LC–MS/MS Profiling and Machine Learning–Guided Virtual Screening of Dual EGFR/HER2 Inhibitors from Ceratonia siliqua L. Pod Extract for Colorectal Cancer.

by

Deli-Bright Nii Tettey Oku (21009414)

submitted in accordance with the requirements for
the degree of

Master of Science

in Chemistry

at the

UNIVERSITY OF SOUTH AFRICA

SUPERVISOR: Dr Ramakwala Christinah Chokwe

CO-SUPERVISOR: Dr Yannick Belo Nuapia

September 2025

DECLARATION

Name: Deli-Bright Nii Tettey Oku

Student number: 21009414

Degree: Master of Science in Chemistry

Exact wording of the title of the dissertation as appearing on the electronic copy submitted for examination:

LC–MS/MS Profiling and Machine Learning–Guided Virtual Screening of Dual EGFR/HER2 Inhibitors from Ceratonia siliqua L. Pod Extract for Colorectal Cancer.

I declare that the above dissertation is my own work and that all the sources that I have used or quoted have been indicated and acknowledged by means of complete references.

I further declare that I submitted the dissertation to originality checking software and that it falls within the accepted requirements for originality.

I further declare that I have not previously submitted this work, or part of it, for examination at Unisa for another qualification or at any other higher education institution.

(The dissertation will not be examined unless this statement has been submitted.)

SIGNATURE DATE: September 2025

UNIVERSITY OF SOUTH AFRICA

KEY TERMS DESCRIBING THE TOPIC OF A DISSERTATION/THESIS

The Executive Committee of Senate decided that in order to assist the library with retrieval of information, master's and doctoral students must list approximately ten key terms which describe the topic of the dissertation/thesis at the end of the summary of the dissertation/thesis.

If the dissertation/thesis is not written in English, the key terms in English must be listed at the end of the English summary.

The following is an example of key terms used for a thesis/dissertation:

Title of thesis/dissertation:

LC–MS/MS Profiling and Machine Learning–Guided Virtual Screening of Dual EGFR/HER2 Inhibitors from Ceratonia siliqua L. Pod Extract for Colorectal Cancer.

KEY TERMS:

EGFR; HER2 receptor; LC–MS/MS; machine learning; stacking ensemble; anti-cancer; natural product; phytochemical profiling

ACKNOWLEDGEMENTS

- I am profoundly thankful to Dr Ramakwala Christinah Chokwe and Dr Yannick Belo Nuapia for their constant guidance, encouragement, and support throughout my MSc journey. Their mentorship and meaningful input were instrumental in the successful completion of this research project.
- The financial assistance from the University of South Africa's College of Science and Technology (CSET) played a vital role in the successful completion of this project, and I remain deeply appreciative of their generous support
- I would also like to acknowledge the Department of Chemistry at the University of South Africa (UNISA) for providing the essential facilities and resources that made this research possible.
- I wish to extend my heartfelt appreciation to Dr Garland More for his invaluable support with the antioxidant and cytotoxicity analyses, and to Mr Damilare Babatunde for his outstanding expertise in molecular docking studies. My sincere thanks also go to Mr Belete Gebreyohannes for his support with the LC–MS/MS analysis. In addition, I gratefully acknowledge Kazimingi Nursery Farm (https://www.kazimingi.co.za/) for generously donating fresh samples of *Ceratonia siliqua* L. used in this project. Their combined efforts were crucial to the successful completion of this research, and I remain genuinely appreciative of their unwavering support.
- I am profoundly thankful to my family and friends (my mother, Bernice Adoley Hammond; my father, Roger Okuley Oku; my grandfather, Oloboi Commodore; my beloved siblings, Naa Nyokor Akweley and Nii Mensah Oko Oku; and my dear friend, Lindelwa Buthelezi) for their constant love, encouragement, and support throughout this journey. Their enduring presence

has been a pillar of strength and inspiration, motivating me to remain determined and focused on achieving my goals and aspirations.

- Lastly, I offer my heartfelt thanks to the Almighty for endowing me with the strength and perseverance to complete this journey. His divine grace has guided me every step of the way, and I remain eternally thankful for His blessings.

PUBLICATION(S)

Accepted publication

1. Oku, Deli-Bright N.T.; Babatunde, Damilare D.; Nuapia, Yannick*; More, Garland K.; Chokwe, Ramakwala Christinah*. "Harnessing Machine Learning for the Virtual Screening of Natural Compounds as Both EGFR and HER2 Inhibitors in Colorectal Cancer: A Novel Therapeutic Approach**",** *ACS Omega, 2025***.** Published, November 19, 2025. https://doi.org/10.1021/acsomega.5c07683
2. Deli-Bright Nii Tettey Oku, Ramakwala Christinah Chokwe. Harnessing machine learning for the virtual screening of natural compounds as both EGFR and HER2 inhibitors in colorectal cancer: A novel therapeutic approach [abstract]. In: *Proceedings of the AACR Special Conference in Cancer Research: Artificial Intelligence and Machine Learning*; 2025 Jul 10-12; Montreal, QC, Canada. Philadelphia (PA): AACR; *Clin. Cancer Res.* 2025; 31(13_Suppl): Abstract nr A001. doi:10.1158/1557-3265.AIMACHINE-A001.

Manuscripts in preparation

3. Oku, Deli-Bright N.T.; Nuapia, Yannick; More, Garland K.; Chokwe, Ramakwala Christinah**.** "Chemical Profiling and Scaffold-Based Drug Discovery Analysis of Bioactive Compounds from *Ceratonia siliqua* L. with Computational and Biological Validation*", Journal of Chemical Information and Modeling, 2026* **(Manuscript submitted).**
4. Olusengu, Samuel; Oku, Deli-Bright N.T.; Nuapia, Yannick; Chokwe, Ramakwala Christinah. *"*Challenges and Advances in Integrative Machine Learning with the Discovery of New Drug Candidates from Medicinal Plants*"*, *Review paper* **(IN PREPARATION).**

CONFERENCE(S)

1. Oku, Deli-Bright N.T.; Nuapia, Yannick; More, Chokwe; Ramakwala, Christinah. "LC–MS/MS, Machine Learning and Virtual Screening of Anti-cancer and Anti-inflammatory Compounds from *Ceratonia siliqua* L. (Carob Tree)." **Oral** presentation at the *SACI Central Young Chemist Symposium*; October 31, 2025; University of South Africa, Florida, South Africa. Awarded 1st Prize – MSc Category.

2. Oku, Deli-Bright N.T.; Babatunde, Damilare D.; Nuapia, Yannick; More, Garland K.; Chokwe, Ramakwala Christinah. *"*Harnessing Machine Learning for the Virtual Screening of Natural Compounds as Both EGFR and HER2 Inhibitors in Colorectal Cancer: A Novel Therapeutic Approach*."* **Poster** presented at the *73rd American Society for Mass Spectrometry (ASMS) Conference on Mass Spectrometry and Allied Topics*; June 3, 2025; Baltimore Convention Center, Baltimore, Maryland, USA. Abstract ID: 323827.

ABSTRACT

Colorectal cancer (CRC), a malignancy that develops in the colon or rectum, is frequently associated with epidermal growth factor receptor (EGFR) overexpression, observed in up to 85% of cases, whereas human epidermal growth factor receptor 2 (HER2) amplification or overexpression is present in a smaller subset (approximately 2–6%, with increased prevalence in selected molecular subgroups). These molecular alterations highlight the therapeutic relevance of targeting EGFR and HER2 pathways in CRC management. Despite progress in targeted therapy development, most existing treatment approaches focus on inhibiting either EGFR or HER2 individually. These single-target therapies frequently demonstrate reduced efficacy because of alterations in downstream effectors such as the Kirsten rat sarcoma viral oncogene homolog (KRAS) and the activation of alternative signalling pathways that promote continued tumour growth. Consequently, designing therapeutic approaches that can concurrently block both EGFR and HER2 represents an essential and promising area of research.

In this research, an innovative machine learning (ML)-driven stacking ensemble framework was established to accurately identify dual EGFR and HER2 inhibitors based on Simplified Molecular-Input Line-Entry System (SMILES) representations. A comprehensive benchmark dataset comprising active and inactive compounds targeting EGFR and HER2 was compiled from the ChEMBL database. Utilising this dataset, forty baseline models were developed and fine-tuned using various molecular descriptors and ML algorithms. The predictions from these models were then integrated through logistic regression (LR) to produce a highly reliable stacking ensemble classifier.

This predictive model was further applied to natural bioactive compounds obtained from liquid chromatography–tandem mass spectrometry (LC–MS/MS) profiling of *Ceratonia siliqua* L. pod

extract, which were annotated using established spectral libraries and subsequently subjected to machine learning–guided virtual screening. The cytotoxic activity of the *Ceratonia siliqua* L. pod extract was confirmed experimentally using the MTT (3-(4,5-dimethylthiazol-2-yl)-2,5-diphenyl tetrazolium bromide) assay against human colorectal carcinoma cell line (HCT116) and non-cancerous Vero cells. The extract exhibited an IC_{50} (half maximal inhibitory concentration) value of 13.32 ± 1.09 μg/mL in HCT116 cells, underscoring its notable anti-cancer potential.

