are treated as identical and simplifies the subsequent calculations. After completing this process, duplicate chemical compounds with identical information were removed and records with missing data were excluded. This resulted in a curated dataset of 1,906 chemical compounds. An IC₅₀ threshold filter of 10 μM was then applied to the dataset. This process is used to label each record, ensuring accurate classification for further analysis. Consequently, 825 of these compounds were categorized as “Active” and 1,081 as “Not Active”. To generate features for machine learning models, the RDKit package was used in a Python environment to compute molecular descriptors. *Chem.SmilesMolSupplier*, a function in the RDKit library, was utilized to read chemical structures from SMILES strings of each chemical compound in the dataset. Then, *Descriptors._descList*, a list within the RDKit library containing tuples of descriptor names and their functions, was used to compute 208 molecular descriptors. These descriptors are predefined in RDKit to represent properties of molecules, such as the *MolWt*, which corresponds to their molecular weight.

Table 2.1 lists such representative examples. Their definitions and interpretations are available and can be explored through relevant literature [91].

Table 2.1 Representative examples of RDKit molecular descriptors.

RDKit: Examples of Molecular Descriptors		
MolWt	BertzCT	NumHAcceptors
MaxAbsEstateIndex	Chi0	MolLogP
qed	Kappa1	fr_Al_COO
NumValenceElectrons	LabuteASA	FpDensityMorgan1
MaxPartialCharge	PEO_VSA1	BCUT2D_MRLOW
BalabanJ	NumAromaticRings	HallKierAlpha

At this point of the thesis, I would like to acknowledge *Dr. Matsoukas Minos* and *PhD candidate Panagiotopoulos Vasilios* for providing the dataset, appropriately processing it, and calculating the molecular descriptors.

2.2.2 Practical implementation

The ultimate goal is to build a high-accuracy machine learning model that predicts the activity of chemical compounds against MC4R. Initially, the dataset was examined to identify and remove zero-variance features. *Zero-variance features*, or *constant features*, have the same value for every record within the dataset, rendering them uninformative. In the utilized dataset, molecular descriptors with all zero values were observed and removed to improve the model's performance.

In this case, as evident, we are dealing with a binary classification problem. The first step for implementing the machine learning process is to split the initial dataset into training and test subsets: the training set is used to fit training models, whereas the test set is used to evaluate the performance of the trained model. Typically, the training set comprises the largest portion of the dataset. During the first process, the utilized classification algorithm sees the input features and the corresponding labels of each record in the data. Then, the trained model is used to predict the labels of unseen data from the test set to assess its generalization ability. Finally, these predictions are compared to the actual labels of the test set to measure the model's *accuracy* or other relevant metrics, such as *sensitivity* or *specificity*. The data was split using a 70-30 ratio to approach the problem.

Subsequently, the *preprocessing.StandardScaler* function from the *scikit-learn* library was used to scale each feature. This function standardizes features by removing the mean and scaling to unit variance, as described in subsection 1.3.2. The scaler was first applied to the training set using the *fit_transform* method. The same scaler was then applied to the test set using the *transform* method, ensuring that the test set was scaled based on the training set's parameters, preventing any biases.

After scaling the data, the next step involved conducting feature selection to pinpoint the most significant features and eliminate those that don't contribute to the model's predictive power. Specifically, tree-based algorithms, such as CART or ensembles of decision trees like Random Forest, offer an approach that evaluates the importance of features based on the reduction in the criterion applied to split points, like Entropy or Gini described in subsection 1.3.1. The importance of the features is determined by the *feature_importances_* attribute, calculated as the mean and standard deviation of accumulation of the impurity reduction within each tree [92], [93]. The function was executed for 50 epochs, during which the top features were retained for each epoch. The frequency of appearance was evaluated by counting the occurrences of each feature, as a way to rank their significance. By applying this feature selection technique, the top 10 features were identified and selected, which were subsequently used to execute the machine learning algorithms, as an approach to enhance the performance of our models.

To systematically explore and evaluate all possible combinations of the 10 available features, we employed the $\binom{n}{k}$ binomial coefficient method.

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}, \quad (2.1)$$

where n represents the total number of available features, and k corresponds to the number of features chosen in each combination.

A custom function was developed utilizing the *combinations* object from the *itertools* module, which generates all possible feature combinations to be used as input for the algorithms.

In total, 11 classifiers from the scikit-learn library were tested on the training data, and their performance was evaluated using K-fold Cross-Validation over 10 folds. This procedure was conducted to search for the best feature combination, based on the *accuracy* metric, ultimately selecting the classifier with the optimal performance.

The classifiers, their module and respective class are presented in the following table (Table 2.2).

Table 2.2 List of employed classifiers.

Module	Classifiers (Class)
<i>sklearn.neighbors</i>	<ul style="list-style-type: none"> • NearestCentroid • KNeighborsClassifier
<i>sklearn.naive_bayes</i>	<ul style="list-style-type: none"> • GaussianNB
<i>sklearn.discriminant_analysis</i>	<ul style="list-style-type: none"> • LinearDiscriminantAnalysis
<i>sklearn.linear_model</i>	<ul style="list-style-type: none"> • LogisticRegression • Perceptron
<i>sklearn.svm</i>	<ul style="list-style-type: none"> • LinearSVC
<i>sklearn.tree</i>	<ul style="list-style-type: none"> • DecisionTreeClassifier
<i>sklearn.ensemble</i>	<ul style="list-style-type: none"> • RandomForestClassifier • GradientBoostingClassifier • ExtraTreesClassifier

The selected model was trained and then further evaluated on the test set over 10 epochs to assess its performance. This evaluation aimed to detect any biases, overfitting, or underfitting. In addition, this helped in understanding how well the model generalized on unseen data estimating the performance metrics by averaging the results over multiple epochs.

Finally, statistical analysis was performed for the molecular descriptors in the optimal feature combination. While it is not always mandatory, it is generally beneficial for the features within a model to present statistical difference between classes in binary classification. To identify features that exhibit a significant statistical difference between “Active” and “Not Active” chemical compounds, the Mann-Whitney U test was applied with a significance threshold of $p \leq 0.001$. This approach helps to identify key features for further investigation of their potential biological and chemical relevance.

2.3 Molecular Docking Experiments

In the present study, the molecular docking experiments conducted to advance our understanding of MC4R are presented. As mentioned above, this technique is essential in drug design and discovery since it helps identify potential interactions between small molecules and their target proteins. By simulating these interactions, we aim to identify potential ligands in an effort to understand their mechanisms of action. This approach complements our previous work with machine learning models by providing detailed insights into the behavior of molecules.

2.3.1 Crystal structure preparation

To conduct the molecular docking experiments, the available bibliography was researched to identify crystal structures of the human MC4R (hMC4R). This involved a thorough review of scientific literature and databases to ensure that we employed accurate and relevant structural data for the purposes of this thesis. As a result, the crystal structure with entry ID **6W25** from the *RCSB Protein Data Bank* was selected.

The *Protein Data Bank (PDB)* was founded in 1971 as the first open-access digital data repository, housing 3D structures of biomolecules such as proteins and nucleic acids. The archive of this database (Figure 2.2) is managed by the Worldwide Protein Data Bank (wwPDB) partnership [94]. The *Research Collaboratory for Structural Bioinformatics Protein Data Bank (RCSB PDB)* serves as the United States data center for the global PDB archive. Funded by the National Science Foundation, the National Institutes of Health, and the US Department of Energy, the RCSB PDB has become an indispensable tool for both research and education in fields such as biology, health, and biotechnology. Understanding the 3D structures of biological macromolecules is crucial, particularly for studies aiming to elucidate their functions in human health and disease. These 3D structure data are primarily obtained through NMR, electron microscopy, and macromolecular crystallography [95].

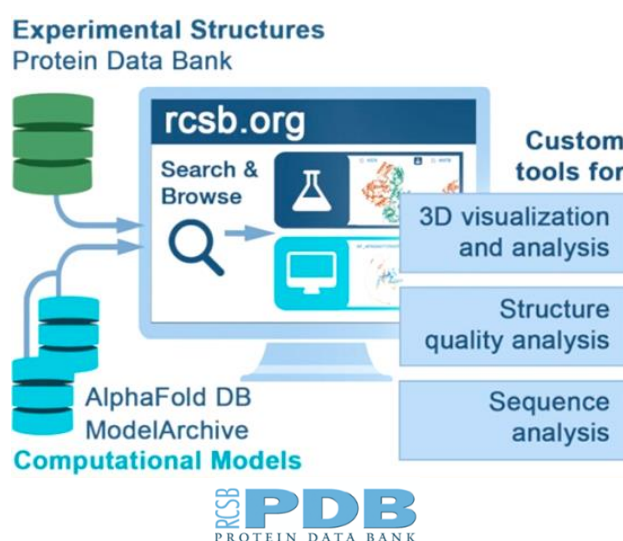


Fig. 2.2: Illustration of the PDB (adapted from [96]).

The crystal structure with entry ID 6W25 corresponds to MC4R in complex with SHU9119, obtained through X-ray diffraction at a resolution of 2.75 Å.

SHU9119 (C₅₄H₇₁N₁₅O₉) is an antagonistic peptide of MC4R. It is a shortened, modified, and circularized derivative of NDP- α -MSH, displaying a related binding mode. SHU9119 binds in the upper part of MC4R ligand-binding pocket (LBP) in a pattern similar to agonists, making it suitable for experiment conduction in this study. It acts as an antagonist due to specific interactions that prevent receptor activation [22].

The 3D illustration of this complex, as provided in the PDB, and the structure of SHU9119, are presented in Figure 2.3.



Fig. 2.3: 6W25: 3D crystal structure of MC4R in complex with SHU9119 (left) and chemical structure of SHU9119 (right) [97].

To ensure the successful execution of subsequent experiments, the crystal structure was prepared using the *Maestro* platform, a comprehensive Graphical User Interface (GUI) for molecular modeling. It is part of the Schrödinger Suite - a collection of computational tools designed for drug discovery. It integrates various functionalities to perform computational tasks, including molecular visualization, structure preparation, and molecular docking. The *Protein Preparation Wizard*, a tool within *Maestro*, addresses common structural issues and generates reliable all-atom protein models [98]. During this process, the pH was adjusted to 7.0 ± 0.5 . Non-receptor entities not required for the docking analysis, such as the ligand and water molecules not involved in bridging interactions between the ligand and target amino acids were removed. Missing hydrogen atoms were added, and protonation states for ionizable residues were assigned. Additionally, the receptor underwent energy minimization using the OPLS3 force field to achieve a stable, low-energy configuration.

2.3.2 Software applications

In the molecular docking experiments, two distinct approaches were adopted. The first approach involved docking studies into the crystallized MC4R with natural compounds possessing a molecular weight below 500 Da, which aligns with Lipinski's Rule of 5 for rational drug design. The second approach concentrated on

docking with natural compounds exceeding 500 Da, reflecting the molecular weights of approved medicinal products targeting the same receptor (for example, Setmelanotide).

The selection of natural compounds for this study was performed with purpose and strategic intent. Natural compounds include organic molecules produced by organisms, especially microbes, and plants, as secondary metabolites. They are known for their non-toxic properties and bioactivity, making them ideal candidates for drug discovery and development [99].

In this thesis, 2000 natural compounds were tested against hMC4R. These natural compounds were retrieved from the *ZINC database* in Structure Data File (SDF) format, with a significant proportion originating from the natural products library of the Specs Company, a leading provider of compound management services and research compounds [100].

The *ZINC database* is a comprehensive collection of chemical compounds formatted for research applications, such as virtual screening software and molecular binding experiments. Developed by the Irwin and Shoichet Laboratories at the Department of Pharmaceutical Chemistry, University of California, San Francisco (UCSF), ZINC provides access to biologically relevant 3D molecular structures. The database offers over 750 million commercially available compounds, organized into subsets and catalogs with advanced filtering options to facilitate small molecule searches. A key feature of ZINC is its integration of market-available compounds with high-value substances, including metabolites, natural products, and drugs from scientific literature. ZINC is regularly updated with new data and is accessible online in various versions, such as ZINC15 and ZINC20 [101], [102].

For a thorough analysis and validation, three *in silico* tools were employed in the molecular docking experiments. These tools were selected for their complementary features and robust capabilities in molecular docking:

- i. **Webina.** Webina (<https://durrantlab.pitt.edu/webina/>) is an open-source library and web application developed by the Department of Biological Sciences of the University of Pittsburgh. It runs AutoDock Vina, a popular docking program, entirely in the web browser, producing ligand poses and the corresponding docking scores. To be more precise, docking calculations take place on the user's computer instead of a remote server, allowing the user to visualize results (poses) on their browser [103].

In addition, Grid-based Ligand Docking with Energetics (Glide, for short), a molecular docking software within Maestro developed by Schrödinger, Inc., was utilized to further investigate the chemical compounds that were shortlisted using Webina. Glide uses a series of hierarchical filters to search for possible locations of the ligand in the active-site region of the receptor. The shape and properties of the receptor are represented on a grid by different sets of fields that provide progressively more accurate scoring of the ligand pose. As a result, a combination of force-field-based scoring functions predicts the binding affinity of the ligand to the target receptor [104]. This thesis incorporates two Glide docking methodologies:

- ii. **Glide-SP.** Glide-SP or Standard Precision Glide is a “softer”, more forgiving function adept at identifying ligands with a reasonable propensity to bind, even in cases where the Glide pose has notable imperfections. This version minimizes false negatives, providing a balance between computational

efficiency and accuracy, when docking on a large number of chemical compounds is necessary.

- iii. ***Glide-XP***. *Glide-XP* or Extra Precision *Glide* imposes severe penalties for poses violating established physical chemistry principles, such as ensuring that charged and strongly polar groups are properly exposed to solvent. It is more adept at reducing false positives and particularly useful in lead optimization or other studies where only a limited number of compounds are considered experimentally, requiring each computationally identified compound to be as high in quality as possible [105], [106].

It should be noted that the natural compounds were prepared accordingly for the docking experiments. Each software has its method for converting SDF files into the required format. *Webina* automatically formulates the molecular structures using the *PDBQTConverter App*. Similarly, *Maestro* employs *LigPrep* to prepare the molecules in Protein Data Bank, Partial Charge (Q) and Atom Type (T) (PDBQT) format, which is compatible with *Glide*. During the conversion of files from SDF to PDBQT format, hydrogen atoms are added to the molecule, and partial atomic charges to the atoms are assigned to the selected pH (7.0 ± 0.5) accordingly. Hydrogen atoms are significant for the geometry of the compounds, and the charges are essential for the electrostatic interactions modeling.

2.3.3 Validation process

The objective of the molecular docking experiments in this thesis is to identify a list of chemical compounds that act as potential ligands to hMC4R. To accomplish this, various validation steps are conducted to confirm the accuracy of the docking results.