To support the experimental outcomes, molecular docking and in silico ADMET (absorption, distribution, metabolism, excretion, and toxicity) evaluations were carried out on the compounds identified from the LC–MS/MS dataset using the stacking model, alongside four Food and Drug Administration (FDA) approved anticancer drugs for comparative analysis. Among all screened molecules, NCGC00385704-01, identified from LC–MS/MS spectral data through library matching, exhibited strong dual inhibitory potential against both EGFR and HER2. Overall, this study highlights *Ceratonia siliqua* L. as a valuable source of potential lead molecules for colorectal cancer therapy through dual EGFR/HER2 inhibition and underscores the power of integrating computational and experimental approaches in natural product-based drug discovery.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF EQUATIONS

LIST OF ABBREVIATIONS AND ACRONYMS

AdaBoost	Adaptive Boosting
ADMET	Absorption, Distribution, Metabolism, Excretion, and Toxicity
AI	Artificial Intelligence
ALogP	Computational method developed by Ghose-Crippen-Viswanadhan to calculate the logarithm of the octanol–water partition coefficient (logP)
AUC-ROC	Area under the receiver operating characteristic curve
C. siliqua	*Ceratonia siliqua* L.
CRC	Colorectal cancer
DAD	Diode array detector
DMSO	Dimethyl sulfoxide
DNA	Deoxyribonucleic acid
DPPH	2,2-diphenyl-1-picrylhydrazyl
DT	Decision Tree
EGFR	Epidermal growth factor receptor
ET	Extra Trees (short for Extremely Randomised Trees)
FDA	Food and Drug Administration
GBM	Gradient Boosting Machine
GLOBOCAN (now GCO)	Global Cancer Observatory (an interactive web-based platform developed by the International Agency for Research on Cancer (IARC)
GNPS	Global Natural Product Social Molecular Networking

HBAs	Hydrogen bond acceptors
HBDs	Hydrogen bond donors
HER2	Human epidermal growth factor receptor 2
HPLC	High-performance liquid chromatography
HPLC–MS/MS	High-performance liquid chromatography–tandem mass spectrometry
*k*NN	*k*-Nearest Neighbours
KRAS	Kirsten rat sarcoma 2 viral oncogene homolog
LBVS	Ligand-based virtual screening
LC–MS/MS	Liquid chromatography–tandem mass spectrometry
LGBM	Light Gradient Boosting Machine
logP	Logarithm of the octanol–water partition coefficient
LOTUS	LOTUS Natural Products Database
LR	Logistic regression
MACCS	Molecular ACCess System
ML	Machine learning
MLP	Multilayer Perceptron
MTT	3-(4,5-dimethylthiazol-2-yl)-2,5-diphenyl tetrazolium bromide
NB	Naive Bayes
nHBA	Number of hydrogen bond acceptors
nHBD	Number of hydrogen bond donors
NPs	Natural products
RF	Random Forest

ROS	Reactive Oxygen Species
RSA	Radical Scavenging Activity
RTKs	Receptor Tyrosine Kinases
SBVS	Structure-Based Virtual Screening
SMILES	Simplified Molecular-Input Line-Entry System
SVM	Support Vector Machine
VS	Virtual Screening

CHAPTER 1: INTRODUCTION

PREAMBLE

This chapter presents a concise background on modern drug discovery, the role of machine learning, approaches in virtual screening, and the contribution of natural product chemistry in drug development, with particular attention to the biological activities of *Ceratonia siliqua* L. It further outlines the research problem, objectives, hypothesis, and rationale for undertaking the study. Finally, a summary of the dissertation structure is provided, highlighting the content of each subsequent chapter.

1.1. Background

The evolution of drug discovery highlights plants as valuable reservoirs of novel bioactive molecules, which continue to present significant challenges for the contemporary pharmaceutical sector [1]. Natural products (NPs) derived from these plants continue to offer invaluable chemical diversity and inspiration, serving as a foundation for the development of drugs [2]. Natural products have historically been central to the development of new drugs, especially for the treatment of infectious illnesses and cancer [3]. Newman and Cragg (2016) reported that approximately 28–34% of all approved small-molecule drugs are derived from or inspired by natural products, with a significantly higher contribution observed in anticancer drug development, ranging from approximately 49–75% [4]. This underscores the importance of exploring novel reservoirs of bioactive compounds to support the development of drugs targeting a range of infectious diseases [5]. In this context, screening natural compounds from traditional medicine could serve as a promising approach for new drug development. Although there is an urgent demand, producing novel drugs within a short timeframe is not feasible. This is because the conventional drug discovery process is time-intensive, requires substantial investment, and has historically demonstrated a relatively low success rate [2]. Consequently, the rise of artificial intelligence (AI) and machine learning (ML) has captivated the minds of biochemical researchers.

In the past few years, notable advancements in the field of ML have contributed to the development of computer-based approaches that learn from data and can simulate intricate chemical and biological processes with greater precision [6]. Machine learning techniques have been effectively applied to a range of challenges, including in computational chemistry, analysis of pharmaceutical data, and predicting protein–ligand binding affinities, as well as exploring the molecular basis of protein functions and biochemical reactions [7-10]. As a branch of AI, ML provides an effective approach to virtual screening for identifying potential therapeutic candidates and is now widely used throughout almost all phases of drug discovery and development [2], [11]. Machine learning–based virtual screening (VS) provides a faster and more efficient method of developing novel drugs within a limited timeframe because it enables comprehensive in silico evaluation of millions of compounds, thereby enhancing the identification of promising therapeutic candidates. Therefore, applying machine learning to virtual screening proves considerably more efficient than conventional drug discovery methods. For example, during the SARS-CoV-2 virus outbreak, virtual screening provided a quick and useful technique for discovering novel antiviral compounds suitable for further optimisation [2].

Various medicinal plants are utilized extensively to cure illnesses like cancer [12], diabetes [13], and inflammation [14]. Cancer, in particular, remains a major global cause of morbidity and mortality, creating a constant demand for safer and more effective new therapies [15]. Naturally obtained compounds have gained growing demand as a result of their perceived lower toxicity compared to conventional treatments like chemotherapy [16]. Numerous plant species have shown promising anti-cancer activity, especially those used in traditional medicine in developing countries. Between 1940 and 2002, about 54% of approved anti-cancer drugs were derived from natural products. Notable examples are the Vinca alkaloids extracted from *Catharanthus roseus* and paclitaxel obtained from *Taxus baccata* [17].

Plant-derived molecules, including polyphenols, taxols, and brassinosteroids, have demonstrated potent anti-cancer effects [16]. Polyphenolic compounds such as flavonoids, tannins, as well as curcumin, resveratrol, and gallocatechin (including gallic acid, epigallocatechin, epicatechin-3-gallate, and epigallocatechin-3-gallate) are recognised as anti-cancer agents [18]. Hence, it was recently suggested that dietary polyphenols can support health and lower cancer risk through their natural antioxidant activity, since their cytotoxic effects on various cancer cell types have been shown and their antioxidant capabilities have been established [16]. These compounds may reduce cancer risk by counteracting both inflammation and oxidative stress which are important factors in the development of cancer [19].

Given the growing interest in plant-derived compounds for drug discovery, *Ceratonia siliqua* L. (often referred to as the carob tree) emerges as a promising natural product with significant pharmaceutical potential. This plant is known to be an abundant source of bioactive compounds with considerable economic importance, contributing to a variety of sectors, such as pharmaceuticals, agriculture, and cosmetics [20]. The pulp of *Ceratonia siliqua* L. contains a diverse array of physiologically active constituents such as sugars, cyclitols, polyphenols, minerals, fibres, and amino acids while its seeds are particularly rich in gum, polyphenols, and proteins. Plant polyphenols, which are abundant in *Ceratonia siliqua* L., have been substantially recognised for their chemopreventive and therapeutic properties against cancer [18], suggesting that this plant may harbour compounds with anti-cancer potential. Hence, the identification of such compounds from *Ceratonia siliqua* L., together with the development of therapeutic strategies that concurrently target both EGFR and HER2 receptors in colorectal cancer, represents an emerging and promising area of research. Although therapies targeting the epidermal growth factor receptor (EGFR) have demonstrated clinical benefits in colorectal cancer, their long-term effectiveness is often limited by the development of resistance mechanisms. These

include activating mutations in downstream signalling pathways, such as Kirsten rat sarcoma 2 viral oncogene homolog (KRAS), as well as the activation of compensatory receptor tyrosine kinases, particularly human epidermal growth factor receptor 2 (HER2) [21]. HER2-mediated signalling can bypass EGFR inhibition and sustain tumour cell proliferation and survival, thereby reducing the efficacy of EGFR-targeted monotherapies [22]. Consequently, dual inhibition of EGFR and HER2 has emerged as a promising therapeutic strategy. By simultaneously targeting both receptors, this approach can block parallel and compensatory signalling pathways, this may overcome resistance mechanisms and enhance treatment efficacy compared to single-target therapies. This provides a strong rationale for the exploration of dual EGFR/HER2 inhibitors in colorectal cancer. Therefore, this study aimed to virtually screen potential anti-cancer drug candidates from *Ceratonia siliqua* L. pods targeting EGFR and HER2 pathways using a high-throughput virtual screening approach. By integrating liquid chromatography–tandem mass spectrometry (LC–MS/MS) analysis and machine learning models, compounds extracted from *Ceratonia siliqua* L. pods were analysed, their molecular fingerprints were computed, and novel structures with desirable bioactivity profiles were predicted. This approach holds promise for identifying lead compounds that could help in the creation of more effective and safe treatments.

1.2. Problem statement

Colorectal cancer (CRC) remains a major malignancy of the digestive system and a significant global health burden. In 2020, CRC accounted for over 1.9 million new cases and approximately 935 000 deaths, ranking as the third most commonly diagnosed cancer and the second leading cause of cancer-related mortality worldwide [23]. In South Africa alone, over 108 000 new cancer cases and nearly 57 000 deaths were recorded in 2020 [24]. Despite decades of progress in chemotherapy and other systemic cancer treatments such as hormone therapy, targeted therapy, and immunotherapy, current

therapies are still frequently associated with severe toxicities, including cardiotoxicity, nephrotoxicity, and pulmonary toxicity, which compromise patient safety and overall quality of life [25-30]. Furthermore, the emergence of drug resistance, particularly due to mutations in oncogenes such as Kirsten rat sarcoma 2 viral oncogene homolog (KRAS) and compensatory signalling pathways, substantially reduces the effectiveness of targeted monotherapies, including those aimed at the epidermal growth factor receptor (EGFR) or human epidermal growth factor receptor 2 (HER2) in colorectal cancer (CRC) proliferation [21]. Therefore, the development of therapeutic strategies of concurrently addressing both EGFR and HER2 receptors in CRC is an emerging field of research and clinical investigation. Natural products present a promising alternative to synthetic drugs, offering a reservoir of structurally diverse, bioactive compounds with potential chemoprotective effects [15]. However, the challenge is that drug discovery and development from natural sources remains a time-consuming, capital-intensive process, with high failure rates in clinical development. A study conducted in 2014 estimated that the average cost of developing a prescription drug was $2.87 billion [31], and has likely increased since then. Compounding this challenge is the under-documentation of plant-based anti-cancer agents in regions rich in biodiversity, such as South Africa [32]. In this context, the full potential of *Ceratonia siliqua* L. (carob tree) remains relatively underexplored in the context of machine learning–driven virtual screening for dual EFR/HER2 inhibition in colorectal cancer.

To overcome the limitations of traditional drug discovery, ML-driven virtual screening has emerged as a transformative approach. Machine learning can rapidly and accurately analyse large compound libraries, predict bioactivity, and identify promising drug candidates with reduced cost and time [33]. Recent developments in machine learning have made it possible to build efficient predictive models, including stacking ensemble methods that combine multiple algorithms for improved performance.

Therefore, this study aimed to address the urgent need for more effective and safer anticancer therapies by combining LC–MS/MS phytochemical profiling of *Ceratonia siliqua* L. pods with ML-based virtual screening to identify potential lead compounds for specific biological targets. Specifically, this

strategy seeks to identify natural compounds with dual inhibitory potential against EGFR and HER2, the two overexpressed and therapeutically relevant targets in CRC, using a novel stacking ensemble ML framework. This integrative approach not only accelerates drug discovery from natural sources but also contributes to the development of multi-targeted cancer therapies.

1.3. Aim and objectives

The aim of this study was to combine LC–MS/MS with machine learning techniques for the virtual screening of compounds derived from the pods of *Ceratonia siliqua* L. with potential anti-cancer properties, particularly targeting EGFR and HER2 receptors

The specific objectives were as follows:

- To perform phytochemical profiling of *Ceratonia siliqua* L. pods using LC–MS/MS.
- To compile and curate a dataset of known anti-cancer compounds, particularly those with inhibitory activity against EGFR and HER2, from relevant databases.
- To develop and optimise machine learning models, including a stacking ensemble approach, for the prediction of potential anti-cancer compounds based on molecular descriptors.
- To apply the developed stacking ensemble models for the virtual screening of phytochemicals identified from *Ceratonia siliqua* L. pod extracts, prioritising those with the highest predicted dual EGFR and HER2 inhibitory potential.
- To compare the predicted active compounds to known drugs approved by the FDA.
- To evaluate the cytotoxic and antioxidant activities of *Ceratonia siliqua* L. pod extracts through in vitro assays.

1.4. Rationale/Justification for the study

The increasing global burden of cancer necessitates the development of safer and more effective therapeutic strategies. Although significant progress has been made in cancer treatment, limitations such as drug resistance, toxicity, and reduced efficacy of single-target therapies continue to hinder clinical outcomes. Natural products remain a valuable source of structurally diverse bioactive compounds for drug discovery due to their pharmacological relevance and chemical diversity.

Ceratonia siliqua L. (carob tree) was selected for this study due to its rich polyphenolic composition and reported biological activities, including antioxidant and cytotoxic effects, which suggest potential relevance in cancer therapy. However, its phytochemical constituents have not been extensively explored using machine learning–driven virtual screening approaches for targeted cancer therapy.

In colorectal cancer, resistance to epidermal growth factor receptor (EGFR)-targeted therapies is frequently associated with compensatory activation of alternative signalling pathways, particularly involving human epidermal growth factor receptor 2 (HER2). As a result, single-target therapies often show limited long-term efficacy. Dual inhibition of EGFR and HER2 offers a promising strategy to overcome resistance mechanisms and enhance therapeutic effectiveness by simultaneously blocking parallel signalling pathways involved in tumour progression.

Therefore, this study integrates LC–MS/MS-based phytochemical profiling with machine learning–guided virtual screening to identify *Ceratonia siliqua* L. compounds with potential dual EGFR/HER2 inhibitory activity, providing a faster and cost-effective approach for prioritising promising anticancer drug candidates.

Hypothesis testing:

H_0 (null hypothesis): Bioactive compounds derived from *Ceratonia siliqua* L. pods do not exhibit significant anti-cancer properties, and machine learning models cannot accurately predict their bioactivity.

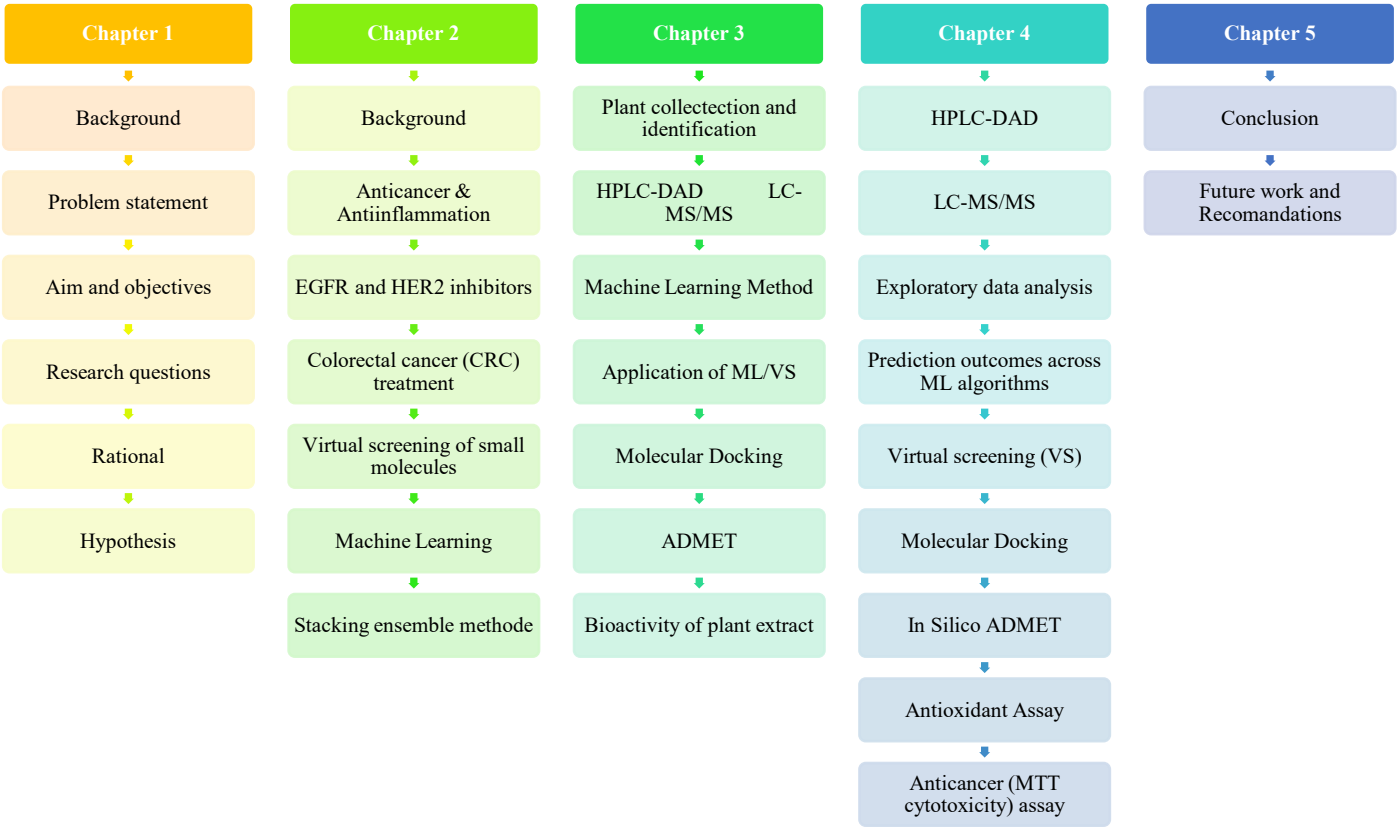
H_1 (alternative hypothesis): Bioactive compounds derived from *Ceratonia siliqua* L. pods possess potential anti-cancer properties, and their bioactivity can be accurately predicted using machine learning models.

1.5. Research questions

- Which bioactive phytochemicals are present in the pods (seed and pulp) of *Ceratonia siliqua* L., as identified through LC–MS/MS analysis?
- How can a curated dataset of known EGFR and HER2 inhibitors be utilised to train a machine learning model for predicting the bioactivity of natural compounds?
- Can machine learning models be developed and applied to identify *Ceratonia siliqua* L. compounds with dual EGFR and HER2 inhibitory potential for colorectal cancer treatment?
- To what extent can the stacking ensemble learning approach effectively prioritise bioactive compounds from *Ceratonia siliqua* L. with potential dual EGFR and HER2 inhibitory activity?
- Do extracts from *Ceratonia siliqua* L. pods demonstrate significant cytotoxic and antioxidant activities in preliminary in vitro evaluations?

1.6. Dissertation outline

This dissertation is structured into five chapters, as illustrated in the flow chart below, which summarizes the logical progression of the study from background to conclusion.


Chapter 1
Background
Problem statement
Aim and objectives
Research questions
Rational
Hypothesis
Chapter 2
Background
Anticancer & Antiinflammation
EGFR and HER2 inhibitors
Colorectal cancer (CRC) treatment
Virtual screening of small molecules
Machine Learning
Stacking ensemble methode
Chapter 3
Plant collectection and identification
HPLC-DAD LC-MS/MS
Machine Learning Method
Application of ML/VS
Molecular Docking
ADMET
Bioactivity of plant extract
Chapter 4
HPLC-DAD
LC-MS/MS
Exploratory data analysis
Prediction outcomes across ML algorithms
Virtual screening (VS)
Molecular Docking
In Silico ADMET
Antioxidant Assay
Anticancer (MTT cytotoxicity) assay
Chapter 5
Conclusion
Future work and Recomandations

CHAPTER 2: LITERATURE REVIEW

PREAMBLE

This section offers an in-depth review of recent studies and scholarly works concerning *Ceratonia siliqua* L. (carob tree), its morphology, traditional uses, and phytochemistry. Additionally, the chapter discusses anti-inflammatory and anti-cancer activities, emphasising their inter-relationship and collective relevance in the context of cancer. It also provides an overview of modern medicinal applications, especially in the development of therapeutic strategies for colorectal cancer (CRC) using virtual screening and machine learning approaches. The chapter further discusses the role of epidermal growth factor receptor (EGFR) and human epidermal growth factor receptor 2 (HER2) inhibitors in CRC treatment. The chapter concludes by elaborating on the theoretical frameworks applied in machine learning and ensemble techniques in drug discovery.

2.1. Background

2.1.1. *Ceratonia siliqua* L. (Carob tree)

The carob tree is among the oldest and most valuable plants recognised by mankind [34]. It is taxonomically identified as *Ceratonia siliqua* L. [35]. The carob tree is also referred to by common names such as St John's Bread or locust bean [36]. It is a typical evergreen member of the legume family, Fabaceae.

The carob tree originates from the Middle East, where it has been grown since antiquity. From that region, it was introduced by the early Greeks into Greece and subsequently distributed along the North African coastline, later reaching Spain and Portugal through Arab influence. During later periods, the carob tree spread beyond its native range into regions with Mediterranean-type climates. Its distribution extended to North Africa, including Morocco; parts of North and South America such as

California, Arizona, Mexico, Chile, and Argentina which was introduced mainly by Spanish settlers; to Australia through migrants from Mediterranean areas; and to South Africa and India via British influence [37], [38] (**Figure 2.1**).

The carob tree has been grown for many years across most countries of the Middle East and the Eastern Mediterranean region, serving both as animal feed and as nourishment for humans. In recent years, this species has gained considerable interest and economic significance. Its pods and seeds serve as essential inputs in the nutritional, pharmaceutical, and cosmetic sectors [32].