Firstly, the superposition of the ligand SHU9119 was performed using all three *in silico* tools to find a configuration that closely matches the one in the crystal structure of 6W25. The binding affinity of this configuration was recorded for each noted software. These alignments were used as a reference point for the docking experiments with natural compounds.

To define the region of interest for the target protein hMC4R, an appropriately sized grid box was generated. Generally, it is advisable to make this 3D box as small as possible while still encompassing the protein's active site, as defined by the ligand in complex. This approach provides a more accurate measure of the effective search space. The grid box is specified in terms of the Cartesian coordinate system. In this case, for the box center is located at $(x, y, z) = (135, 4, 103)$ and for the box size the dimensions in Å are $(x, y, z) = (17, 17, 18)$.

Additionally, the interaction pattern between SHU9119 and hMC4R was documented to facilitate the selection process of the final molecules. The interaction pattern between each natural product and hMC4R was recorded and compared to that of the ligand in complex. To achieve this, the open-access *Protein-Ligand Interaction Profiler* (PLIP, <https://plip-tool.biotec.tu-dresden.de/plip-web/plip/index>) was used for analyzing the docking results from *Webina*. The PLIP web server provides detailed results for all binding sites in the input structure, including atom-level binding information, and a 3D interactive visualization. For the docking results obtained from *Glide* for the shortlisted molecules, as previously noted, a tool within *Maestro* was used to generate a 2D projection of the protein-ligand interactions. The

Ligand Interaction Diagram visually presents how the ligand interacts with the target protein, highlighting both the binding interactions and the solvent-exposed regions of the molecules. A 4 Å cut-off was used. This approach captures the majority of electrostatic interactions and any hydrogen interactions that are significant in our protein-ligand complex.

2.4 Machine Learning and Docking Hybridization

A significant focus of this thesis is to explore how the combination of machine learning models with molecular docking results can enhance the drug design research process. To accomplish this, the model that provided the strongest classification was employed to categorize the final chemical compounds obtained from the *in silico* experiments.

For this part of the study, *MetaboAnalyst 6.0*, a freely accessible platform for comprehensive metabolomics data analysis and interpretation (<https://www.metaboanalyst.ca/>) was employed. The platform offers various applications including statistical analysis, biomarker analysis, and dose response analysis. It was selected for its automated procedures and ability to facilitate the analysis of complex data, including results derived from machine learning models.

Initially, for the natural compounds that were distinguished among the 2000 as potential ligands of hMC4R, the RDKit molecular descriptors were calculated, as described in subsection 2.1.1. A new CSV file was created containing the IDs of the chemical compounds from the ChEMBL dataset, their labels, and the molecular descriptors included in the best feature combination. This file was appended with the IDs and the identical molecular descriptors of the final natural compounds.

The CSV file was uploaded to MetaboAnalyst's Biomarker Analysis module, which offers the ROC curve-based evaluation approach to potential biomarkers identification and model performance evaluation.

MetaboAnalyst detected 3 groups of data in the CSV: Active, Not Active and Unlabeled chemical compounds. In this module, the platform provides multiple data normalization procedures. For consistency, we opted for the Auto Scaling procedure. The best combination of biomarkers was manually selected to create the biomarker models using the Random Forest classifier. This module allows users to hold out a subset of samples for extra validation purposes, as well as to predict class for new samples (Unlabeled). This performs 100 cross-validations and averages the results to produce the ROC curve. Additionally, it calculates the predictive accuracy based on these validations and predicts the labels for the final natural compounds based on their probability scores.

2.5 ADMET Prediction

To further validate and cross-check the results, two different open-access computational tools were employed to study the metabolic properties of the final molecules. The SMILES strings were used to generate the ADMET properties. Both computational tools utilize a neural network framework to generate predictions and present the results in a tabular format. We aimed to compare the characteristics of our selected natural compounds with those of the FDA-approved drug Setmelanotide.

The first tool we employed is the *ADMETlab 3.0* web server (<https://admetlab3.scbdd.com/>), which predicts pharmacokinetics and toxicity

properties of molecules with comprehensive and precise models, covering major endpoints. ADMETlab 3.0 is developed and maintained by the CBDD team of Central South University, the HIT team of National University of Defense Technology, and the Innovation Institute for Artificial Intelligence in Medicine of Zhejiang University [107].

In addition, we decided to utilize a novel ADMET prediction tool, *ADMET-AI 1.3.1* (<https://admet.ai.greenstonebio.com/>), developed by the Department of Computer Science at Stanford University. This tool offers an efficient method for predicting ADMET properties, providing results at a faster rate [108].

Chapter 3: Results

3.1 Machine Learning Results

In this section, the results from the Machine Learning models are presented.

3.1.1 Optimal feature combination

Among the 10 features identified through the feature selection process, a combination of 7 RDKit molecular descriptors proved to be the most effective for building a robust machine learning model.

Table 3.1 lists the final results of the K-fold cross-validation procedure, highlighting the molecular descriptors that achieved optimal model performance. The classifier that attained the highest efficiency is the *RandomForestClassifier*, using the default parameters from the scikit-learn library. Furthermore, a brief interpretation is provided [109] - [111].

Table 3.1 *Optimal feature combination.*

RDKit Descriptor	Brief Interpretation
<i>VSA_EState6</i>	The 6 th of the 10 VSA_EState molecular descriptors. They quantify the surface area contributions of different types of atoms or bonds within a molecule.
<i>MaxAbsEStateIndex</i>	It refers to the maximum absolute value of the E-State indices across all atoms in the molecule.
<i>PEOE_VSA8</i>	The 8 th of the 14 PEOE_VSA molecular descriptors. They intend to capture the direct electrostatic interactions within a certain range of atomic partial charges of $0 \leq x \leq 0.05$.
<i>Kappa2</i>	The 2 nd of the 3 kappa shape indexes. It is a measure of molecular branching and connectivity.
<i>MolMR</i>	The Molecular Weight of a molecule expressed in Da.
<i>BCUT2D_MRLOW</i>	Lowest eigenvalue weighted by Crippen Molar Refractivity (Crippen MRR)
<i>Kappa3</i>	The 3 rd of the 3 kappa shape indexes. It involves complex connectivity information and is influenced by the presence of rings or cyclic structures.

To provide a deeper understanding of the concepts listed in Table 3.1, a more detailed examination is necessary:

The *VSA_Estate* and *MaxAbsEStateIndex* are molecular descriptors that use E-State indices and surface area contributions. EState or E-state is a concept developed by Kier and Hall [112] that corresponds to an Electrotological State

index, which is used to describe the electronic environment and topology of atoms within a molecule. Thus, the E-State index is a measure of the electronic accessibility of a specific atom and can be interpreted as a probability of interaction with another molecule. However, this type of descriptors is not considered electronic, but descriptors of atom polarity and steric accessibility [109].

The Van der Waals Surface Area (VSA) is a value obtained by considering the shape of each atom to be a sphere with a radius equal to that of Van der Waals. At this point, it is important to note that the surface area of an atom in a molecule is the amount of surface area of that atom not contained in any other atom of the molecule [113].

The atomic partial charge is calculated using the Partial Equalization of Orbital Electronegativity (PEOE) method, which was developed by Marsili and Gasteiger [114] through a topological iterative approach. In other words, partial charges are assigned to atoms of a molecule based on their electronegativities and neighboring atoms. As evident, the PEOE_VSA descriptors are numerical values that correspond to the electron density distribution across molecular surface areas.

The Kier alpha-modified shape or kappa descriptors are a group of molecular descriptors developed by Kier and Hall, which are associated with the different shape contribution of heteroatoms and hybridization states. As a result, they offer a way to describe the structural characteristics of molecules, which is crucial for drug design.

Finally, Burden-Cas-University of Texas eigenvalues (BCUT) are based on the Burden approach, which considers three matrices whose diagonal elements correspond to i) atomic charge-related values, ii) atomic polarizability-related values, and iii) atomic H-bond abilities. The BCUT2D descriptors are a specific type of BCUT descriptors calculated based on a 2D representation of the molecular structure [109].

Table 3.2 displays the mean and standard deviation of the performance metrics obtained from the K-fold cross-validation for the *RandomForestClassifier* with the optimal feature set. As evident, the scores are relatively high, demonstrating the effectiveness of the selected feature combination and the classifier.

Table 3.2 Average performance metrics from K-fold cross-validation for the *RandomForestClassifier*.

Metrics	MEAN	STD.DEV
Sensitivity	90.75 %	1.53 %
Specificity	94.19 %	0.83 %
Accuracy	92.71 %	1.07 %
F1-score	0.91	0.01
Precision	0.92	0.01
AUC	0.97	0.01

3.1.2 Model validation

As previously noted, the model was separately trained over 10 epochs to provide a comprehensive view of its performance, to detect biases, underfitting or overfitting.

Table 3.3 presents the average performance metrics of the test set evaluated over 10 epochs for the selected classifier. Furthermore, Figure 3.1 illustrates the ROC curve for the final results across 10 epochs, providing a visual representation of the performance.

Table 3.3 Average performance metrics over 10 epochs for the *RandomForestClassifier*.

Metrics	MEAN (%)	STD.DEV (%)
Sensitivity	93.02	0.87
Specificity	95.25	0.89
Accuracy	94.28	0.44
F1-score	0.93	0.00
Precision	0.94	0.01
AUC	0.98	0.00

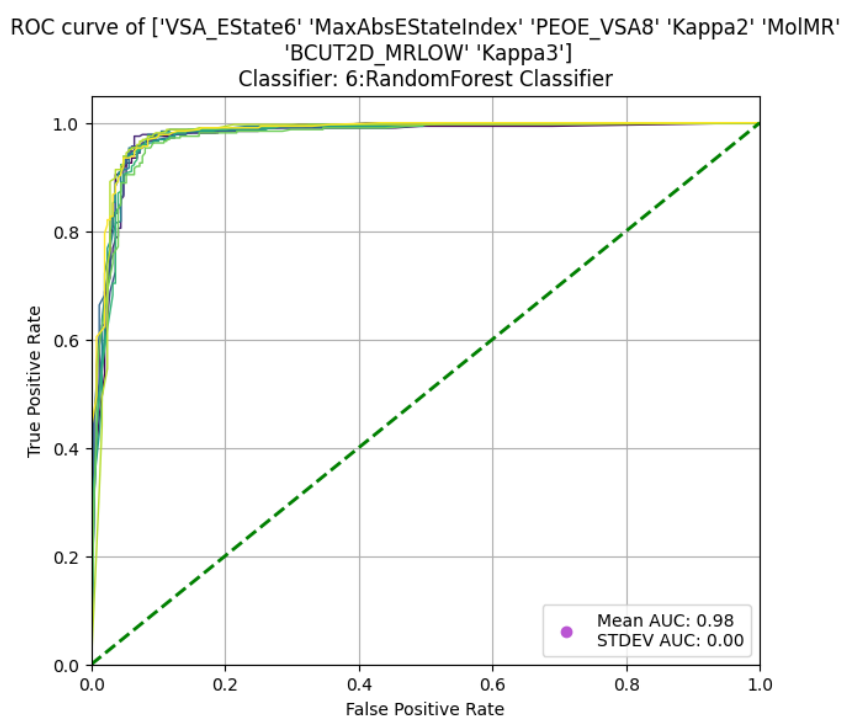


Fig. 3.1: ROC curve of the optimal feature combination, using *RandomForestClassifier* over 10 epochs (AUC=0.98).

For comparative analysis, Table 3.4 presents the mean and standard deviation of the performance metrics, for all classifiers used in this thesis, evaluated over 10 epochs, highlighting the Random Forest classifier.

Table 3.4 Average performance metrics (mean and standard deviation) over 10 epochs for all classifiers.

Model	Sensitivity (%)	Specificity (%)	Accuracy (%)	F1	Precision	AUC
KNN	91.45 1.69	91.64 0.79	91.56 1.00	0.90 0.01	0.89 0.01	0.95 0.01
Bayes.	92.62 5.50	80.77 6.79	85.91 1.95	0.85 0.01	0.79 0.05	0.92 0.01
LDA	88.63 1.56	87.62 1.96	88.06 1.20	0.87 0.01	0.85 0.02	0.94 0.01
LogReg	87.18 1.35	89.78 1.13	88.65 0.45	0.87 0.01	0.87 0.01	0.94 0.01
Percep.	74.15 20.21	84.07 6.32	79.77 10.64	0.75 0.16	0.77 0.12	0.84 0.14
SVM	87.22 1.57	89.07 1.10	88.27 1.08	0.87 0.01	0.86 0.01	0.94 0.01
RF	93.02 0.87	95.25 0.89	94.28 0.44	0.93 0.00	0.94 0.01	0.98 0.00
CART	89.80 1.49	92.22 1.24	91.17 0.60	0.90 0.01	0.90 0.01	0.91 0.01
XGB	92.18 1.99	94.88 0.90	93.71 1.00	0.93 0.01	0.93 0.01	0.97 0.01
Ada	92.54 1.21	91.60 0.9	92.01 0.86	0.91 0.01	0.89 0.01	0.97 0.01
GBT	93.87 1.59	94.14 0.88	94.02 0.69	0.93 0.01	0.92 0.01	0.97 0.00
ET	93.79 1.99	94.26 1.29	94.06 0.99	0.93 0.01	0.93 0.01	0.97 0.01

3.2 Statistical Analysis Results

The following figures (Figure 3.2-3.5) illustrate the results of the statistical analysis for the molecular descriptors detailed in Table 3.1, specifically comparing the two classes of a single feature. The figures include boxplots and ROC curves for the features where there is a statistically significant difference, and the AUC exceeds a threshold of 0.8. As evident these features are **MaxAbsEstateIndex**, **PEOE_VSA8**, **Kappa_2** and **BCUT2D_MRLOW**.