The latest report by the Food and Agriculture Organization (FAOSTAT, 2020–2023) of the United Nations indicates that indicate that global carob production ranges between approximately 49,000 and 56,000 tonnes annually, with cultivation largely concentrated in Mediterranean countries [34]. In addition to its geographical distribution, *Ceratonia siliqua L.* exhibits notable genetic diversity across different growing regions. Based on this variation, four principal genetic groups of *Ceratonia siliqua* L. have been recognised: southern Spain, southern Morocco, eastern Mediterranean, as well as the

central Mediterranean region which encompasses genotypes from Portugal, Algeria, France, Sardinia, Sicily, and the Balearic Islands [39].

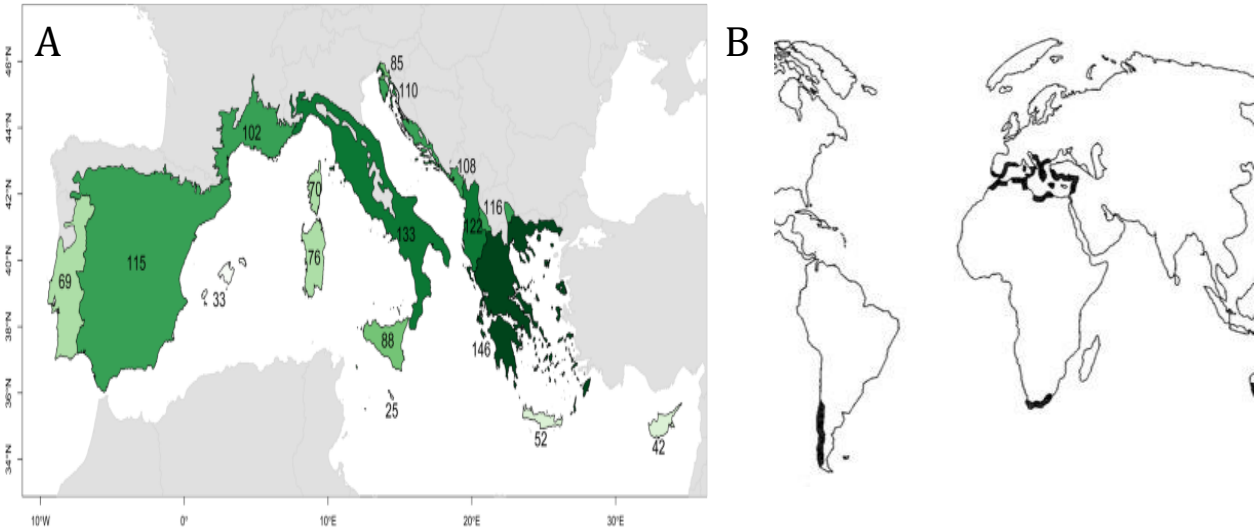


Figure 2.1: A map illustrating the distribution of the carob tree: (A) in Mediterranean region; and (B) around the world [38].

2.1.2. Morphology of *Ceratonia siliqua* L.

The *Ceratonia siliqua* L (carob tree) is an evergreen perennial species that typically grows as a shrub or small tree, capable of attaining heights of up to 10 m, characterised by a wide, dome-shaped crown, a sturdy trunk, coarse brown bark, and strong branches [38] **(Figure 2.2).** The carob fruit is produced primarily by female or hermaphrodite (self-pollinating) trees, as male trees do not bear fruit. The mature pod is hard and fibrous, containing multiple seeds enclosed within a tough, dark brown, elongated, flat, arched, or straight outer shell with thick sutures. Their edges are either smoothly curved or blunt, and their length varies between 10 and 30 cm [34], [40] **(Figure 2.3A).** The carob pod primarily consists of two components: the pulp, which makes up about 90%, and the seeds, accounting for the remaining 10%, which are key sources of the bioactive compounds investigated in this study. The pulp is made up of leathery outermost covering (known as pericarp) and a softer internal layer

(also known as mesocarp). This mesocarp separates the seeds, which are arranged transversely within the pod **(Figure 2.3B).** They are very hard and abundant, measuring 8–10 mm in length, 7–8 mm in width, and 3–5 mm in thickness; the testa is smooth, glossy, and brown in colour. Leaves measure 10–20 cm in length, with alternate pinnate arrangements that may or may not include a terminal leaflet **(Figure 2.3).** When present, the leaflets measure approximately 3–7 cm in length and exhibit an ovate to elliptical form. They generally appear in 4–10 opposite pairs, possessing a thick, leathery texture. The upper surface is dark green and glossy, while the underside is lighter in colour. The leaves display fine venation, slightly wavy edges, and small stipules [34], [38].

Figure 2.2: Image of the carob tree (*Ceratonia siliqua* L.) and its pods, showing both mature (brown) and immature (green) stages.

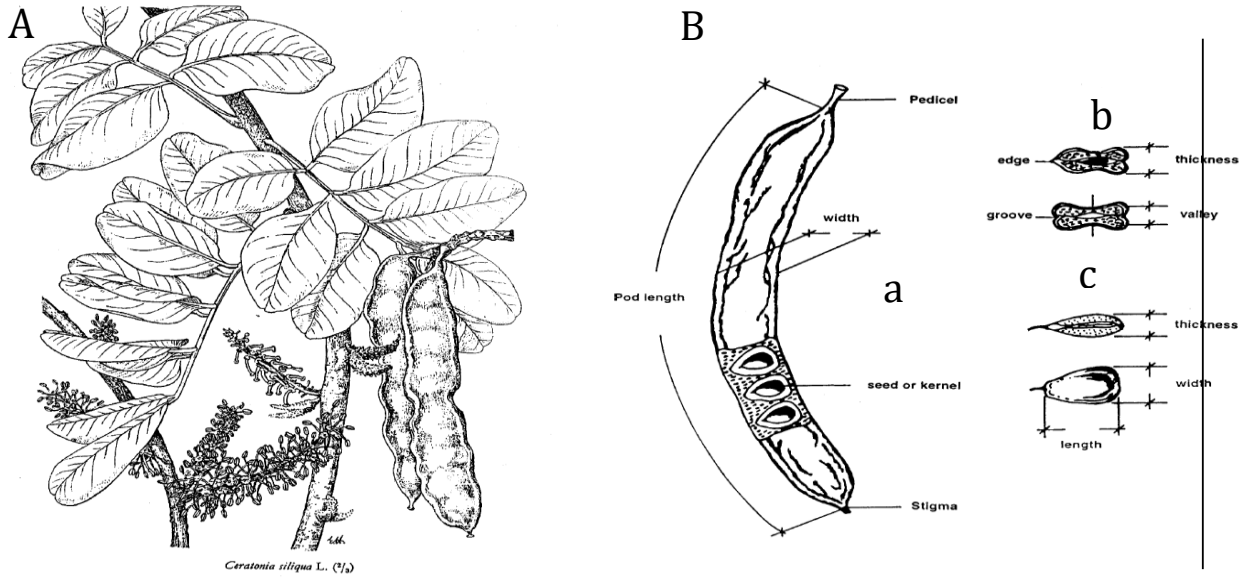


Figure 2.3: (A) Morphological features of *Ceratonia siliqua* L., including shoots, compound leaves, individual leaflets, pods, both male and female flower; and (B) key parts of *Ceratonia siliqua* L.: (a) Mature carob pod highlighting the outer structure; (b) section of the inflorescence showing pod attachment and arrangement; and (c) carob seed [41].

2.1.3. Traditional uses of *Ceratonia siliqua* L.

The carob tree is valued for its edible pods that contain a naturally sweet, fleshy pulp [42]. This sweet pulp has traditionally been employed as feed for livestock [39]. In terms of its conventional applications in folk and integrative medicine, the carob pods and seeds are used in Palestine to cure hypertension, and an infusion made from the leaves is utilized as an emetic for severe poisoning. Also, Tunisians use carob pods to treat diarrhoea, while medicines made from the tree are used to treat digestive disorders. Historically carob leaves are utilised to treat diarrhoea in Turkey, both the leaves and the seeds are used to manage diabetes in Morocco. In Southern Italy, carob pods are used as an expectorant, ingested by both humans and animals, and prescribed to treat intestinal irritation [43]. Furthermore, carob wood was traditionally widely utilised to make charcoal that burned slowly [44].

2.1.4. Phytochemistry of *Ceratonia siliqua* L.

There are several factors that affect the carob tree's phytochemical makeup, including genetic makeup, geographical origin, environmental conditions, physiological state, and cultivation practices. Darwish et al. (2021) identified flavonoids, polyphenols, alkaloids, carbohydrates, and amino acids within the carob pods [45]. The occurrence of these bioactive constituents accounts for the recognition of carob (*Ceratonia siliqua* L.) as an important medicinal plant for managing different ailments. These compounds were extracted using water and characterised using colorimetric techniques and high-performance liquid chromatography (HPLC). Among the detected compounds, gallic acid was found to be the predominant phenolic component in the aqueous pod extract, followed by catechin, protocatechuic acid, and cinnamic acid, while compounds such as p-coumaric acid, rutin, gentisic acid, p-hydroxybenzoic acid, vanillic acid, and ferulic acid were present in lower amounts [45]. This outcome is consistent with earlier findings, which also highlighted gallic acid as the main phenolic component found in carob pods [46].

In a 2011 study, Fadel et al. used HPLC to examine the presence of phenolic compounds in aqueous acetone extracts of carob pods (pulp and seed). The plant samples of *Ceratonia siliqua* L. were collected around Izouika and Reggada in southwest Morocco [47]. The chromatographic analysis demonstrated that the extracts were rich in phenolic compounds. In both regions, the pulp extracts contained higher levels of phenolics compared to the seeds, both qualitatively and quantitatively. The phenolic profile of the pulp was dominated by coumaric acid (20.52% in Izouika vs. 17.05% in Reggada) and gallic acid (17.8% in Izouika vs. 12.57% in Reggada). Similarly, in the seed extracts, coumaric acid and gallic acid were also the major phenolic acids, with coumaric acid comprising 8.07% in Izouika and 8.18% in Reggada, while gallic acid accounted for 5.01% in Izouika and 3.95% in Reggada. Additional benzoic acids detected in the carob tree include syringic acid, 4-hydroxybenzoic acid, and gentisic acid [44], [47].