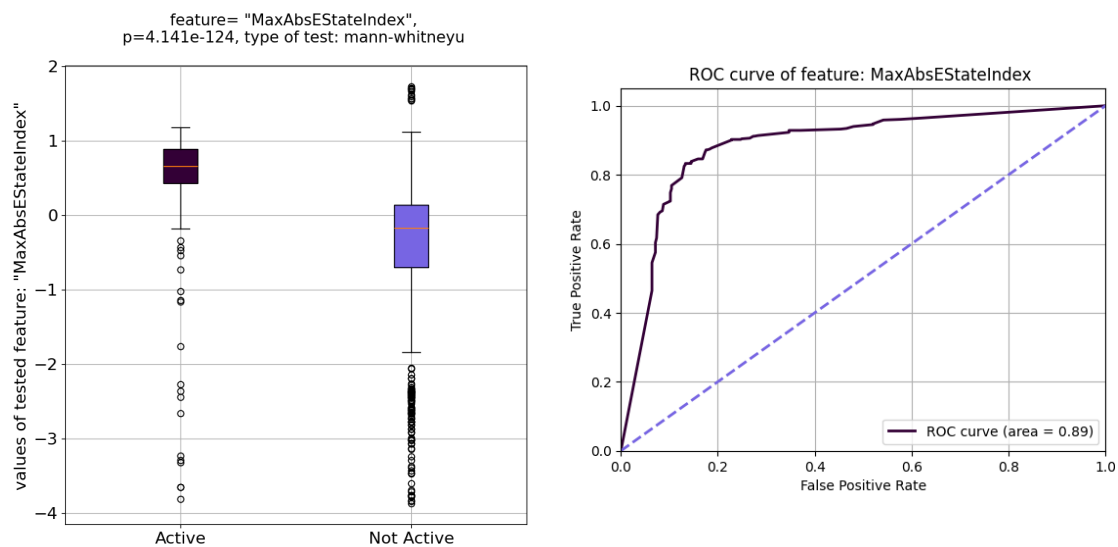


Fig. 3.2: *MaxAbsEstateIndex* boxplots (left) and ROC curve (AUC=0.89) between classes (right).

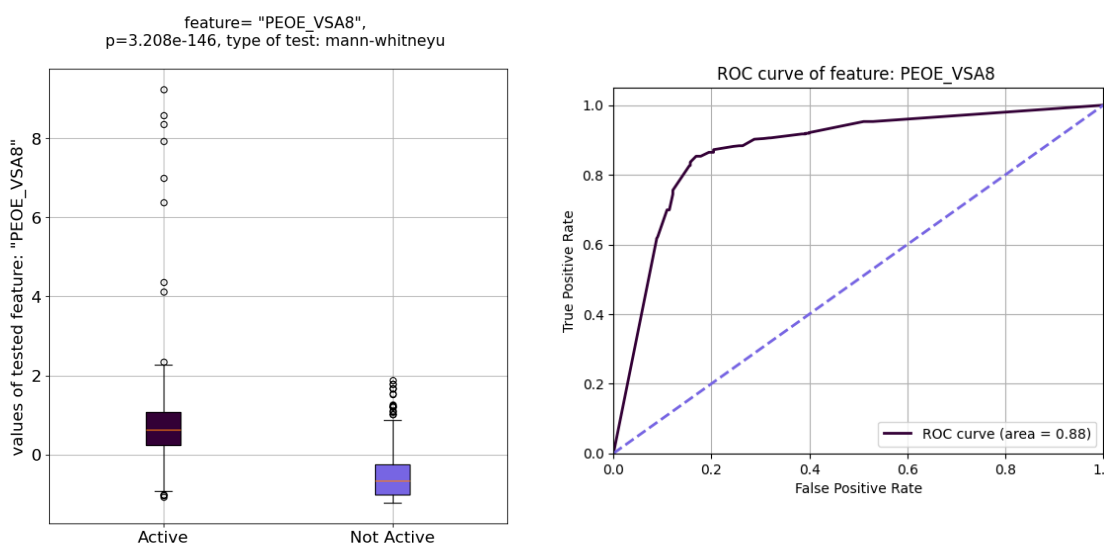


Fig. 3.3: *PEOE_VSA8* boxplots (left) and ROC curve (AUC=0.88) between classes (right).

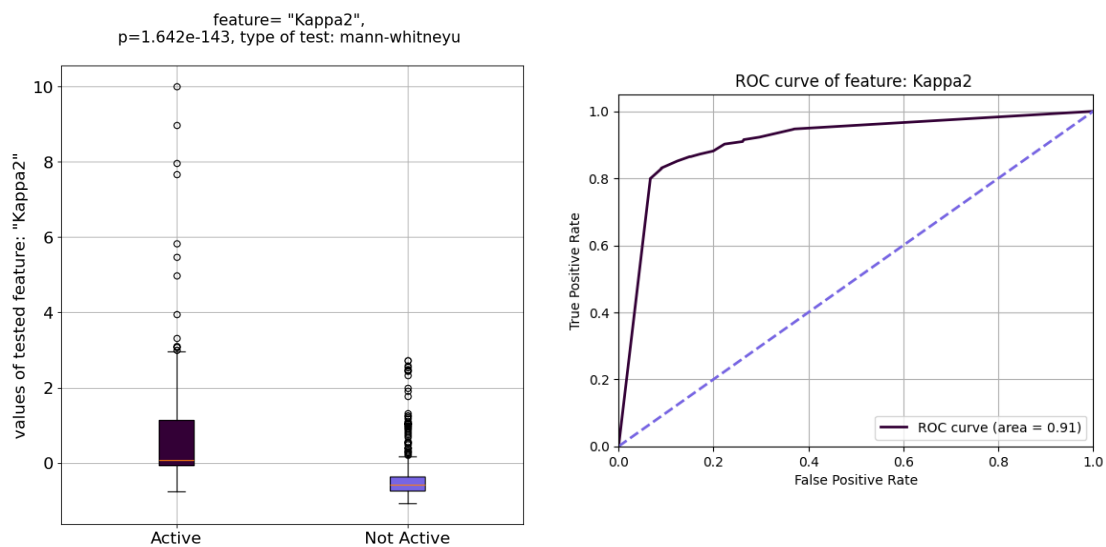


Fig. 3.4: *Kappa2* boxplots (left) and ROC curve (AUC=0.91) between classes (right).

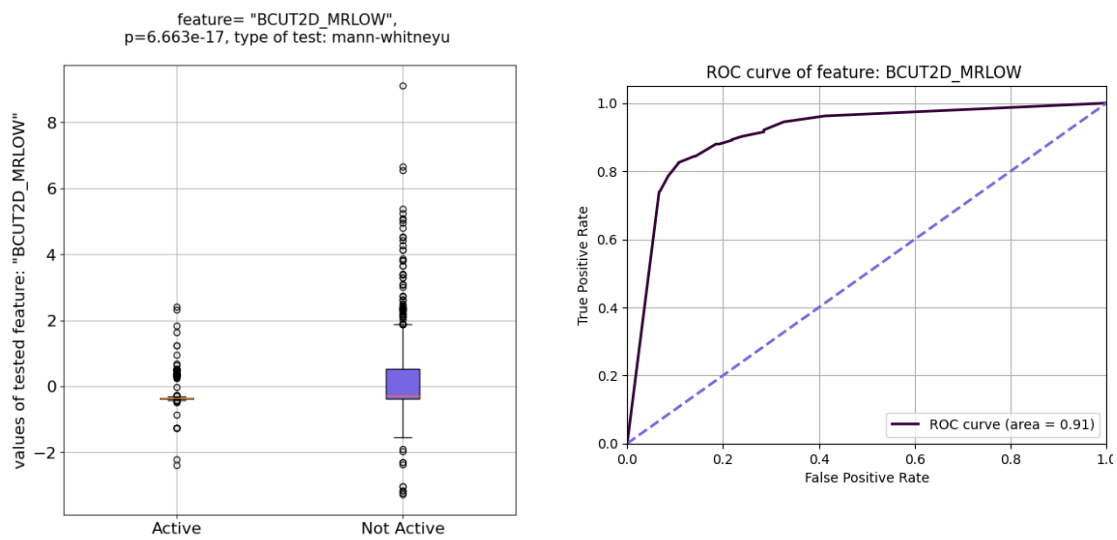


Fig. 3.5: *BCUT2D_MRLow* boxplots (left) and ROC curve (AUC=0.91) between classes (right).

3.3 Molecular Docking Results

Maestro and PLIP were utilized to visualize and analyze the interactions between hMC4R and SHU9119. The derived results were further validated against the existing literature.

Table 3.5 highlights the direct ligand interactions with the target protein. It is worth noting that the hydrophobic interactions are not presented since all the transmembrane helices, the N-terminus, and ECL2 regions are involved, rendering them quite expansive. These regions are illustrated in Figure 1.5.

Table 3.5 MC4R-SHU9119 (PDB: 6W25) interaction pattern.

MC4R-SHU9119 Interactions	
<i>Salt Bridge (SB)</i>	ASP126
	GLU100
	THR101
	ASN123
<i>Hydrogen Bond(HB)</i>	SER188
	HIS264
<i>π-π Interactions (π-π)</i>	PHEN51
	TYR268

Figure 3.6 displays SHU9119 within the binding site of hMC4R, as visualized using the Maestro platform. In the figure, hydrogen bonds are depicted in yellow, salt bridges in pink, and pi-pi interactions in cyan dashed lines.

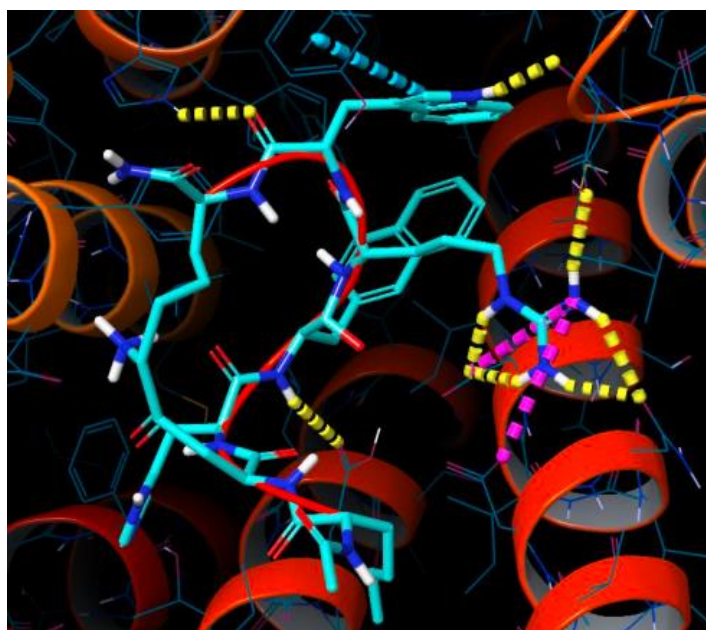


Fig. 3.6: SHU9119 in the binding site of hMC4R (PDB: 6W25), as visualized using Maestro. SBs are presented in pink color, HBs in yellow color and pi-pi interactions in cyan dashed lines.

It should be noted that MC4R-SHU9119 complex exhibits a classical seven-transmembrane helical bundle with a small orthosteric binding pocket containing SHU9119 and Ca^{2+} , a metal ion with strong electron density. The Ca^{2+} is coordinated by two main-chain carbonyl oxygen atoms in SHU9119 and three negatively charged residues in MC4R. Several studies suggest that the coordination of the Ca^{2+} plays a critical role as a cofactor in the ligand-binding process [115].

The following tables (Table 3.6-3.7) summarize the most significant findings from the molecular docking experiments.

Table 3.6 lists the Docking Scores (kcal/mol) of the chemical compounds qualified as potential ligands from the set of 2,000 natural compounds, for each software employed in this thesis. Additionally, the Docking Score of SHU9119 is presented, which was used to validate our selections. The first compound (ZINC000000487423) resulted from the first approach, which involved docking natural compounds with a molecular weight below 500 Da. The remaining compounds, which exceed this threshold, were identified through the second docking approach.

Table 3.6 Docking Scores (kcal/mol) of the final selection of chemical compounds with MC4R (PDB: 6W25).

Compounds		Docking Score (kcal/mol)		
		Glide-SP	Glide-XP	Webina
SHU9119 (validation)		-11.710	-14.580	-7.800
1	ZINC000000487423	-6.968	-5.671	-7.149
2	ZINC000004349406	-5.787	-10.771	-8.178
3	ZINC000169724085	-7.737	-11.953	-8.360
4	ZINC000299817569	-6.286	-8.875	-7.086
5	ZINC000095913431	-6.999	-10.300	-8.774
6	ZINC000095913799	-5.548	-9.382	-8.189

Figure 3.7 demonstrates the validation process, which is the optimal superimposition of the 3D crystal configurations with the docked structure of SHU9119 for Glide-SP (-11.710 kcal/mol) and Glide-XP (-14.580 kcal/mol).

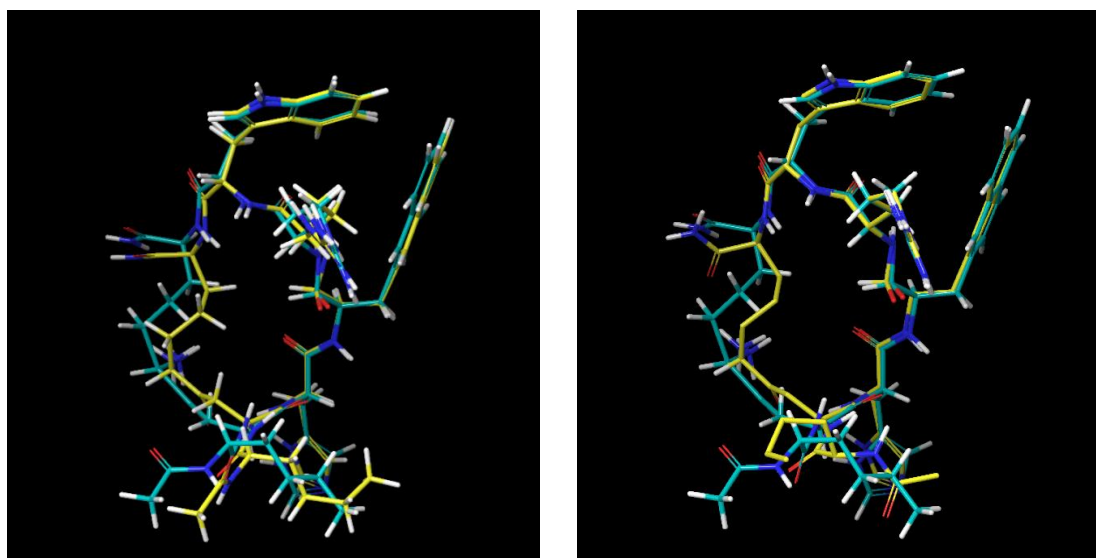


Fig. 3.7: Validation. Optimal superimposition of the 3D crystal configurations with the docked structure of SHU9119 using Glide-SP (left) and Glide-XP (right).

Table 3.7 presents the interaction patterns of the selected compounds into hMC4R as identified by each software. The common amino acids that form interactions, compared to the co-crystallized ligand are highlighted in blue. These common amino acids and the Docking Scores were used as criteria for the final selection of chemical compounds in the molecular docking results.

Table 3.7 Interaction patterns of the final selection of natural compounds into hMC4R (PDB: 6W25).

Compounds		Interaction Pattern		
		Glide-SP	Glide-XP	PLIP
SHU9119 (validation)		HB: GLU100, THR101, ASN123, SER188, HIS264 SB: ASP126 pi-pi: PHE51, TYR268		
1	ZINC000000487423	HB: PHE184, SER188	HB: ASP126 , SER188	HB: ASP126 (2) , ILE129, CYS130, SER188
2	ZINC000004349406	HB: ASP122, ASP126 (2), SER188(2) , ASP189 pi-pi: HIS264 , TYR268	HB: GLU100 , ASP122, ASP126 , PHE184, HIS264	HB: GLU100 , ASN123 , ASP126 , SER188 , ASP189(3), TYR268

Table 3.7 (Continued).

3	ZINC000169724085	HB: GLU100, ASP126(2), HIS264	HB: ASP122 (2), PHE184, HIS264, TYR268, ASN285	HB: ASN123, SER188(2), ASP189, HIS264 pi-pi: PHE284
4	ZINC000299817569	HB: ASP122, ASP126, PHE184, SER188, LEU288 pi-pi: HIS264	HB: ASP122, ASN123, PHE184, HIS264, ASN285	HB: GLN43, ASN123, ASP126, PHE184, SER188(2), TYR268, ASN285 pi-pi: TYR268
5	ZINC000095913431	HB: ASP122, SER188, ASN123, PHE184, HIS261, ASN285	HB: GLU100, ASP122(3), HIS264 pi-pi: TYR268	HB: GLU100, ASN123, ILE129, SER188(3), HIS264, TYR268 pi-pi: PHE284
6	ZINC000095913799	HB: GLU100, ASP122, ASN123, ASP126, SER188, ASP189, HIS264	HB: GLU100, ASP122, ASP126(2), SER188, ASP189, TYR268	HB: GLU100, ASN123, ASP126, ILE129, SER188, ASP189(3), TYR268

To provide clarity, the subsequent figures (Figure 3.8-3.13) display the chemical structures of the final natural compounds and their respective ZINC library IDs.

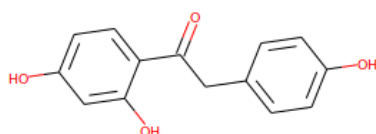


Fig. 3.8: ZINC000000487423.

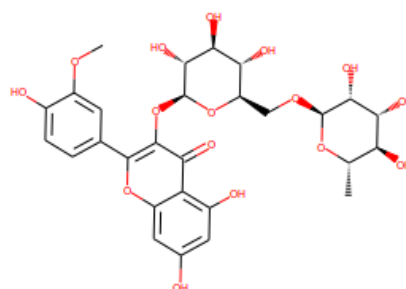


Fig. 3.9: ZINC0000004349406.

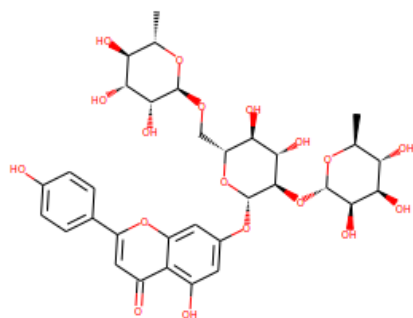


Fig. 3.10: ZINC000169724085.

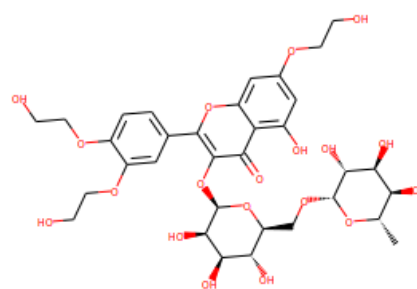


Fig. 3.11: ZINC000299817569.

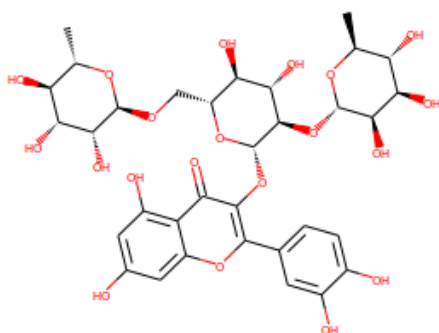


Fig. 3.12: ZINC000095913431.

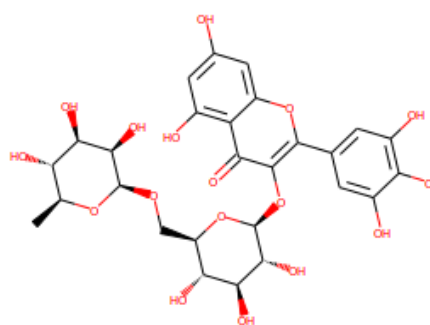


Fig. 3.13: ZINC000095913799.

As indicated by the previous figures, five out of the six natural compounds identified as potential ligands of hMC4R share a common scaffold that belongs to a specific category of polyphenolic compounds. These compounds are classified as flavonoids, specifically flavones. Flavonoids derive from the secondary metabolism of plants, typically vascular plants and some mosses. The basic skeleton of flavones is distinguished by a non-saturated 3-C chain and a double bond between C-2 and C-3 [116].

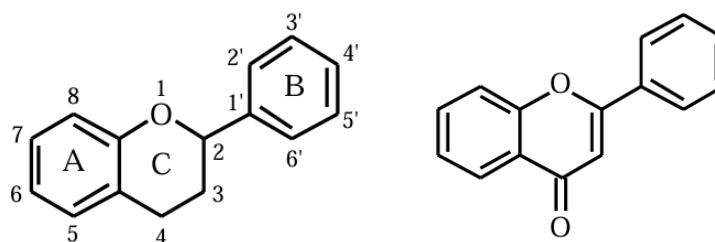


Fig. 3.14: Basic structure and numbering system of flavonoids (left) and flavones (right) [116].

3.3.1 Protein-ligand binding pose prediction using PLIP

The following figures (Figure 3.15-3.20) provide visual representations of 3D binding poses of the final molecules with the target protein, as generated using the

PLIP open-access platform. The yellow color corresponds to the potential ligand, whereas the blue formations belong to the target protein.

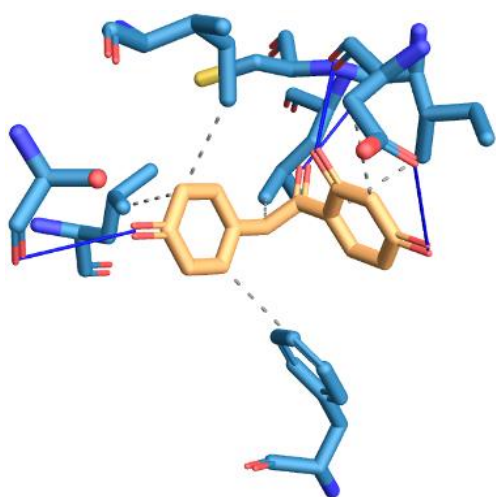


Fig. 3.15: Representative 3D binding pose of ZINC000000487423 at MC4R binding site (PDB: 6W25) using PLIP.

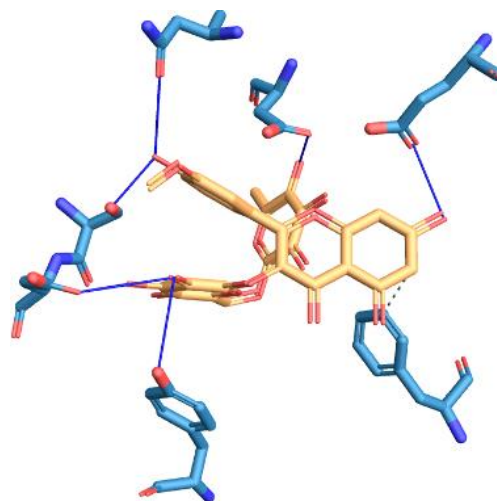


Fig. 3.16: Representative 3D binding pose of ZINC000004349406 at MC4R binding site (PDB: 6W25) using PLIP.

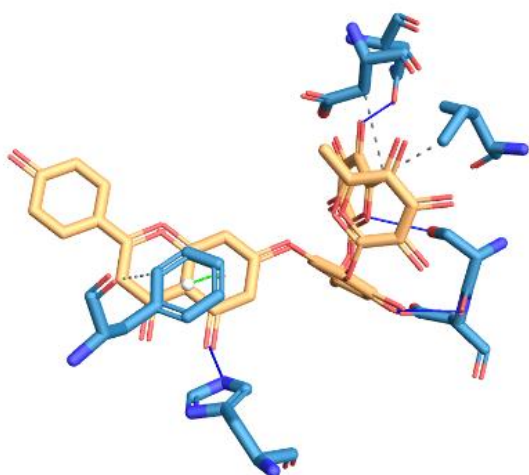


Fig. 3.17: Representative 3D binding pose of ZINC000169724085 at MC4R binding site (PDB: 6W25) using PLIP.

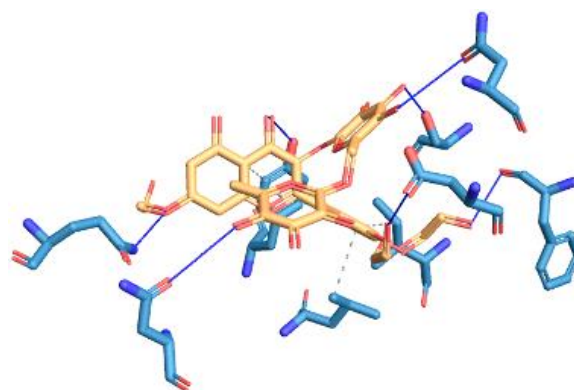


Fig. 3.18: Representative 3D binding pose of ZINC000299817569 at MC4R binding site (PDB: 6W25) using PLIP.

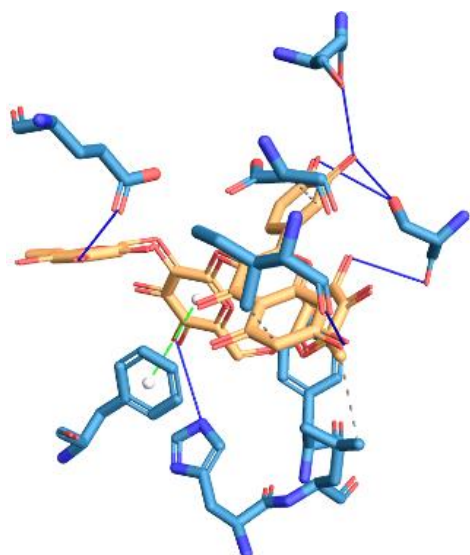


Fig. 3.19: Representative 3D binding pose of ZINC000095913431 at MC4R binding site (PDB: 6W25) using PLIP.

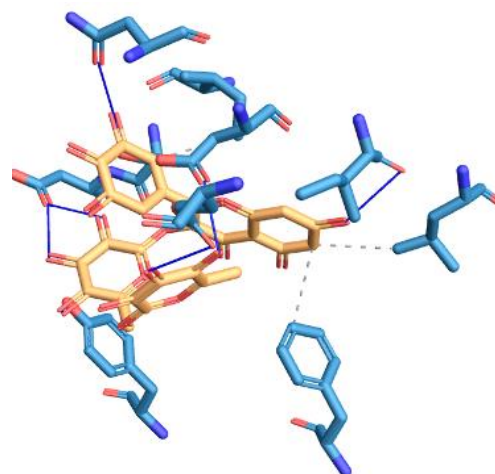
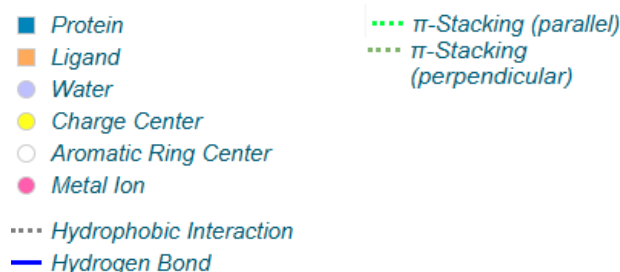


Fig. 3.20: Representative 3D binding pose of ZINC000095913799 at MC4R binding site (PDB: 6W25) using PLIP



3.3.2 Protein-ligand binding pose prediction using Maestro

The following figures (3.21-3.26) provide visual representations of the 2D Ligand Interaction Diagram, and the 3D binding poses of the final selection with the target protein, as generated using the Maestro platform. These figures provide information about how the ligand is interacting with the protein of interest.

When examining the Ligand Interaction Diagram, we can analyze the interactions between the ligand and the surrounding residues within the binding pocket. The protein residues are depicted as being connected along a black line, with varying orientations. If a residue points away from the ligand, it indicates that the backbone of the residue is facing the ligand. Conversely, if a residue points toward the ligand, it signifies that the side chain of that residue is oriented toward the ligand. Finally, regions shaded in grey represent solvent-exposed areas.

For the 3D interaction poses, as previously noted, the yellow color corresponds to the hydrogen bonds and the cyan to the pi-pi interactions.

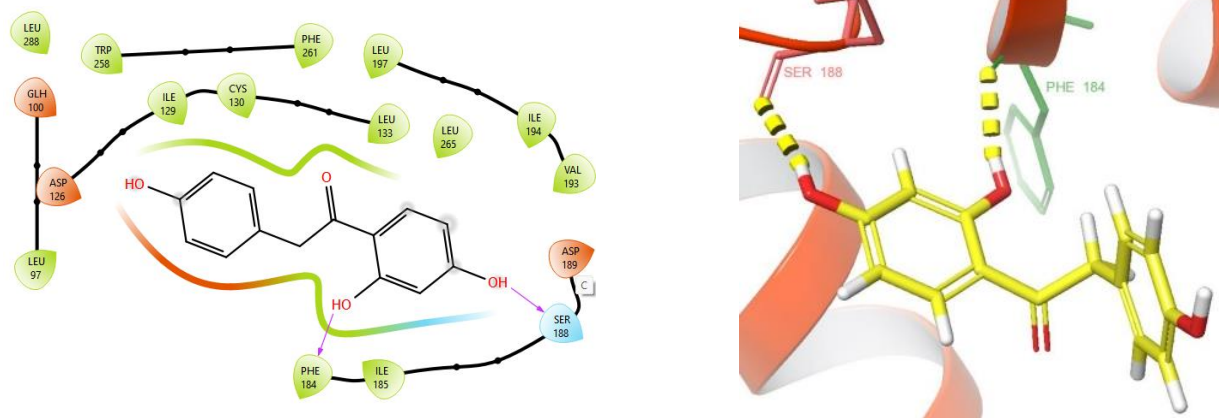


Fig 3.21: Representative 2D Ligand Interaction Diagram (left) and 3D binding pose (right) of ZINC000000487423 into the hMC4R binding site (PDB: 6W25) using Maestro. HBs are represented in yellow dashed lines.