Avallone et al. (1997) used HPLC to assess polyphenols in carob pods and succeeded in identifying condensed tannins (Proanthocyanins). Galloyl esters, gallic acid, epicatechin gallate, epigallocatechin gallate, quercetin glycosides, and other flavan-3-ol groups were among the proanthocyanins identified. Carob pods were also found to contain hydrolyzable tannins (0.95 mg of gallotannins and ellagitannins) [48]. Carob fibre was shown to contain a wide range of phenolic constituents, with 24 polyphenols identified, amounting to 3.94 g (dry weight). Gallic acid dominated the composition in various forms, including methyl gallate (1%), gallotannins (29%), and free gallic acid (42% of the total polyphenols). About 2% of the profile was made up of simple phenols, primarily cinnamic acid, whereas quercetin-3-O-α-L-rhamnoside and myricetin were the primary flavonoids found [49].

In a different study, Rtibi et al. (2016) carried out HPLC analysis and found that the main constituents of mature carob pods were tannic acid (9.01 ± 1.40%), catechin (19.10 ± 2.11%), and pyrogallol (48.02 ± 3.55%). Additionally, the aqueous extract of *Ceratonia siliqua* L. leaves was shown to be rich in phenolic compounds especially kaempferol (77 ± 2.43%), tannic acid (13 ± 0.45%), and catechin hydrate (4.30 ± 0.34%) [50]. Alkaloids and saponins were not present in the ethyl acetate and methanol extracts of carob bark in its crude form, but flavonoids and tannins were present, according to phytochemical analysis. According to this investigation, the methanol extract's phenolic content was higher than that of the ethyl acetate extract, which is in line with the fact that phenolic and different bioactive substances are typically soluble in polar solvents [51].

The carob tree is especially rich in flavonols such as quercetin, myricetin, kaempferol, and their glycosidic forms (Figure 2.4). Among these, quercetin and myricetin rhamnosides are typically the most dominant flavonoids in *Ceratonia siliqua* L. [52]. According to a semi-quantitative ultra-performance liquid chromatography (UPLC) analysis performed on the leaves of *Ceratonia siliqua* L.

from southern Morocco (Tafraoute), the primary compounds in the aqueous extract were luteolin-7-glucoside, followed by epicatechin, apigenin-7-glucoside, quercetin-3-O-glucoside, caffeic acid, gallic acid, and chlorogenic acid. These findings confirm that *Ceratonia siliqua* L. leaves serve as a valuable natural source of bioactive compounds [44].

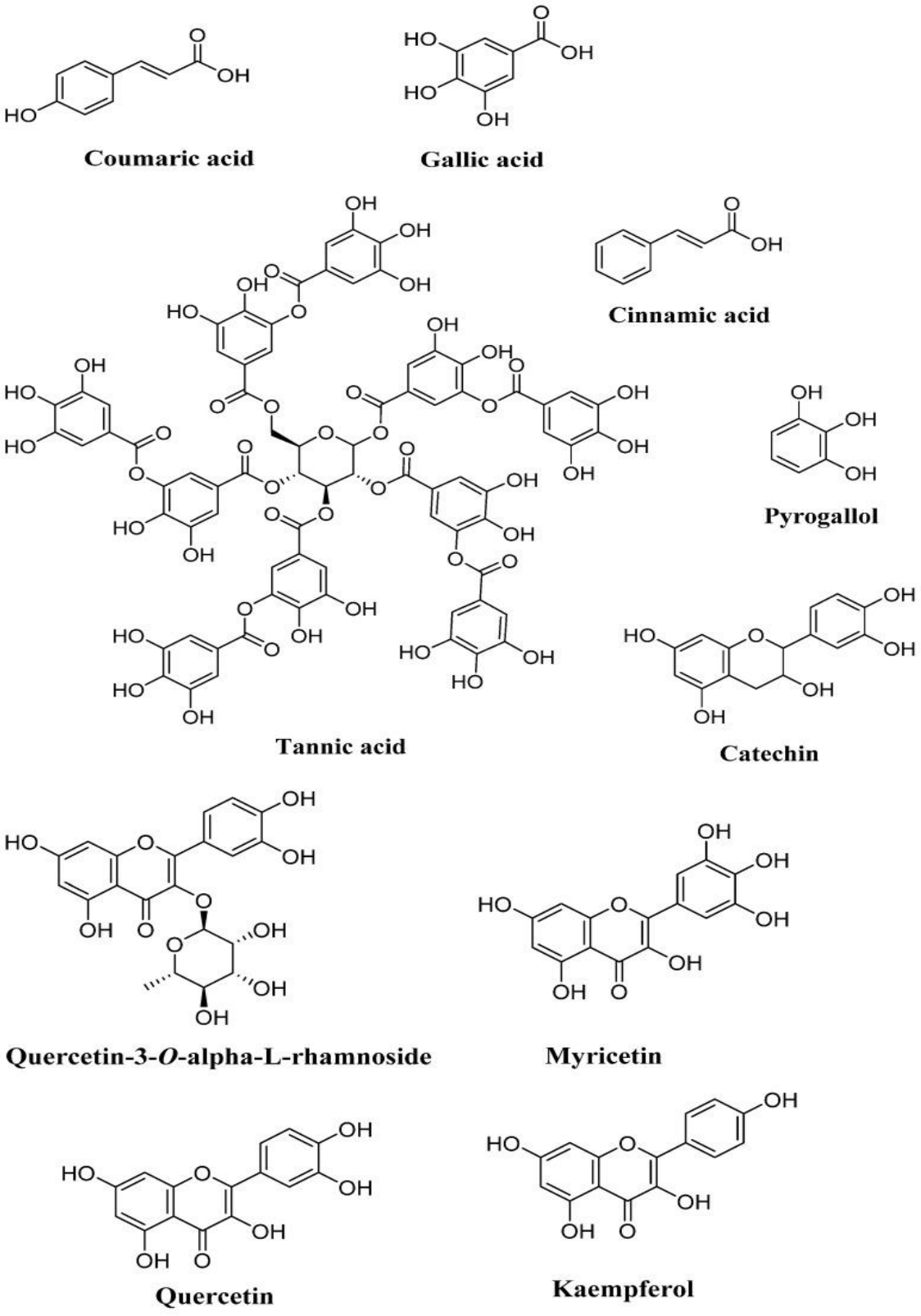


Figure 2.4: Chemical constituents identified in *Ceratonia siliqua* L. [53]

2.1.5. Chemical composition of *Ceratonia siliqua* L.

Studies have explored the chemical composition of different parts of *Ceratonia siliqua* L., revealing a broad spectrum of bioactive molecules. These include groups such as polyphenols, phenolic acids, and flavonoids, which were extracted using various techniques including maceration, Soxhlet, and ultrasound-assisted methods. **Table 2.1** provides a summary of the identified compounds, their respective plant parts, extraction approaches, and solvents employed in extraction.

Among the different carob parts examined, the leaves, pods, pulp, and seeds of *Ceratonia siliqua* L. were shown to contain diverse phenolic compounds and flavonoids. Notably, phenolic acids like gallic acid, syringic acid, and cinnamic acid were consistently reported across all plant parts, reflecting their widespread distribution throughout the species. Flavonoids such as quercetin, catechin, epicatechin, and kaempferol were particularly abundant in the leaves and pulp. Maceration appeared as the most frequently applied extraction method, commonly paired with solvents like ethanol, methanol, or water, while ultrasound-assisted and Soxhlet extractions were also used to enhance compound recovery. The chemical diversity observed in carob leaves, pods, pulp, and seeds highlights its potential for pharmacological and nutraceutical applications.

Table 2.1: The chemical constituents of the plant parts of carob (*Ceratonia siliqua* L.)

Class	Extraction method	Solvent	Compounds	Part of the plant	References
Phenol	Maceration	Methanol, acetone, water (6/6/7; v/v/v)	Resorcinol	Leaves, pods, pulp, seeds	[54]
Phenol	Maceration	n-hexane, methanol–water (6:4)	2,4-bis(dimethylbenzyl)-6-tert-butylphenol	Leaves	[54],[55]
Phenol	Maceration Soxhlet	EtOH 30%	bis(2,3,4-trihydroxyphenyl)methanone	Pod	[56], [57]

Phenolic acid	Ultrasound-assisted method Soxhlet	Ethanol and water	Vanillic acid, gentisic acid, caffeic acid, and hydroxybenzoic acid	Leaves, pods	[58], [59], [54], [60], [61]
Phenolic acid	-	-	Tannic acid	Leaves, pods, seeds	[62]
Phenolic acid	-	-	Ellagic acid, rosmarinic acid	Pods, pulp, seeds	[57], [61]
Phenolic acid	Maceration Soxhlet	EtOH 30%	Sinapic acid	Pulp, seeds	[57]
Phenolic acid	Maceration Soxhlet	EtOH 30% n-hexane, methanol–water (6:4)	Pyrogallol, methyl gallate, benzoic acid, protocatechuic acid	Pods, seeds	[56], [62], [58], [61], [57]

Phenolic acid	Ultrasound-assisted method Soxhlet	Ethanol and water	4-Hydroxy-coumaric acid	Leaves	[60]
Phenolic acid	Maceration Soxhlet	EtOH 30%	5-Caffeoylquinic acid, myristic acid, ascorbic acid	Pulp	[57]
Flavonoids	Maceration	Methanol, acetone, water (6/6/7; v/v/v)	Epicatechin, quercetin, kaempferol, luteolin, catechin, apigenin Epigallocatechin gallate, rutin, myricetin, naringenin	Leaves, pods, pulp, seeds	[62], [58], [56], [54], [59]
Flavonoids	Maceration Soxhlet	EtOH 30%	Iso-rhamnetin	Leaves, pods, seeds	[56], [57], [61]
Flavonoids	-	-	Leucoanthocyanins	Leaves, pulp, seeds	[63]

Flavonoids	-	-	Genistein	Leaves, pods	[58], [61]
Flavonoids	Ultrasound-Assisted Method Soxhlet	Ethanol and water	Quercitrin, catechin tannins	Leaves, pulps	[60]
Flavonoids	-	-	Anthocyanins	Pods, pulp, seeds	[58]
Flavonoids	Ultrasound-assisted method Soxhlet Maceration	Ethanol and water, Methanol, acetone,	Myricitrin, daidzein, flavonol, morin	Leaves	[54], [60]

Flavonoids	Maceration	Water	Rhamnosides, chrysoeriol, tricetin dimethyl ether, (iso)schaftoside-4′-*O*-glucoside, gallocatechin, chrysoeriol-*O*-deoxyheoxoside, dihydroxyflavanone hexoside, tetrahydroxy flavanone, trihydroxy flavone (apigenin isomer), kaempferide, methoxykaempferol, dihydroxy flavanone, tricetin dimethyl ether, cirsiliol, flavone glycosides, hydroxytyrosol	Pods	[58], [56], [64], [61]
Flavonoids	-	-	Cirsimaritin, catechol, isoquercitrin, flavonols 3′,4′,5,7-OH, 2-hexadecanol scutellarin tetramethyl ether, silybin	Pulp	[62]