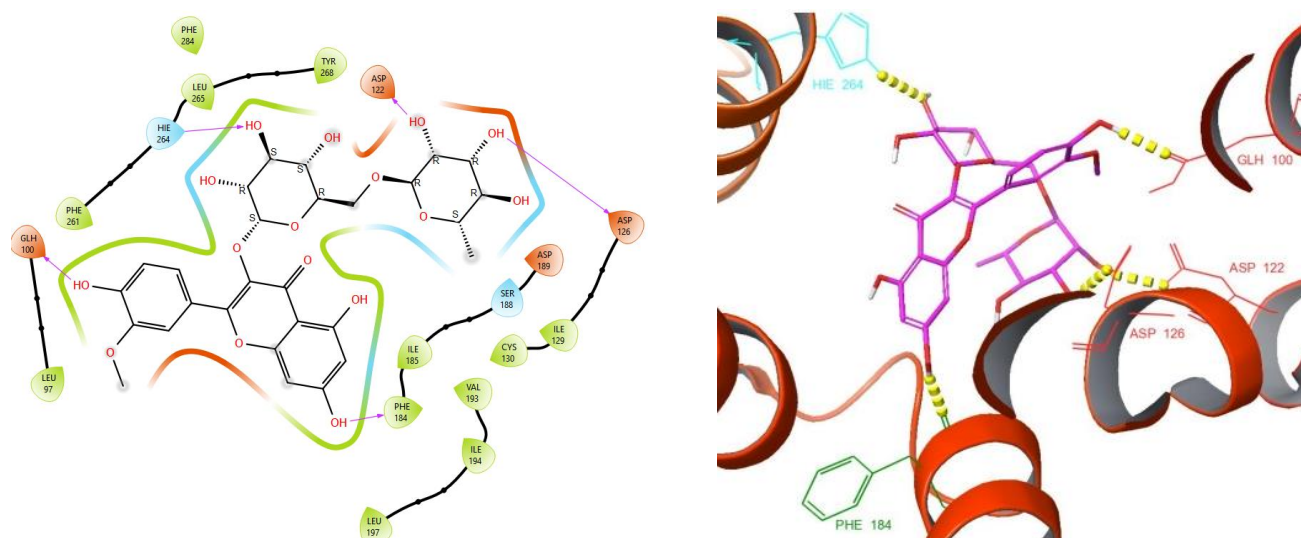


Fig. 3.22: Representative 2D Ligand Interaction Diagram (left) and 3D binding pose (right) of ZINC0000004349406 into the hMC4R binding site (PDB: 6W25) using Maestro. HBs are represented in yellow dashed lines.

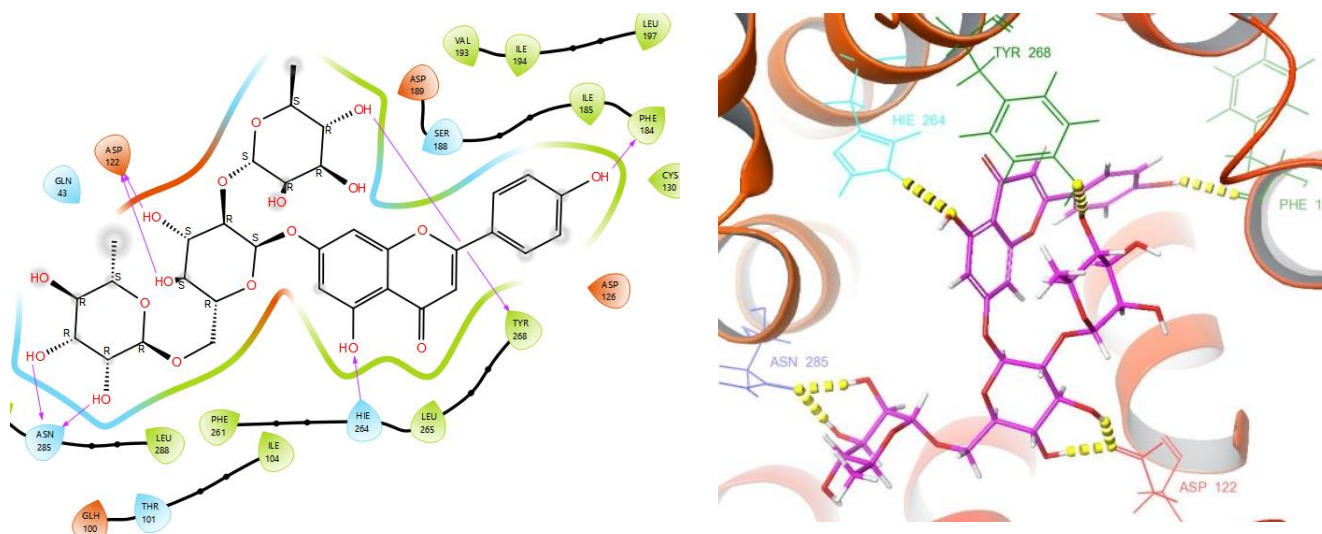


Fig. 3.23: Representative 2D Ligand Interaction Diagram (left) and 3D binding pose (right) of ZINC000169724085 into the hMC4R binding site (PDB: 6W25) using Maestro. HBs are represented in yellow dashed lines.

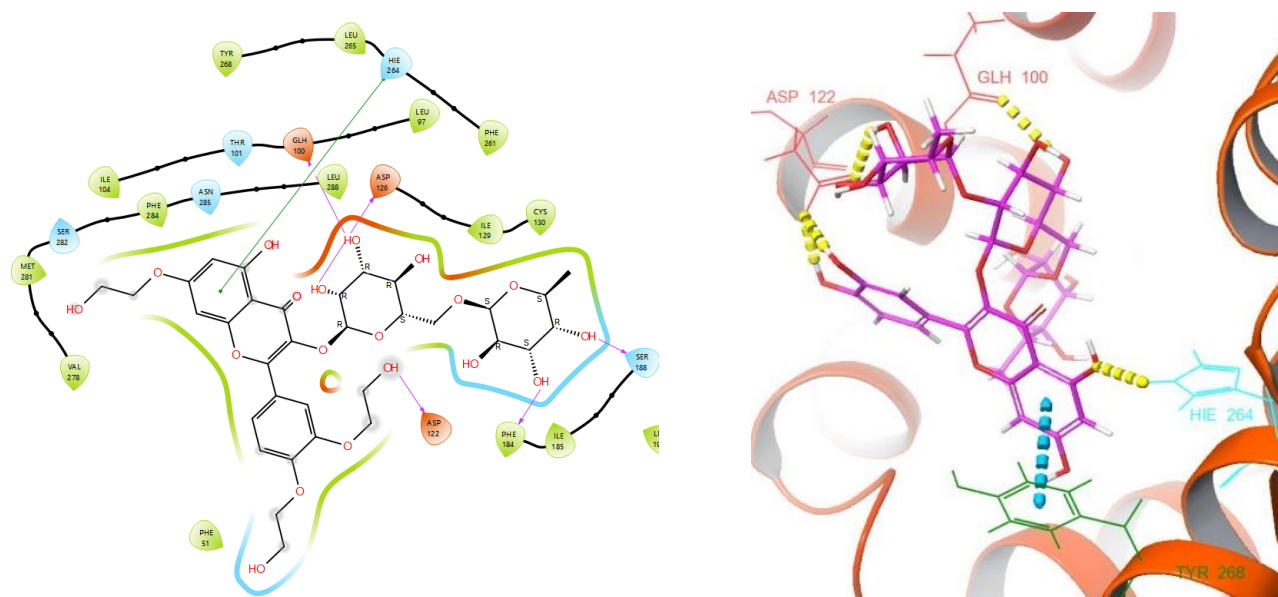


Fig. 3.24: Representative 2D Ligand Interaction Diagram (left) and 3D binding pose (right) of ZINC000299817569 into the hMC4R binding site (PDB: 6W25) using Maestro. HBs are represented in yellow and pi-pi interactions in cyan dashed lines.

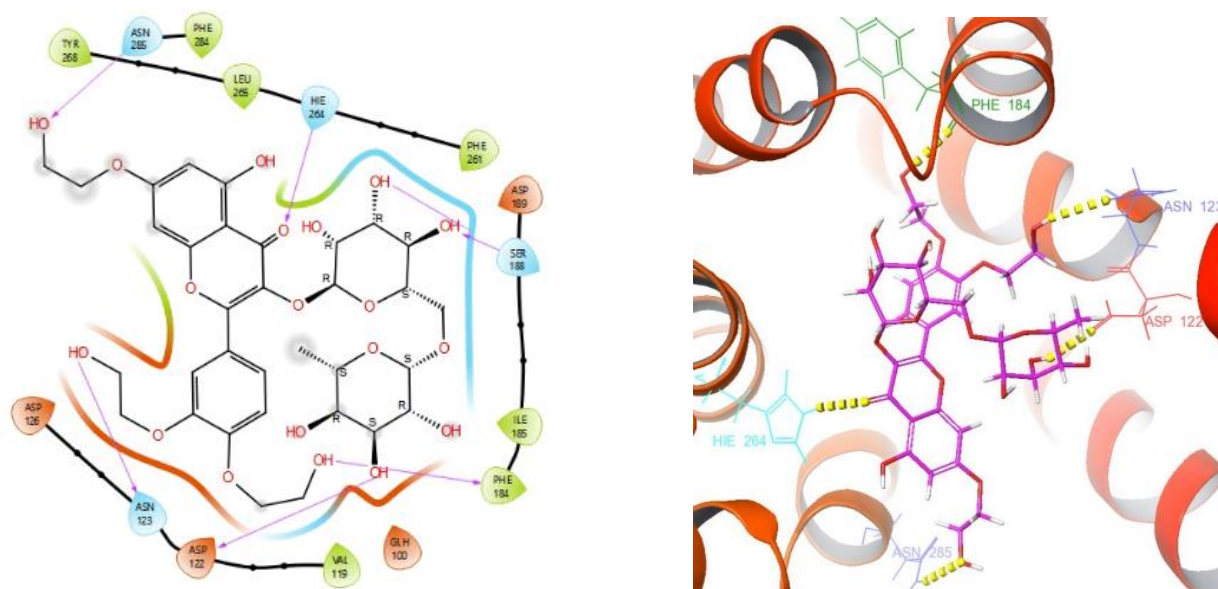


Fig. 3.25: Representative 2D Ligand Interaction Diagram (left) and 3D binding pose (right) of ZINC000095913431 into the hMC4R binding site (PDB: 6W25) using Maestro. HBs are represented in yellow dashed lines.

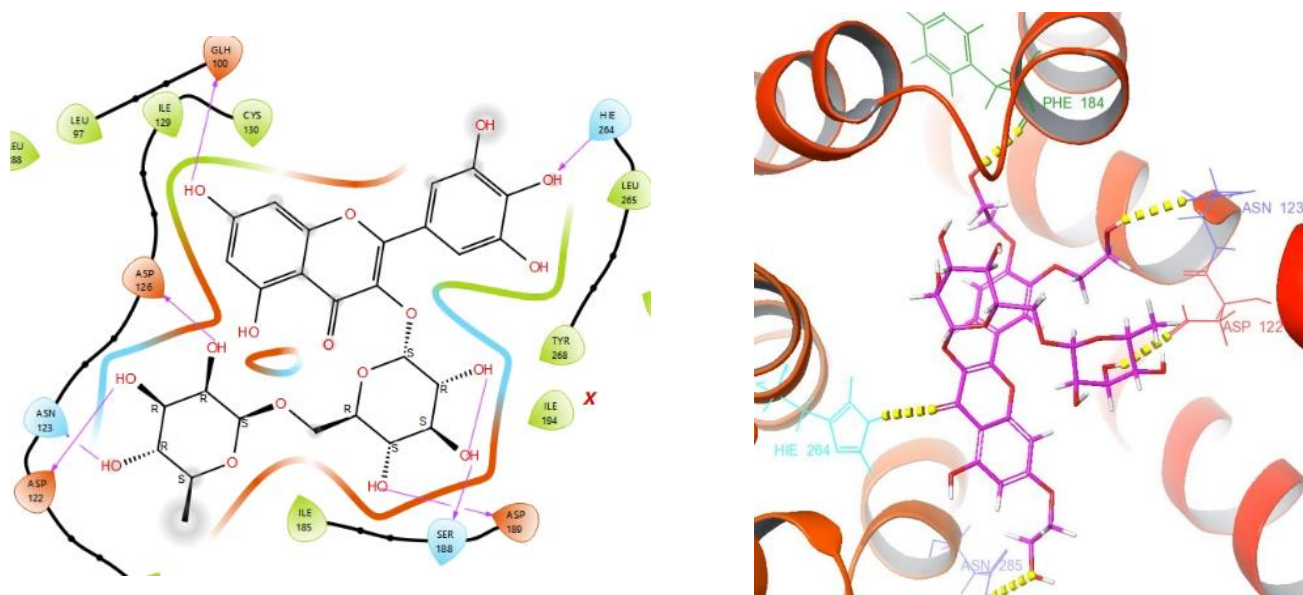
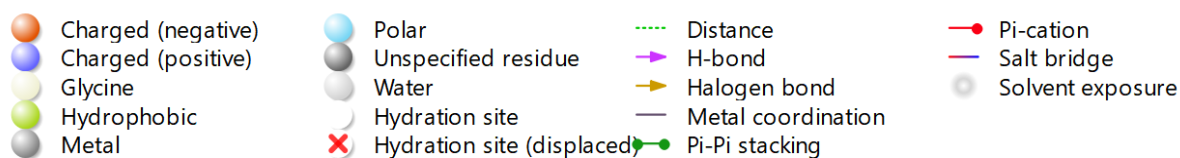


Fig. 3.26: Representative 2D Ligand Interaction Diagram (left) and 3D binding pose (right) of ZINC000095913799 into the hMC4R binding site (PDB: 6W25) using Maestro. HBs are represented in yellow dashed lines.



3.4 MetaboAnalyst Results

The figures presented (Figure 3.27-3.28) below are obtained through the MetaboAnalyst 6.0 open-access software. As previously noted, we manually selected the combination of features listed in Table 3.1 to create the biomarker models using the Random Forest classifier.

Figure 3.27 illustrates the ROC curve for 100 cross-validations. The results were averaged to generate the plot.

The first diagram in Figure 3.28 shows the average of predicted class probabilities of each sample across the 100 cross-validations. The algorithm uses a balanced sub-sampling approach, the classification boundary is located at the center ($x = 0.5$, the dotted line). Additionally, the box plot provides a visual representation of the predictive accuracy. The average accuracy based on 100 cross validations is 0.936.

Table 3.8 Prediction of final natural compounds.

Compounds	Probability	Class
ZINC000000487423	0.99	Not Active
ZINC0000004349406	0.69	Not Active
ZINC000169724085	0.83	Active
ZINC000299817569	0.71	Active
ZINC000095913431	0.76	Active
ZINC000095913799	0.77	Not Active

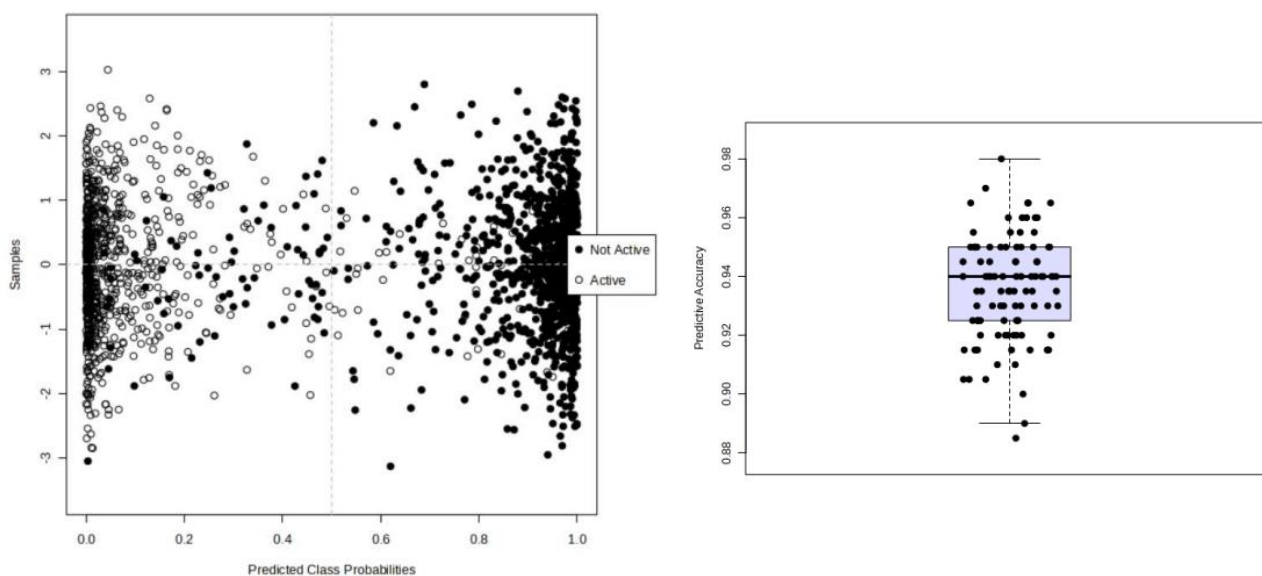


Fig. 3.27: Predicted class probabilities (left) and box plot of predictive accuracy (right) using MetaboAnalyst.

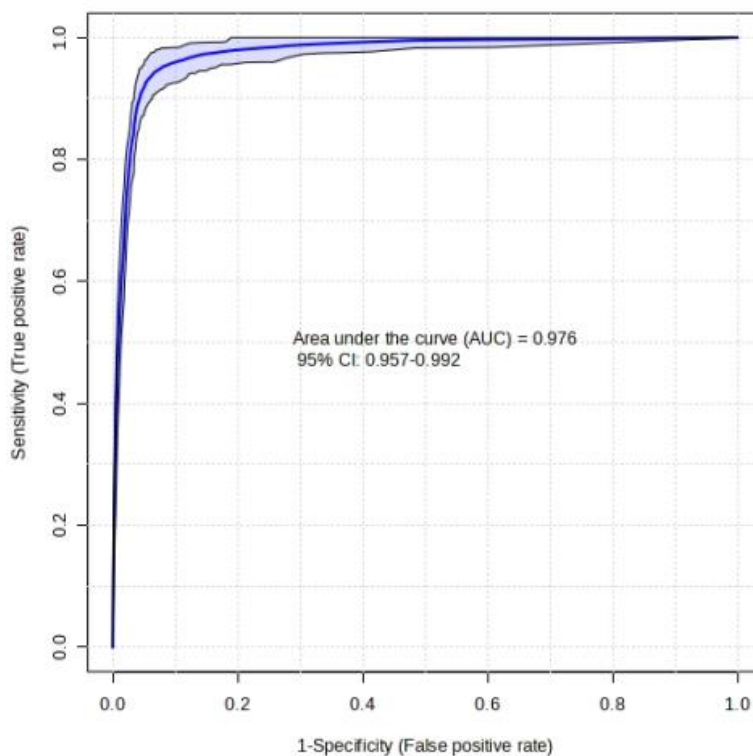


Fig. 3.28: ROC curve for 100 cross-validations using MetaboAnalyst.

3.5 ADMET Results

The following tables (Table 3.9-3.12) display the ADMET prediction results obtained from both tools for the natural compounds identified by our model as Active against MC4R. We have focused on key properties that we deemed significant for evaluating the chemical compound's characteristics. It is evident from our results that these tools employ different underlying methods to make their predictions.

Table 3.9 illustrates the predictions for Setmelanotide, while the subsequent tables (Table 3.10-3.12) detail the predictions for the compounds we have evaluated.

The Human Intestinal Absorption (HIA) is considered as a basic requirement for a molecule's effectiveness, since it is considered as an indicator of oral bioavailability. A molecule with an absorbance below 0.3 is regarded as having poor absorption.

The BBB Penetration indicates a molecule's ability to cross the BBB. The output presented in the tables corresponds to the probability of a chemical compound crossing this barrier and accumulating in the brain within the range of 0 and 1 [117].

Cytochromes P450 or CYPs are isozymes in the liver, which serve as major drug-metabolizing enzymes. We selected to present the CYP3A4 enzyme, since it's the key enzyme responsible for Phase I metabolism [118]. The output presented is the probability of a chemical compound to be considered an inhibitor or substrate within the range of 0 and 1.

The drug's half-life ($T_{1/2}$) is an alternative measure of clearance, corresponding to the reduction of its active substance by 50%.

Finally, Drug Induced Liver Injury (DILI) is another measure to assess hepatotoxicity. The output listed in the tables is the probability of a molecule to be DILI-positive within the range of 0 and 1 [117].

Table 3.9 *Setmelanotide ADMET prediction.*

SETMELANOTIDE		
	ADMETlab	ADMET-AI
Physicochemical Properties		
MW (Da)	1116.49	1117.33
LogP	-0.535	-3.18
Absorption		
HIA	0-0.1	0.51
Distribution		
BBB Penetration (cm/s)	0-0.1	0.11
Metabolism		
CYP3A4 Inhibitor	0-0.1	0.62
CYP3A4 Substrate	0.9-1.0	0.78
Excretion		
T _{1/2} (hr)	3.083	33.63
Toxicity		
DILI	0.942	0.48

Table 3.10 *ZINC000169724085 ADMET prediction.*

ZINC000169724085		
	ADMETlab	ADMET-AI
Physicochemical Properties		
MW (Da)	724.22	724.67
LogP	0.938	-2.25
Absorption		
HIA	0.7-0.9	0.05
Distribution		
BBB Penetration (cm/s)	0-0.1	0.08
Metabolism		
CYP3A4 Inhibitor	0-0.1	0.01
CYP3A4 Substrate	0-0.1	0.02
Excretion		
T _{1/2} (hr)	4.708	40.97
Toxicity		
DILI	0.972	0.57

Table 3.11 ZINC000299817569 ADMET prediction.

ZINC000299817569		
	ADMETlab	ADMET-AI
Physicochemical Properties		
MW (Da)	742.23	742.28
LogP	-0.072	-2.69
Absorption		
HIA	0.9-1.0	0.08
Distribution		
BBB Penetration (cm/s)	0-0.1	0.05
Metabolism		
CYP3A4 Inhibitor	0-0.1	0.03
CYP3A4 Substrate	0-0.1	0.45
Excretion		
T _{1/2} (hr)	4.055	58.84
Toxicity		
DILI	0.867	0.70

Table 3.12 ZINC000095913431 ADMET prediction.

ZINC000095913431		
	ADMETlab	ADMET-AI
Physicochemical Properties		
MW (Da)	756.21	756.66
LogP	0.83	-2.84
Absorption		
HIA	0.5-0.7	0.05
Distribution		
BBB Penetration (cm/s)	0-0.1	0.06
Metabolism		
CYP3A4 Inhibitor	0-0.1	0.01
CYP3A4 Substrate	0-0.1	0.44
Excretion		
T _{1/2} (hr)	5.31	46.99
Toxicity		
DILI	0.908	0.68

There are slight differences in the outputs of each computational tool. ADMETlab provides its predictions as probability ranges, while ADMET-AI presents them as distinct values. While MW is practically equivalent, the same cannot be said for LogP. In the case of HIA, the predictions are opposite. The BBB penetration values predicted by both tools are approximately equal. According to the ADMETlab documentation, these values are considered empirically “excellent”. The metabolism of the final compounds, when compared with Setmelanotide, is suboptimal. The half-life predictions in ADMET-AI are approximately 10 times greater than those from ADMETlab. According to the results, both Setmelanotide and the natural compounds are classified as high DILI-positive, indicating a significant risk of liver injury.

Chapter 4: Discussion

Obesity remains one of the most pressing global health challenges, despite extensive efforts to address the issue. In this study, we focused on the severe type of monogenic obesity and its association with MC4R, which plays a crucial role in its appearance during early childhood. Our goal was to combine machine learning models with molecular modeling techniques, focusing on molecular docking, to identify potential ligands for MC4R that may act as agonists to regulate appetite. By employing this hybrid approach, we aimed to discover chemical compound scaffolds that could be further investigated *in vitro*, contributing to the drug design process.

The first significant findings emerged from the application of machine learning techniques and the development of predictive models that estimate the activity of chemical compounds against the protein target of interest. After thorough investigation, we successfully developed a model with effective performance, resulting from the combination of 7 molecular descriptors out of the 208 available of the RDKit Library: **VSA_EState6**, **MaxAbsEstateIndex**, **PEOE_VSA8**, **Kappa2**, **MolMR**, **BCUT2D_MRLOW**, and **Kappa3**. Understanding the chemical and biological significance of these molecular descriptors is essential for interpreting our results. Our model considers the MW of molecules, their chemical structure and attributes associated with atom polarity and steric accessibility.

We developed a model that achieved an impressive accuracy of 92.71% with a standard deviation of 1.07% over 10 epochs, using the Random Forest Classifier. It demonstrated a precision of 0.92, with a standard deviation of 0.01, and a high F1-score, indicating reliable positive predictions and minimal false positives. It needs to be noted that in the context of this thesis, the "Active" chemical compounds are considered the positive class, while the "Not Active" compounds are the negative class. Additionally, the model yielded an AUC of 0.98 with no standard deviation, highlighting its exceptional ability to distinguish between the two classes for these specific molecular descriptors.

The subsequent statistical analysis using the Mann-Whitney U-test indicates that most of the molecular descriptors in our model show statistically significant differences between the two classes. Additionally, by examining each feature with ROC curve analysis, we visually assessed the separability of molecular descriptors between classes. This separability is further illustrated by the box plots we present. Although there are some outliers, the boxes for the two classes do not overlap, and the median values are distinctly higher for one class compared to the other. We determined the descriptors with a $p\text{-value} \leq 0.001$ and an $\text{AUC} > 0.8$. **MaxAbsEstateIndex**, **PEOE_VSA8**, **Kappa2**, and **BCUT2D_MRLOW**, exhibit both statistically significant differences and strong separability between "Active" and "Not Active" chemical compounds.

In parallel with the machine learning analysis, we performed molecular docking studies to investigate the interaction patterns and binding affinities of chemical compounds as potential MC4R ligands. We selected six compounds for further examination based on docking simulations conducted with three different computational tools. Initially, we focused on hydrogen bonds, since they indicate strong binding, and subsequently on potential pi-pi interactions. We also evaluated whether the compounds formed bonds with amino acids common to SHU9119.

Our first approach of molecular docking on natural compounds with a MW below 500 Da did not yield promising results, with only one compound

(ZINC000000487423) falling into this category. However, our second approach, which involved molecules with MW greater than 500 Da, demonstrated more potential.

We noticed that the molecules that interact best with the receptor are flavonoids, particularly flavones. This category of compounds is abundant in plants, fruits and vegetables. Recent research has explored the role of flavonoids in appetite regulation, metabolic enhancement, and overall obesity management. These findings emerged from the compelling need to search for natural and safe alternatives of commonly used anti-obesity medications that do not cause side effects such as high blood pressure, heart palpitations, or depression. In particular, flavones have been studied for their antioxidant and anti-inflammatory properties, and it has also been proven that they can influence the metabolism of adipose tissue. A number of *in vitro* and *in vivo* studies indicate that a diet high in flavones reduces visceral adiposity by inhibiting adipogenesis. Furthermore, these studies suggest that flavones could be effective as dietary supplements for modulating the feeling of satiety, in general. Nevertheless, clear evidence is still lacking regarding the effectiveness of flavones in regulating obesity related to CNS gene mutations. Our findings suggest that flavones may act as regulators for MC4R [119], [120].

However, it is important to note that these natural compounds we identified from the ZINC library contain glucosinolate components, which may undergo hydrolysis. As shown in Figures 3.22 to 3.26, these compounds are exposed to solvents, which could facilitate this process. Nevertheless, a significant number of hydrogen bonds are formed with the flavone portion of the molecules. This suggests that even if the glucosinolate component is hydrolyzed, the flavones will remain intact, maintaining their functional interactions.

Taken together, the machine learning model was used to predict the activity of these six chemical compounds against MC4R. Our model predicted that only three of the six final compounds are active against the obesity-associated receptor. These compounds, ranked in descending order of predicted activity, are **ZINC000169724085** with a probability of 0.83, **ZINC000095913431** with a probability of 0.76, and **ZINC000299817569** with a probability of 0.71.

To further assess the ADMET properties of these molecules, we utilized two different open-access software platforms. We compared their predicted characteristics to those of the FDA-approved drug Setmelanotide, offering a benchmark for evaluating the potential of these compounds. However, upon examining the results in Tables 3.9 to 3.12, we identified some limitations. Notably, there are significant discrepancies in the probability predictions for certain properties between the two tools. In particular, these differences are evident in **HIA**, **CYP3A4 Substrate**, and **T_{1/2}**. Additionally, these computational tools failed to accurately predict the T_{1/2} of Setmelanotide to a high degree, which is known to be approximately 11 hours [121]. We believe that this discrepancy may be attributed to differences in the neural network frameworks employed by each platform, as well as the distinct datasets used during their training. This suggests that we cannot derive fully accurate results from this computational experiment.

In summary, this thesis contributes valuable insights to the potential role of flavones as ligands of MC4R to regulate the disease of obesity as a novel therapeutic option. In addition, it underscores the necessity to explore further the genetic

influences and predisposition that contribute to the complexity of obesity as a multifactorial condition.