			B, hydroxytyrosol, catechin gallate		
Flavonoids	-	-	Apigenin flavone, chrysin aglycones	Seeds	[61]

- (not specified)

2.1.6. Pharmacological studies conducted on *Ceratonia siliqua* L.

The cosmetic, food, and pharmaceutical industries make use of various parts of the carob tree for different applications. Carob gum, derived from the seeds of *Ceratonia siliqua* L., is utilised in the pharmaceutical sector in formulations such as pomades, anti-celiac preparations, tablets, and dental paste. The bioactive compounds present in the carob tree exhibit potential roles as anti-cancer, antioxidant, anti-inflammatory, anti-reflux, antidiabetic, antidiarrheal, antihyperlipidemic, antibacterial, antimicrobial, and antifungal agents [34], [39]. Several studies have reported that *Ceratonia siliqua* L. (carob) and its derived products, including carob powder and gum, can support human health and lower the chance of developing chronic diseases because they are rich in dietary fibre, polyphenols, and flavonoids compounds [65]. Kaïs Rtibi et al. (2017) found that the carob tree demonstrates antioxidant, anti-inflammatory, antibacterial, antidiarrheal, anti-ulcer, and gastrointestinal actions that restrict the absorption of glucose. The researchers came to the conclusion that *Ceratonia siliqua* L. has significant therapeutic and preventative potential, especially for gastrointestinal health, based on its chemical composition and pharmacological characteristics [66]. *Ceratonia siliqua* L. is primarily used industrially to extract endosperm from its seeds which is then ground to produce endosperm flour (locust bean gum), and germ flour as a by-product (Dakia 2011a). Locust bean gum is frequently used as a thickening agent in the food industry due to its abundance of galactomannans (β-D-mannose and α-D-galactose units) [44].

Extensive pharmacological investigations have been carried out on different parts of the *C. siliqua* plant. Some of these studies are presented in **Table 2.2** below.

Table 2.2: Pharmacological studies conducted on the different parts of *Ceratonia siliqua* L. and their bioactivity

Bioactive compounds	Activity	Plant parts	Reference
Polyphenols, gallic acid, epigallocatechin, catechin, quercetin, myricetin, kaempferol and rutin	Antioxidant, anti-inflammatory, antibacterial activities, effective against neurogenerative disorders and antitumoral activities	Carob pulp, carob fibre, carob pod and carob seed extract	[67], [68], [69]
D-pinitol	Anti-cancer and antidiabetic effect	Carob pulp extract	[70]
Tannins	Antidiarrheal effect	Carob bean juice	[71]
Cinnamic acid	Antioxidant activities and hepatoprotective effect	Carob fruit extract	[72], [73]
Galactomannan	Effective in gastrointestinal health	Carob pod endosperm	[74]
Chlorogenic acid and epicatechin	Reduce intestinal glucose absorption and laxative activities	Carob seed	[75]

Bioactive compounds	Activity	Plant parts	Reference
Flavonol glycosides	Increased lipid metabolism and reduce LDL level	Carob pulp extract	[76]

2.2. Anti-inflammation activities

Numerous chronic illnesses, including cancer, diabetes, gastrointestinal disorders, arthritis, and cardiovascular disorders, are frequently brought on by inflammation [77]. Anti-inflammatory agents are substances that reduce inflammation in the body, mainly by blocking compounds that trigger inflammatory responses. Aspirin, which originates from plants (willows), is one of the earliest anti-inflammatory drugs, it functions by irreversibly inhibiting cyclooxygenase (COX) enzymes; this action prevents the synthesis of prostaglandins and thromboxane, making it the most extensively used therapeutic agent. Current medications for inflammatory diseases are generally classified as steroidal and nonsteroidal anti-inflammatory drugs (NSAIDs). Steroidal drugs include corticosteroids, specifically glucocorticoids, which suppress inflammation while regulating the metabolism of fats, carbohydrates, and proteins. They are used in managing conditions such as arthritis, colitis, asthma, bronchitis, and skin rashes. Nonsteroidal anti-inflammatory drugs (NSAIDs) act primarily by inhibiting cyclooxygenase (COX) enzymes, thereby reducing the synthesis of prostaglandins involved in inflammation, pain, and fever, and are commonly used to treat conditions such as pain, fever, and inflammatory disorders [78].

Among phytochemicals derived from plants, compounds like flavonoids, saponins and alkaloids have been linked with anti-inflammatory properties in rheumatoid arthritis treatment and other systemic inflammatory illnesses. Several natural plant-based compounds are also recognised for their therapeutic role in reducing inflammation. For instance, quercetin has been shown to downregulate inflammatory cytokines [79]. Similarly, curcumin, derived from turmeric, and gingerols, from ginger, inhibit the release of inflammatory enzymes and cytokines, thereby reducing inflammation [80]. Although these studies demonstrate the anti-inflammatory potential of plant-derived compounds, including flavonoids present in *Ceratonia siliqua L.*, their role in modulating cancer-related signalling pathways, particularly receptor tyrosine kinases such as EGFR and HER2—remains insufficiently explored.

2.3. Antioxidant and anti-cancer mechanisms of plant-derived dietary polyphenols

Given the established relationship between chronic inflammation and carcinogenesis, antioxidant mechanisms represent a critical link between inflammation and cancer progression. Polyphenolic compounds, which are abundant in *Ceratonia siliqua L.*, are therefore of particular interest due to their dual role in reducing oxidative stress and modulating pathways associated with tumour development. Inflammation produces harmful free radicals which cause oxidative stress that in turn cause cell damage. Antioxidants counter this by neutralising the free radicals resulting in the reduction of inflammation and prevention of cell damage. Polyphenolic compounds can work as antioxidants, primarily through their ability to transfer hydrogen atoms or electrons, therefore counteracting radicals that are unstable and decreasing oxidative damage. This aspect is highly significant in cancer chemoprevention, as oxidative stress contributes critically to DNA damage,

mutations, and the eventual formation of tumours via mechanisms such as radical scavenging and binding of divalent cations involved in Fenton chemistry [81]. A diet rich in polyphenols can enhance overall health and decrease cancer risk by acting as natural antioxidants. These dietary compounds may also help prevent the harm that unstable radicals cause and inhibit the initiation phase of carcinogenesis [82].

The carob fibre, obtained from the pods of *Ceratonia siliqua* L. (macromolecular matrices), can interact with polyphenols, permitting them to enter the colon, where they maintain the health of the gut and have positive effects on the digestive system [83]. Therefore, by limiting oxidative damage caused by free radicals, natural antioxidants such as those present in *Ceratonia siliqua* L, together with its fibre, may contribute to lowering cancer risk and supporting CRC management.

Carob fiber and phenolic compounds can bind together via covalent bonds like ester bonds that bind phenolic acids to polysaccharides, hydrophobic interactions, or hydrogen bonding (between the hydroxyl groups of polyphenols and the oxygen atoms of glycosidic linkages in polysaccharides, i.e., carob fiber) [84]**(Figure 2.5).**

Despite these well-established mechanisms, most studies on *Ceratonia siliqua L.* have focused on general antioxidant and chemopreventive effects, with limited investigation into their potential to directly target key molecular drivers of colorectal cancer such as EGFR and HER2.

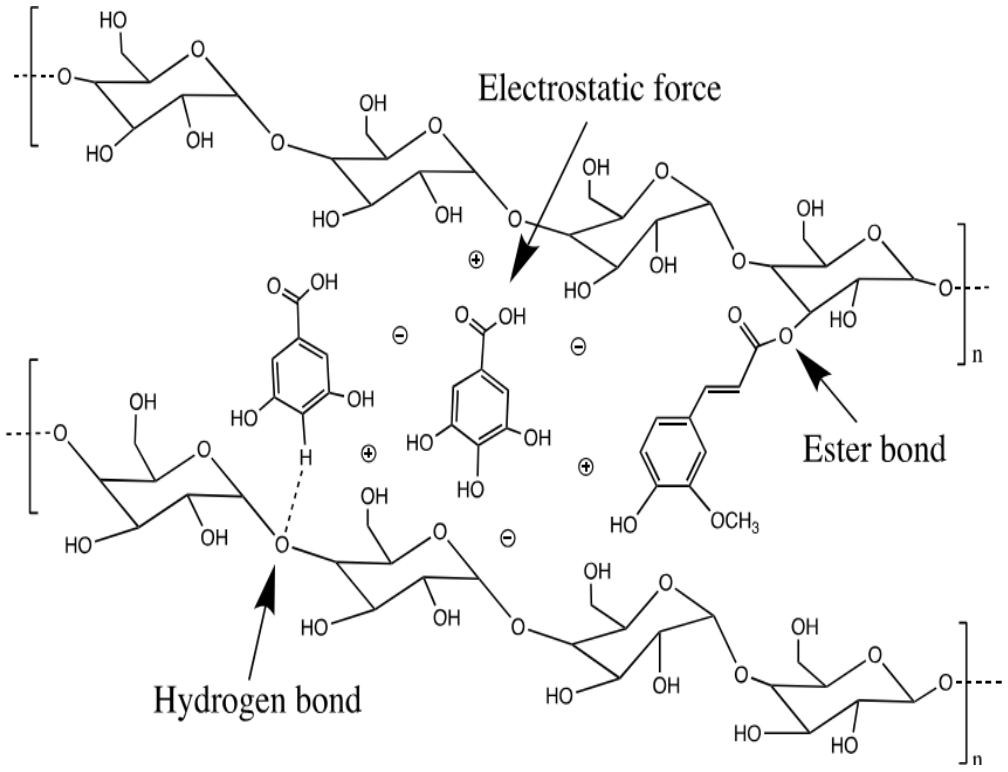


Figure 2.5: Different ways that phenolic chemicals and dietary fibre interact.

2.4. Anti-cancer activities

Cancer occurs as a genetic disorder marked by the abnormal cell division that originates from a single normal cell that has undergone malignant change. This abnormal cell division accelerates rapidly, spreading uncontrollably and reaching other cell types through body fluids like lymph and blood [12]. When cancer cells invade surrounding tissues, they can migrate to distant regions of of the body or adjacent lymph nodes, resulting in cancerous growths, a process referred to as metastasis or carcinogenesis [85-86].