4.1 Future Steps

To further evaluate our findings and better understand how flavones impact obesity, future steps should include:

- i. ***Molecular Dynamics (MDs) Simulations.*** MDs simulations will provide a detailed, atomic-level understanding of the receptor's behavior and reveal how flavones interact with it at a molecular level.
- ii. ***In vitro experiments.*** Laboratory experiments will further validate the effects of flavones on biological processes, such as adipose tissue formation.
- iii. ***Molecular Docking on peptides.*** This approach will allow us to explore a different library of compounds and investigate their mechanisms, focusing on those with structural similarities to the native ligands of hMC4R.

These steps will help clarify the potential and limitations of flavones in obesity regulation and guide the development of targeted therapeutic strategies.

Bibliography

- [1] N. S. Mitchell, V. A. Catenacci, H. R. Wyatt, and J. O. Hill, "Obesity: Overview of an epidemic," *Psychiatric Clinics of North America*, vol. 34, no. 4, pp. 717–732, Dec. 2011. doi:10.1016/j.psc.2011.08.005
- [2] "Obesity and overweight," World Health Organization, <https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight> (accessed Aug. 3, 2024).
- [3] "The challenge of obesity," World Health Organization, <https://www.who.int/europe/news-room/fact-sheets/item/the-challenge-of-obesity> (accessed Aug. 3, 2024).
- [4] H. C. Kadouh and A. Acosta, "Current paradigms in the etiology of obesity," *Techniques in Gastrointestinal Endoscopy*, vol. 19, no. 1, pp. 2–11, Jan. 2017. doi:10.1016/j.tgie.2016.12.001
- [5] T. Rankinen *et al.*, "The human obesity gene map: The 2005 update," *Obesity*, vol. 14, no. 4, pp. 529–644, Apr. 2006. doi:10.1038/oby.2006.71
- [6] K. R. Rao, N. Lal, and N. V. Giridharan, "Genetic & epigenetic approach to human obesity," *Indian J Med Res.*, vol. 140, no. 5, pp. 589–603, Nov. 2014. PMID: 25579139 PMCID: PMC4311311
- [7] C. T. Montague *et al.*, "Congenital leptin deficiency is associated with severe early-onset obesity in humans," *Nature*, vol. 387, no. 6636, pp. 903–908, Jun. 1997. doi:10.1038/43185
- [8] H. Huvenne, B. Dubern, K. Clément, and C. Poitou, "Rare genetic forms of obesity: Clinical approach and current treatments in 2016," *Obesity Facts*, vol. 9, no. 3, pp. 158–173, 2016. doi:10.1159/000445061
- [9] J. L. Tymoczko, J. M. Berg, and L. Stryer, *Biochemistry: A Short Course*, 3rd ed. New York, NY: W.H. Freeman & Company, 2015.
- [10] "Amino acids which do not have any charge on them are neutral amino acids.," bartleby, <https://www.bartleby.com/subject/science/chemistry/concepts/neutral-amino-acids> (accessed Aug. 3, 2024).
- [11] M. E. Sahin, T. Can, and C. D. Son, "GPCRsorT—responding to the next generation Sequencing Data Challenge: Prediction of G protein-coupled receptor classes using only structural region lengths," *OMICS: A Journal of Integrative Biology*, vol. 18, no. 10, pp. 636–644, Oct. 2014. doi:10.1089/omi.2014.0073
- [12] J. V. Zhang, L. Li, Q. Huang, and P.-G. Ren, "Obestatin receptor in energy homeostasis and obesity pathogenesis," *Progress in Molecular Biology and Translational Science*, pp. 89–107, 2013. doi:10.1016/b978-0-12-386933-3.00003-0
- [13] V. J., C. L., and van A., "Melanocortins and their receptors and antagonists," *Current Drug Targets*, vol. 4, no. 7, pp. 586–597, Oct. 2003. doi:10.2174/1389450033490858
- [14] I. Gantz and T. M. Fong, "The melanocortin system," *American Journal of Physiology-Endocrinology and Metabolism*, vol. 284, no. 3, Mar. 2003. doi:10.1152/ajpendo.00434.2002
- [15] A. I. Bima *et al.*, "Molecular profiling of melanocortin 4 receptor variants and agouti-related peptide interactions in morbid obese phenotype: A novel paradigm from molecular docking and Dynamics Simulations," *Biologia*, vol. 77, no. 5, pp. 1481–1496, Feb. 2022. doi:10.1007/s11756-022-01037-3

- [16] J. P. Gonçalves, D. Palmer, and M. Meldal, “MC4R agonists: Structural overview on Antiobesity Therapeutics,” *Trends in Pharmacological Sciences*, vol. 39, no. 4, pp. 402–423, Apr. 2018. doi:10.1016/j.tips.2018.01.004
- [17] T. Kalanathan *et al.*, “The melanocortin system in Atlantic Salmon (*Salmo salar* L.) and its role in appetite control,” *Frontiers in Neuroanatomy*, vol. 14, Aug. 2020. doi:10.3389/fnana.2020.00048
- [18] B.-X. Chai *et al.*, “Receptor–antagonist interactions in the complexes of agouti and agouti-related protein with human melanocortin 1 and 4 receptors,” *Biochemistry*, vol. 44, no. 9, pp. 3418–3431, Feb. 2005. doi:10.1021/bi0478704
- [19] A. Jais and J. C. Brüning, “Arcuate nucleus-dependent regulation of metabolism—pathways to obesity and diabetes mellitus,” *Endocrine Reviews*, vol. 43, no. 2, pp. 314–328, Sep. 2021. doi:10.1210/endrev/bnab025
- [20] D. Huszar *et al.*, “Targeted disruption of the melanocortin-4 receptor results in obesity in mice,” *Cell*, vol. 88, no. 1, pp. 131–141, Jan. 1997. doi:10.1016/s0092-8674(00)81865-6
- [21] M. M. Hammad *et al.*, “Structural analysis of setmelanotide binding to MC4R variants in comparison to wild-type receptor,” *Life Sciences*, vol. 307, p. 120857, Oct. 2022. doi:10.1016/j.lfs.2022.120857
- [22] N. A. Heyder *et al.*, “Structures of active melanocortin-4 receptor–GS-protein complexes with NDP- α -MSH and setmelanotide,” *Cell Research*, vol. 31, no. 11, pp. 1176–1189, Sep. 2021. doi:10.1038/s41422-021-00569-8
- [23] L. Wang, N. Wang, Z. Yan, Z. Huang, and C. Fu, “Peptide and peptide-based drugs,” *Privileged Scaffolds in Drug Discovery*, pp. 795–815, 2023. doi:10.1016/b978-0-443-18611-0.00015-2
- [24] S. M. Jayatilake and G. U. Ganegoda, “Involvement of machine learning tools in Healthcare Decision making,” *Journal of Healthcare Engineering*, vol. 2021, pp. 1–20, Jan. 2021. doi:10.1155/2021/6679512
- [25] J. Grus, *Data Science From Scratch, 2nd Edition*. O’Reilly Media, Inc, 2019.
- [26] S. Raschka and V. Mirjalili, *Python Machine Learning: Machine Learning and Deep Learning with Python, Scikit-Learn, and Tensorflow 2, 3rd Edition*. Birmingham: Packt Publishing, Limited, 2019.
- [27] R. Y. Choi, A. S. Coyner, J. Kalpathy-Cramer, M. F. Chiang, and J. P. Campbell, “Introduction to Machine Learning, Neural Networks, and Deep Learning,” *Translational Vision Science & Technology*, vol. 9, no. 2, Feb. 2020. doi:https://doi.org/10.1167/tvst.9.2.14
- [28] R. C. Deo, “Machine learning in medicine,” *Circulation*, vol. 132, no. 20, pp. 1920–1930, Nov. 2015. doi:10.1161/circulationaha.115.001593
- [29] I. H. Sarker, “Machine learning: Algorithms, real-world applications and Research Directions,” *SN Computer Science*, vol. 2, no. 3, Mar. 2021. doi:10.1007/s42979-021-00592-x
- [30] “Reinforcement learning - simply explained,” Data Basecamp, <https://databasecamp.de/en/ml/reinforcement-learnings> (accessed Aug. 17, 2024).
- [31] Δ. Γκλώτσος (2023) “8. Μηχανική Μάθηση (Machine Learning),” in *Συστήματα Υποστήριξης Απόφασης*.
- [32] Δ. Κάβουρας (2023) “MM_Διάλεξη_02_Python_2023_24,” in *Μηχανική Μάθηση σε Python*.

- [33] D. Gong, “Top 6 machine learning algorithms for classification,” Medium, <https://towardsdatascience.com/top-machine-learning-algorithms-for-classification-2197870ff501> (accessed Aug. 17, 2024).
- [34] C. Albon, *Machine Learning with Python Cookbook: Practical Solutions from Preprocessing to Deep Learning*. Sebastopol (CA): O’Reilly, 2020.
- [35] “1.9. naive Bayes,” scikit, https://scikit-learn.org/stable/modules/naive_bayes.html (accessed Aug. 17, 2024).
- [36] Δ. Κάβουρας (2023) “MM_Διάλεξη_03_Python_2023_24.,” in *Μηχανική Μάθηση σε Python*.
- [37] A. Famili, W.-M. Shen, R. Weber, and E. Simoudis, “Data preprocessing and Intelligent Data Analysis,” *Intelligent Data Analysis*, vol. 1, no. 1, pp. 3–23, Jan. 1997. doi:10.3233/ida-1997-1102
- [38] S. Khalid, T. Khalil, and S. Nasreen, “A survey of feature selection and feature extraction techniques in machine learning,” *2014 Science and Information Conference*, Aug. 2014. doi:10.1109/sai.2014.6918213
- [39] Q. Xu, M. Kamel, and M. M. Salama, “Significance test for feature subset selection on image recognition,” *Lecture Notes in Computer Science*, pp. 244–252, 2004. doi:10.1007/978-3-540-30125-7_31
- [40] “Πιθανότητες,” in *Πιθανότητες και Στατιστική για Μηχανικούς*, Αθήνα: ΤΖΙΟΛΑ, 2019
- [41] D. Sarkar, R. Bali, and T. Sharma, *Practical machine learning with python*, pp. 260–261, 2018. doi:10.1007/978-1-4842-3207-1
- [42] Δ. Κάβουρας (2023) “MM_Διάλεξη_04_Python_2023,” in *Μηχανική Μάθηση σε Python*.
- [43] D. M. W. Powers, “Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation,” Internet Archive Wayback Machine, https://web.archive.org/web/20191114213255/https://www.flinders.edu.au/science_engineering/fms/School-CSEM/publications/tech_reps-research_artfcts/TRRA_2007.pdf (accessed Aug. 17, 2024).
- [44] “F1_score,” scikit, https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html (accessed Aug. 17, 2024).
- [45] Unpingco, *Python for Probability, Statistics, and Machine Learning*. Springer International Publishing, 2019.
- [46] F. S. Nahm, “Receiver operating characteristic curve: Overview and practical use for clinicians,” *Korean Journal of Anesthesiology*, vol. 75, no. 1, pp. 25–36, Feb. 2022. doi:10.4097/kja.21209
- [47] “Box and Whisker Plots,” Numeracy, Maths and statistics - academic skills kit, <https://www.ncl.ac.uk/webtemplate/ask-assets/external/maths-resources/statistics/data-presentation/box-and-whisker-plots.html> (accessed Aug. 17, 2024).
- [48] K. Hu, “Become competent within one day in generating Boxplots and violin plots for a novice without prior R experience,” *Methods and Protocols*, vol. 3, no. 4, p. 64, Sep. 2020. doi:10.3390/mps3040064
- [49] Μαυρομούστακος, Θ., Χοντζοπούλου, Ε., Κυριακίδη, Σ., & Ζουμπουλάκης, Π. (2023). Αρχές Υπολογιστικής Χημείας [Μεταπτυχιακό εγχειρίδιο]. Κάλλιπος, Ανοικτές Ακαδημαϊκές Εκδόσεις

- [50] H. Garcia-Molina, J. D. Ullman, and J. Widom, *Database Systems: The Complete Book*. Upper Saddle River, N.J: Pearson Prentice Hall, 2009.
- [51] A. Gaulton *et al.*, “ChEMBL: A large-scale bioactivity database for Drug Discovery,” *Nucleic Acids Research*, vol. 40, no. D1, Sep. 2011. doi:10.1093/nar/gkr777
- [52] “PubChem”, National Center for Biotechnology Information. PubChem Compound Database, <https://pubchem.ncbi.nlm.nih.gov/> (accessed Aug. 23, 2024).
- [53] A. Gaulton *et al.*, “ChEMBL: A large-scale bioactivity database for Drug Discovery,” *Nucleic Acids Research*, vol. 40, no. D1, Sep. 2011. doi:10.1093/nar/gkr777
- [54] F. Grisoni, V. Consonni, and R. Todeschini, “Impact of molecular descriptors on Computational Models,” *Methods in Molecular Biology*, pp. 171–209, 2018. doi:10.1007/978-1-4939-8639-2_5
- [55] M. Ματσούκας (2023) “5.Library-Design,” in *Φαρμακευτική Μηχανική*.
- [56] D. S. Wigh, J. M. Goodman, and A. A. Lapkin, “A review of molecular representation in the age of machine learning,” *WIREs Computational Molecular Science*, vol. 12, no. 5, Feb. 2022. doi:10.1002/wcms.1603
- [57] N. M. O’Boyle, “Towards a universal smiles representation - a standard method to generate canonical smiles based on the inchi,” *Journal of Cheminformatics*, vol. 4, no. 1, Sep. 2012. doi:10.1186/1758-2946-4-22
- [58] D. Weininger, “Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules,” *Journal of Chemical Information and Computer Sciences*, vol. 28, no. 1, pp. 31–36, Feb. 1988. doi:10.1021/ci00057a005
- [59] A. R. Leach and V. J. Gillet, “Chapter 3: Molecular Descriptors,” in *An Introduction to Chemoinformatics-Revised Edition*, Dordrecht, The Netherlands: Springer, 2007, pp. 53–74
- [60] L. Xue and J. Bajorath, “Molecular descriptors in Chemoinformatics, computational combinatorial chemistry, and virtual screening,” *Combinatorial Chemistry & High Throughput Screening*, vol. 3, no. 5, pp. 363–372, Oct. 2000. doi:10.2174/1386207003331454
- [61] A. Cereto-Massagué *et al.*, “Molecular fingerprint similarity search in virtual screening,” *Methods*, vol. 71, pp. 58–63, Jan. 2015. doi:10.1016/j.ymeth.2014.08.005
- [62] I. Mendolia, S. Contino, U. Perricone, R. Pirrone, and E. Ardizzone, “A convolutional neural network for virtual screening of molecular fingerprints,” *Lecture Notes in Computer Science*, pp. 399–409, 2019. doi:10.1007/978-3-030-30642-7_36
- [63] W. P. Walters, M. T. Stahl, and M. A. Murcko, “Virtual screening—an overview,” *Drug Discovery Today*, vol. 3, no. 4, pp. 160–178, Apr. 1998. doi:10.1016/s1359-6446(97)01163-x
- [64] C. Sottriffer, *Virtual Screening: Principles, Challenges, and Practical Guidelines*. Weinheim, Germany: Wiley-VCH, 2011.
- [65] S. F. Sousa, N. M.F.S.A. Cerqueira, P. A. Fernandes, and M. Joao Ramos, “Virtual screening in drug design and development,” *Combinatorial Chemistry & High Throughput Screening*, vol. 13, no. 5, pp. 442–453, Jun. 2010. doi:10.2174/138620710791293001
- [66] C. Pittinger and A. Mohapatra, “Software tools for toxicology and risk assessment,” *Information Resources in Toxicology*, pp. 631–638, 2009. doi:10.1016/b978-0-12-373593-5.00069-0