Carcinogenesis is a multistage sequence that involves three primary phases: initiation, promotion, and progression. These phases are crucial for predicting the outlook of a given cancer type, with

prevention of initiation being the most vital strategy, as it influences the subsequent phases of the development of cancer [87]. The promotion phase represents a crucial point in the management of cancer. If effective inhibitory approaches are discovered to halt or limit tumour spread and if the timing is right, this stage might still be reversed. Therefore, using natural products is considered to be a good way to control tumor invasion and prevent or reverse carcinogenesis [88]. Medicinal plants provide a wide variety of natural products, modified synthetic derivatives, and secondary metabolites such as alkaloids, coumarins, flavonoids, tannins, terpenoids, quinones, phenolic acids, lignanolides, polyphenols, and steroids, all of which have demonstrated strong antioxidant activity and thereby contribute to cancer prevention through anti-inflammatory, antitumor, antimutagenic, and anti-cancer effects [12].

The development of cancer and its response to treatment are significantly influenced by inflammation [78]. In 1863, the German pathologist Rudolf Virchow was the first to propose that cancer originates in areas of persistent inflammation. Based on subsequent epidemiological findings, it is now understood that nearly 30% of all cancers may be linked to inflammation-driven causes, often arising from chronic infections or unresolved inflammatory conditions [89]. Advances in understanding the molecular mechanisms behind carcinogenesis have revealed key molecular targets involved in tumour initiation, progression, and development [87]. Among these, the family of epidermal growth factor receptors (EGFRs) is essential in regulating cell growth, survival, and division. Because abnormal activation of these receptors has been tied to several cancer types, they represent attractive therapeutic targets. Thus, focusing on these receptors offers a promising pathway for developing anti-cancer therapies, particularly for tumours associated with receptor tyrosine kinase (RTK) signalling [90]. While natural products have demonstrated broad anti-cancer activities, their specific application in targeting receptor-mediated pathways in

colorectal cancer remains underexplored. In particular, there is limited evidence on whether bioactive compounds from *Ceratonia siliqua L.* can effectively inhibit EGFR and HER2 signalling pathways, which are critical drivers of tumour growth and therapeutic resistance in CRC.

Overall, although *Ceratonia siliqua L.* has been widely studied for its antioxidant, anti-inflammatory, and anti-cancer properties, there remains a significant gap in understanding its potential in targeted cancer therapy. Specifically, there is a lack of studies integrating phytochemical profiling with advanced computational approaches, such as machine learning-driven virtual screening, to identify compounds with dual inhibitory activity against EGFR and HER2 in colorectal cancer. Addressing this gap forms the basis of the present study.

2.5. Epidermal growth factor receptors (EGFR)

The epidermal growth factor receptor (EGFR) family, also referred to as ErbB receptors, is made up of EGFR, HER2/ErbB2, ErbB3, and ErbB4, a group of receptor tyrosine kinases (RTKs) which have been recognized as key targets for treatment and oncogenic factors in a variety of human types of cancer. These receptors are abundantly exhibited in epithelial tissues, where they are crucial for regulating the survival, growth, and differentiation of cells through interconnected signalling cascades that include activation of the PI3K/Akt and MAPK/ERK1/2 pathways. The ErbB receptors operate by forming dimeric signalling complexes, which in turn activate multiple oncogenic pathways in cancer cells.

In colorectal cancer (CRC), EGFR is one of the most frequently implicated receptor tyrosine kinases. However, reported expression levels vary depending on methodological approaches such as immunohistochemical scoring thresholds and whether protein expression or gene amplification

is assessed. According to Yang et al. (2020), EGFR protein positivity has been reported in a wide range of approximately 25–80% of CRC cases, while gene amplification is observed in a smaller subset of about 5–10% of patients [90]. In contrast, HER2 expression is considerably less frequent in CRC compared to EGFR. Recent evidence consistently indicates that HER2 overexpression occurs in approximately 2–6% of colorectal cancer cases [91][92], while gene amplification is detected in a slightly higher proportion of patients, particularly in molecularly defined subgroups such as RAS/BRAF wild-type tumours [94]. For example, Ivanova et al. (2022) reported HER2 overexpression in approximately 5–6% of CRC cases, with gene amplification observed in about 7% of patients [93]. Similarly, Fraga et al. (2023) reported HER2 positivity in metastatic CRC ranging from 2–6%, increasing to 5–14% in selected molecular subtypes [94-95]. HER2 plays a much more prominent role in breast cancer, where its overexpression is reported in approximately 30% of early-stage cases and is considered one of the major molecular abnormalities linked to tumour development [96].

Overall, while EGFR is broadly expressed across a large proportion of CRC cases, HER2 is present in a smaller but clinically significant subset. This differential expression pattern supports the therapeutic rationale for targeting both receptors, particularly in strategies aimed at overcoming resistance mechanisms associated with EGFR monotherapy.

2.5.1. EGFR and HER2 inhibitors

Colorectal cancer is caused by abnormal cell growth and faulty signalling pathways that primarily regulate how human cells grow and divide. Two key proteins, EGFR (epidermal growth factor

receptor) and HER2 (human epidermal growth factor receptor 2) are mainly associated with these signalling networks [90].

The EGFR antagonists like cetuximab and panitumumab, inhibit EGFR-driven signalling, which is essential for the growth of tumour. However, innate or acquired resistance mechanisms, including mutations in downstream effectors such as the Kirsten rat sarcoma 2 viral oncogene homolog (KRAS) or the triggering of alternative signalling cascades, frequently limit their therapeutic success [97]. Similarly, HER2 is another well-recognised target in colon cancer therapy, particularly in cases where anti-EGFR treatments are ineffective [92], [98]. In metastatic colorectal cancer (mCRC; defined as colorectal cancer that has spread beyond the primary site to distant organs), HER2 amplification has been recognised as a prognostic biomarker that is negative for anti-EGFR therapies; in such instances, drugs targeting HER2, including trastuzumab and lapatinib, have shown therapeutic potential. Furthermore, dual HER2 blockade has demonstrated effectiveness in patients with a specific genetic profile (KRAS wild type) who fail to respond to standard treatment options [99-100].

2.6. Limitations of recent advancements in colorectal cancer (CRC) treatment

Colorectal cancer is one of the leading causes of deaths due to cancer globally, as reported by the Global Cancer Observatory (GCO) in 2024, formerly known as Global Cancer Incidence, Mortality, and Prevalence (GLOBOCAN) [101]. Although recent developments in molecular-targeted therapies for CRC such as anti-EGFR antibodies, anti-vascular endothelial growth factor (anti-VEGF) antibodies, and HER2-directed antibody drugs and antibody–drug conjugates have

led to improved outcomes, their clinical benefits remain limited to just a small percentage of patients [102].

A major obstacle in CRC treatment is the limited effectiveness of monotherapies aimed at EGFR or HER2, which often provide only short-term benefits due to mutations in downstream effectors like KRAS or the triggering of alternative signalling pathways that sustain tumour survival and growth [21]. In addition, cancer cells frequently acquire resistance to single-agent therapies through multiple mechanisms, further diminishing the success of long-term treatment [103]. The utility of EGFR- or HER2-targeted therapies is also constrained by the complexity and heterogeneity of CRC, including variability in receptor expression, making it difficult to identify the patients who are most likely to react [90]. These limitations emphasise the pressing need for innovative and more effective strategies to identify compounds capable of overcoming resistance mechanisms and enhancing therapeutic outcomes.

2.7. Virtual screening (VS) of small molecules

To address the above-mentioned limitations and the challenges of the traditional drug discovery process, Virtual screening (VS) is growing as an invaluable tool for finding novel therapeutic options, including those targeting EGFR and HER2 in CRC. Virtual screening is the technique of prioritising and ranking compounds in vast chemical libraries based on how likely they are to interact with a certain target. Virtual screening is applied to forecast the binding potential of extensive ligand databases against a target to identify compounds with promise for further evaluation [104].

Virtual screening approaches are generally classified into two main groups: structure-based virtual screening (SBVS) and ligand-based virtual screening (LBVS) [105].

- **Structure-based virtual screening (SBVS):** This technique is used when the target protein's three-dimensional (3D) structure is identified. The primary sources of such structural information are X-ray crystallography, nuclear magnetic resonance (NMR), and homology modelling. In SBVS, molecular docking serves as the central approach for examining candidate compounds (potential ligands) by predicting their binding affinity for the specific 3D binding pocket of the biological target [106].
- **Ligand-based virtual screening (LBVS)**: This approach does not depend on the knowledge of the target protein 3D structure but rather on identifying potential drug candidates similar to known active molecules (ligands) based on their chemical structure and properties. It is based on the principle that molecules that are similar in structure to known active compounds are likely to have similar biological activities in protein-ligand interactions. Therefore, it relies on data obtained from previously identified active compounds (ligands) to investigate and discover new compounds that share similar chemical and pharmacological characteristics. The LBVS techniques fall into three categories: 3D shape-based analysis, drug modelling, and 2D fingerprint matching searches [95].

Molecular docking is one of the core techniques in SBVS where virtual screening software simulates the interaction between small molecules, such as ligands, and the binding site of the target protein (such as EGFR and HER2), with the objective of predicting the best fit between the ligand and the protein, and prioritising ligands with high binding activity. Using molecular docking, SBVS ranks compounds by predicting their binding affinity and orientation for a target

protein, thus enabling researchers to virtually screen for molecules that may inhibit cancer-related receptors like EGFR and HER2 [107].

2.8. Machine learning (ML)

With its improved prediction abilities for drug discovery, machine learning has become a vital extension of virtual screening. Machine learning algorithms are highly complementary to both SBVS and LBVS techniques because they can effectively predict the biological properties, binding affinity and the toxicity of small molecules by recognising patterns in existing database information. Within computer-aided drug development, machine learning has become one of the most important and quickly developing fields [2], [11], [108]. The discipline integrates concepts from computer science, mathematics, statistics, and engineering [109]. It is widely regarded as a key application of artificial intelligence (AI), enabling computers, software, and devices to function through cognitive processes [110]. Machine learning relies on algorithms that process and analyse raw data to achieve defined objectives [111]. By applying ML approaches, it becomes possible to develop mathematical connections based on actual observations of small molecules in order to forecast the chemical, biological, and physical properties of novel compounds. The four main categories of machine learning techniques are reinforcement learning, supervised learning, unsupervised learning, and semi-supervised learning. In practice, ML is most often applied through supervised and unsupervised learning methods. Supervised learning techniques are designed to build training models capable of predicting future values of data classes or continuous variables, whereas unsupervised techniques, which allow data grouping without user labelling prior to use, are mostly employed for exploratory research [111-112].

2.9. Theoretical framework

2.9.1. Machine learning models

The application of machine learning (ML) methodologies for model construction from data has witnessed substantial growth in recent years. Effective model development necessitates a thorough comprehension of the actual dataset and an extensive knowledge of the available ML algorithms and their respective characteristics. Depending on their purpose, ML models may be designed to facilitate a deeper interpretation of data or to generate predictive outcomes. Irrespective of their role, however, rigorous evaluation remains one of the most critical stages in the model development process.