- [67] B. J. Neves *et al.*, “QSAR-based virtual screening: Advances and applications in Drug Discovery,” *Frontiers in Pharmacology*, vol. 9, Nov. 2018. doi:10.3389/fphar.2018.01275
- [68] K. Roy, *Advances in QSAR Modeling: Applications in Pharmaceutical, Chemical, Food, Agricultural and Environmental Sciences*. Cham, Switzerland: Springer, 2017.
- [69] T. Seidel, G. Ibis, F. Bendix, and G. Wolber, “Strategies for 3D pharmacophore-based Virtual Screening,” *Drug Discovery Today: Technologies*, vol. 7, no. 4, Dec. 2010. doi:10.1016/j.ddtec.2010.11.004
- [70] S. Raschka, “Molecular docking, estimating free energies of binding, and AutoDock’s semi-empirical force field,” Sebastian Raschka, PhD, https://sebastianraschka.com/Articles/2014_autodock_energycomps.html (accessed Aug. 23, 2024).
- [71] D. Giordano, C. Biancaniello, M. A. Argenio, and A. Facchiano, “Drug design by pharmacophore and virtual screening approach,” *Pharmaceuticals*, vol. 15, no. 5, p. 646, May 2022. doi:10.3390/ph15050646
- [72] Q. Li and S. Shah, “Structure-based Virtual Screening,” *Methods in Molecular Biology*, pp. 111–124, 2017. doi:10.1007/978-1-4939-6783-4_5
- [73] Θ. Μαυρομούστακος και Π. Ζουμπουλάκης, *Μοριακή Μοντελοποίηση: Εφαρμογές Στην Οργανική Και Φαρμακευτική Χημεία*. Ιατρικές Εκδόσεις Γιάννης Β. Παρισσιανός.
- [74] Μ. Ματσούκας (2023) “2-3. Docking –Pharmacophore Lecture,” in *Φαρμακευτική Μηχανική*.
- [75] D. M. George *et al.*, “Prodrugs for colon-restricted delivery: Design, synthesis, and in vivo evaluation of colony stimulating factor 1 receptor (CSF1R) inhibitors,” *PLOS ONE*, vol. 13, no. 9, Sep. 2018. doi:10.1371/journal.pone.0203567
- [76] A. R. Leach and V. J. Gillet, “Chapter 8: Virtual Screening,” in *An Introduction to Chemoinformatics-Revised Edition*, Dordrecht, The Netherlands: Springer, 2007, pp. 159–189
- [77] J. Dong *et al.*, “ADMETlab: A platform for systematic ADMET evaluation based on a comprehensively collected ADMET database,” *Journal of Cheminformatics*, vol. 10, no. 1, Jun. 2018. doi:10.1186/s13321-018-0283-x
- [78] Γ. Κόκοτος και Β. Μαγκριώτη, *Φαρμακοχημεία: Βασικές Έννοιες Της Φαρμακοχημείας*. Κάλλιπος, Ανοιχτές Ακαδημαϊκές Εκδόσεις, 2015.
- [79] M. S. Benedetti *et al.*, “Drug metabolism and pharmacokinetics,” *Drug Metabolism Reviews*, vol. 41, no. 3, pp. 344–390, Jul. 2009. doi:10.1080/10837450902891295
- [80] A. Alahmari, “Blood-brain barrier overview: Structural and functional correlation,” *Neural Plasticity*, vol. 2021, pp. 1–10, Dec. 2021. doi:10.1155/2021/6564585
- [81] Y. Zhang *et al.*, “A combined drug discovery strategy based on machine learning and Molecular Docking,” *Chemical Biology & Drug Design*, vol. 93, no. 5, pp. 685–699, Mar. 2019. doi:10.1111/cbdd.13494
- [82] S. Xia, E. Chen, and Y. Zhang, “Integrated Molecular Modeling and machine learning for drug design,” *Journal of Chemical Theory and Computation*, vol. 19, no. 21, pp. 7478–7495, Oct. 2023. doi:10.1021/acs.jctc.3c00814

- [83] P. M. Andersson and T. Lundstedt, "Hierarchical experimental design exemplified by QSAR evaluation of a chemical library directed towards the melanocortin 4 receptor," *Journal of Chemometrics*, vol. 16, no. 8–10, pp. 490–496, Aug. 2002. doi:10.1002/cem.738
- [84] M. Cai *et al.*, "Design of novel melanotropin agonists and antagonists with high potency and selectivity for human melanocortin receptors," *Peptides*, vol. 26, no. 8, pp. 1481–1485, Aug. 2005. doi:10.1016/j.peptides.2005.03.020
- [85] B. A. Falls and Y. Zhang, "Insights into the allosteric mechanism of setmelanotide (RM-493) as a potent and first-in-class melanocortin-4 receptor (MC4R) agonist to treat rare genetic disorders of obesity through an *in silico* approach," *ACS Chemical Neuroscience*, vol. 10, no. 3, pp. 1055–1065, Jul. 2018. doi:10.1021/acscemneuro.8b00346
- [86] C. Martin *et al.*, "Structure-based design of Melanocortin 4 receptor ligands based on the SHU-9119-HMC4R cocrystal structure," *Journal of Medicinal Chemistry*, vol. 64, no. 1, pp. 357–369, Nov. 2020. doi:10.1021/acs.jmedchem.0c01620
- [87] B. Zdrzil *et al.*, "The ChEMBL database in 2023: A drug discovery platform spanning multiple bioactivity data types and time periods," *Nucleic Acids Research*, vol. 52, no. D1, Nov. 2023. doi:10.1093/nar/gkad1004
- [88] M. Davies *et al.*, "ChEMBL Web Services: Streamlining Access to drug discovery data and utilities," *Nucleic Acids Research*, vol. 43, no. W1, Apr. 2015. doi:10.1093/nar/gkv352
- [89] D. I. Ugwu and J. Conradie, "Anticancer properties of complexes derived from bidentate ligands," *Journal of Inorganic Biochemistry*, vol. 246, p. 112268, Sep. 2023. doi:10.1016/j.jinorgbio.2023.112268
- [90] S. Ulenberg *et al.*, "In vitro approach for identification of a leading cytochrome P450 isoenzyme responsible for biotransformation of novel arylpiperazine drug candidates and their inhibition potency towards CYP3A4," *Acta Poloniae Pharmaceutica - Drug Research*, vol. 77, no. 1, pp. 69–76, Feb. 2020. doi:10.32383/appdr/111813
- [91] Rdkit, "Rdkit/rdkit: The official sources for the RDKit Library," GitHub, <https://github.com/rdkit/rdkit> (accessed Aug. 23, 2024).
- [92] "Feature importances with a forest of trees," scikit-learn.org/stable/auto_examples/ensemble/plot_forest_importances.html (accessed Aug. 23, 2024).
- [93] J. Brownlee, "How to calculate feature importance with python," MachineLearningMastery.com, <https://machinelearningmastery.com/calculate-feature-importance-with-python/> (accessed Aug. 23, 2024).
- [94] S. K. Burley *et al.*, "RCSB Protein Data Bank: Celebrating 50 years of the PDB with new tools for understanding and visualizing biological macromolecules in 3D," *Protein Science*, vol. 31, no. 1, pp. 187–208, Nov. 2021. doi:10.1002/pro.4213
- [95] S. K. Burley *et al.*, "RCSB Protein Data bank: Tools for visualizing and understanding biological macromolecules in 3D," *Protein Science*, vol. 31, no. 12, Nov. 2022. doi:10.1002/pro.4482
- [96] S. K. Burley *et al.*, "RCSB Protein Data Bank (RCSB.org): Delivery of experimentally-determined PDB structures alongside one million computed structure models of proteins from Artificial Intelligence/machine learning," *Nucleic Acids Research*, vol. 51, no. D1, Nov. 2022. doi:10.1093/nar/gkac1077

- [97] R. P. D. Bank, "6W25: Crystal structure of the melanocortin-4 receptor (MC4R) in complex with shu9119," RCSB PDB, <https://www.rcsb.org/structure/6W25> (accessed Aug. 25, 2024).
- [98] "Protein preparation wizard," Schrödinger, <https://www.schrodinger.com/life-science/learn/white-papers/protein-preparation-wizard/> (accessed Aug. 25, 2024).
- [99] K. B. Louie *et al.*, "Mass spectrometry for natural product discovery," *Comprehensive Natural Products III*, pp. 263–265, 2020. doi:10.1016/b978-0-12-409547-2.14834-6
- [100] "Research compounds," Specs, <https://www.specs.net/index.php?page=2019041215290210#naturalproducts> (accessed Aug. 25, 2024).
- [101] T. Sterling and J. J. Irwin, "Zinc 15 – ligand discovery for everyone," *Journal of Chemical Information and Modeling*, vol. 55, no. 11, pp. 2324–2337, Nov. 2015. doi:10.1021/acs.jcim.5b00559
- [102] J. J. Irwin *et al.*, "Zinc20—a free ultralarge-scale chemical database for Ligand Discovery," *Journal of Chemical Information and Modeling*, vol. 60, no. 12, pp. 6065–6073, Oct. 2020. doi:10.1021/acs.jcim.0c00675
- [103] Y. Kochnev, E. Hellemann, K. C. Cassidy, and J. D. Durrant, *Webina: An open-source library and web app that runs autodock vina entirely in the web browser*, Dec. 2019. doi:10.1101/2019.12.18.881789
- [104] R. A. Friesner *et al.*, "Glide: A new approach for rapid, accurate docking and scoring. 1. method and assessment of Docking Accuracy," *Journal of Medicinal Chemistry*, vol. 47, no. 7, pp. 1739–1749, Feb. 2004. doi:10.1021/jm0306430
- [105] "What are the main differences between HTVS, SP, and XP docking?," Schrödinger, <https://support.schrodinger.com/s/article/1013> (accessed Aug. 25, 2024).
- [106] "Docking and scoring," Schrödinger, <https://www.schrodinger.com/life-science/learn/white-papers/docking-and-scoring/> (accessed Aug. 25, 2024).
- [107] L. Fu *et al.*, "ADMETlab 3.0: An updated comprehensive online admet prediction platform enhanced with broader coverage, improved performance, API functionality and decision support," *Nucleic Acids Research*, vol. 52, no. W1, Apr. 2024. doi:10.1093/nar/gkae236
- [108] K. Swanson *et al.*, "ADMET-ai: A machine learning admet platform for evaluation of large-scale chemical libraries," *Bioinformatics*, vol. 40, no. 7, Jun. 2024. doi:10.1093/bioinformatics/btae416
- [109] R. Todeschini and V. Consonni, *Methods and Principles in Medicinal Chemistry-Handbook of Molecular Descriptors*, vol. 11. Weinheim: Wiley-VCH, 2000.
- [110] L. B. Kier, "A shape index from molecular graphs," *Quantitative Structure-Activity Relationships*, vol. 4, no. 3, pp. 109–116, Jan. 1985. doi:10.1002/qsar.19850040303
- [111] L. B. Kier, "Inclusion of symmetry as a shape attribute in Kappa index analysis," *Quantitative Structure-Activity Relationships*, vol. 6, no. 1, pp. 8–12, Jan. 1987. doi:10.1002/qsar.19870060103
- [112] L. H. Hall, Brian. Mohny, and L. B. Kier, "The electrotopological state: Structure information at the atomic level for molecular graphs," *Journal of Chemical Information and Computer Sciences*, vol. 31, no. 1, pp. 76–82, Feb. 1991. doi:10.1021/ci00001a012

- [113] P. Labute, “A widely applicable set of descriptors,” *Journal of Molecular Graphics and Modelling*, vol. 18, no. 4–5, pp. 464–477, 2000. doi:10.1016/s1093-3263(00)00068-1
- [114] J. Gasteiger and M. Marsili, “Iterative partial equalization of orbital electronegativity—a rapid access to atomic charges,” *Tetrahedron*, vol. 36, no. 22, pp. 3219–3228, Jan. 1980. doi:10.1016/0040-4020(80)80168-2
- [115] J. Yu *et al.*, “Determination of the melanocortin-4 receptor structure identifies Ca²⁺ as a cofactor for ligand binding,” *Science*, vol. 368, no. 6489, pp. 428–433, Apr. 2020. doi:10.1126/science.aaz8995
- [116] R. Cermak and S. Wolffram, “The potential of flavonoids to influence drug metabolism and pharmacokinetics by local gastrointestinal mechanisms,” *Current Drug Metabolism*, vol. 7, no. 7, pp. 729–744, Oct. 2006. doi:10.2174/138920006778520570
- [117] ADMETlab 3.0 explanation, <https://admetlab3.scbdd.com/explanation/#/> (accessed Aug. 29, 2024).
- [118] Y. Guttman and Z. Kerem, “Dietary inhibitors of CYP3A4 are revealed using virtual screening by using a new deep-learning classifier,” *Journal of Agricultural and Food Chemistry*, vol. 70, no. 8, pp. 2752–2761, Feb. 2022. doi:10.1021/acs.jafc.2c00237
- [119] V. Sandoval *et al.*, “Metabolic impact of flavonoids consumption in obesity: From central to peripheral,” *Nutrients*, vol. 12, no. 8, p. 2393, Aug. 2020. doi:10.3390/nu12082393
- [120] D. Song, L. Cheng, X. Zhang, Z. Wu, and X. Zheng, “The modulatory effect and the mechanism of flavonoids on obesity,” *Journal of Food Biochemistry*, vol. 43, no. 8, Jun. 2019. doi:10.1111/jfbc.12954
- [121] Hussain A, Farzam K. Setmelanotide “Setmelanotide,” StatPearls [Internet]. <https://www.ncbi.nlm.nih.gov/books/NBK589641/> (accessed Aug. 29, 2024).