This process ascertains whether the objectives of the modelling activity have been met; it facilitates the comparison of various modelling techniques and guides subsequent research endeavours [113]. Some of the ML models are described below:

i. Logistic regression (LR)

Among the supervised machine learning techniques is logistic regression which is employed to handle binary-class classification problems and is useful for modelling the odds or probability of an event occurring [114]. In order to predict a variable with two possible values (like 0 or 1), there must be an attempt to predict the likelihood that the dependent variable has a value of zero or one. The following equation is used:

$$P(Y_i = 1) = 1 / 1 + e^{-(b0+b1X1i+...+bpXpi)} \quad \text{Equation (2.1)}$$

where $P(Y_i = 1)$ is the probability that the dependent variable, Y_i, will take the value of 1 for a given observation, i (this represents the outcome of interest); e is the base of the natural logarithm, approximately 2.71828; b_0 is the intercept of the model; X_{1i}, …, X_{pi} are the independent variables for observation i; $b_1,\ldots,b_p$ are the regression coefficients associated with each independent variable; and $-(b_0 + b_1X_{1i} + \ldots + b_pX_{pi})$ is the linear combination of the predictors [115]. These independent variables may be a range of features relevant to the study, such as molecular descriptors, physicochemical properties, or other pertinent factors. The logistic distribution for the

$$P\,(Y_i = 1) = \exp(\alpha j + \beta X) \,/\, (1 + \exp(\alpha j + \beta X)) \qquad \text{Equation (2.2)}$$

cumulative probability is shown by Equation (2.2 [116]. In ordinal logistic regression, the cumulative probability of a response less than or equal to j is given as:

where $P(Y \leq j)$ is the cumulative probability that the ordinal response y is less than or equal to category j; α_j is the threshold or intercept specific to category j, which varies for each category; βX represents the effect of the predictor variables (X) on the cumulative probability; and exp( ) is the exponential function [115]. The logistic regression equation (Equation (2.1)) can be rearranged to the following equation known as the binary logistic regression model [115]

$$\ln\left(\frac{P\,(Y_i = 1)}{1 - P\,(Y_i = 1)}\right) = b_0 + b_1X_{1i} + b_2X_{2i} + b_3X_{3i} + \ldots + b_kX_{pi} \qquad \text{Equation (2.3)}$$

where ln is the natural logarithm; $P(Y_i = 1)$ is the probability that the dependent variable Y_i for the ith observation equals 1; $1 - P(Y_i = 1)$ is the probability that the dependent variable Y_i equals 0; b_0 is the intercept or baseline log-odds when all predictor variables are zero; b_1, b_2, $b_3 \ldots b_k$ are the

coefficients for each independent (predictor) variable; and X_{1i}, X_{2i}, X_{3i}…X_{pi} are the independent (predictor) variables for the *i*th observation.

ii. *k*-Nearest Neighbours (*k*NN)

The k-nearest neighbours (*k*NN) algorithm is a supervised machine learning technique that is mostly used for classification applications. The *k*NN approach functions as an "instance-based learning" algorithm [117]. This form of lazy or memory-based learning estimates local functions and postpones computation until the classification stage. The *k*NN algorithm serves as a classification or regression method without parameters that organises unidentified occurrences in the feature space according to the k training examples that are closest to it (k should ideally be a small positive integer). The class of the nearest single neighbour is assigned to the unknown occurrence if k = 1 [118]. Euclidean distance is the most commonly used distance metric of the *k*NN algorithm which represents the shortest straight-line distance between two points (new point and other points to be calculated), expressed as follows (Equation (2.4) [119] **(Figure 2.6).**

$$d(p,q) = \sqrt{\sum_{i=1}^{n}(q_i - q_i)^2}$$

$$d(p,q) = \sqrt{(q_1 - q_1)^2 + (q_2 - q_2)^2 + \cdots (q_n - q_n)^2} \qquad \text{Equation (2.4)}$$

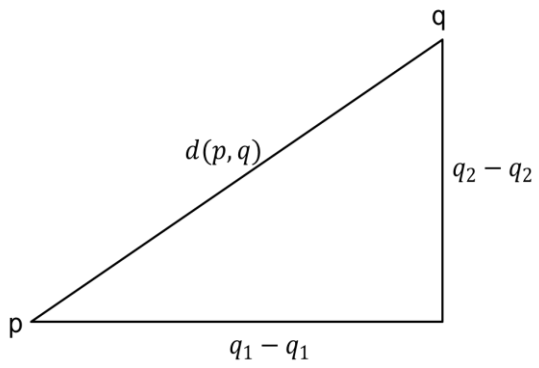


Figure 2.6: Examples of calculating the straight-line distance between two points to identify the neighbours

iii. Support Vector Machine (SVM)

The Support vector machine (SVM) algorithm is also a popular machine learning technique for problems like regression and classification, investigating data, and identifying patterns. It is utilised for both linearly separable data (where classes can be divided using a straight line or hyperplane) and nonlinear data [120], where special kernels such as the radial basis function (RBF) are employed to project data into higher-dimensional space for improved separation [121].

The SVM algorithm has been successfully used across diverse disciplines such as bioinformatics (e.g., classifying protein structures), chemoinformatics (e.g., predicting drug–target interactions), and medical diagnostics (e.g., classifying patients based on gene expression profiles) [122-123]. In drug discovery, SVM models are particularly useful in forecasting the small compounds' biological effect against defined targets like HER2 and EGFR based on molecular descriptors and fingerprints [124].

When datasets contain considerable noise, including overlapping classes, SVM tends to perform poorly [125]. The following equation is a feasible way to predictions in SVM classification [126]:

$$h(x_{\mathrm{i}}) = sign(\sum_{j=1}^{s} \alpha_{\mathrm{j}}\, \gamma_{\mathrm{j}}\, K(x_{\mathrm{j}}, x_{\mathrm{i}}) + b)$$

$$K(v, v') = exp(\frac{||\, v - v'||^2}{2\gamma^2}) \qquad \text{Equation (2.5)}$$

where x_{i} is the vector of the values that are to be predicted, x_{j} represents the support vectors, a subset of training data, γ_{j} is the class label (e.g., +1 or -1 in binary classification) associated with each support vector x_j, α_j are non-negative constants associated with each support vector x_j, b is a bias term, a scalar value that shifts the decision boundary, and s is the number of support vectors [126].

iv. Decision Tree classifier

The Decision Tree (DT) is a machine learning technique that works well for both classification and regression. It functions by progressively dividing the dataset into a hierarchical, tree-structured set of decisions. The tree is constructed through internal or decision nodes (which represent points where input features are evaluated) and leaf nodes (which denote the final predicted outcome or value) [127].

The root node is always located at the top the structure, which identifies the feature that best partitions the dataset based on a selected criterion, like Gini or information gain. From there, the dataset is recursively divided through successive decision nodes, where features are tested under varying conditions. This recursive process continues until the tree reaches a leaf node, which

provides the final prediction [128-129] **(Figure 2.7**Decision Trees can handle both continuous and categorical data, making them incredibly flexible for any set of data. This versatility makes them effective for applications such as identifying molecular characteristics relevant to therapeutic efficacy, categorizing compounds according to their biological activity, and forecasting drug–target interactions [130]. **Figure 2.7** below illustrates the relationship between decision nodes and leaf nodes within a DT [129].

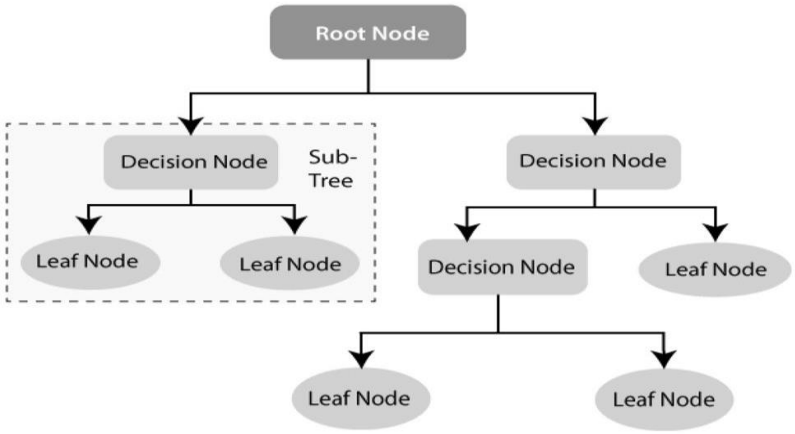


Figure 2.7: Illustration of a Decision Tree

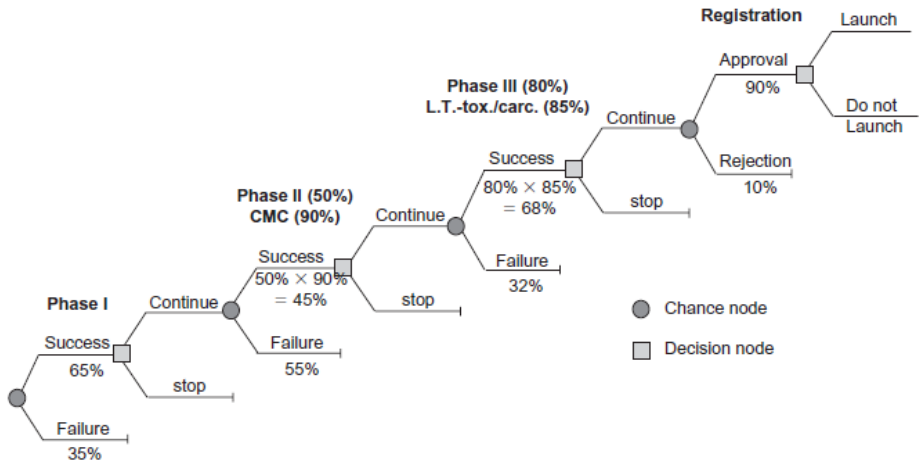


Figure 2.8: Example of a Decision Tree used in evaluating a drug candidate for clinical development [131]

A study conducted by Bode-Greuel and Greuel (2005) further demonstrated the use of Decision Trees in clinical drug research, as shown in **Figure 2.8** [131]. Decision Trees are designed to capture uncertain outcomes that influence managerial actions and ultimately affect project value. Data from independent studies conducted in parallel, such as clinical trials and long-term animal toxicology, are integrated into development milestones, which illustrate the project's optionality. In the diagram, chance nodes denote possible outcomes, whereas decision nodes represent managerial choices. Each Decision Tree is modelled independently, with variations in the number of decision points and overall complexity. At each development milestone, clinical probabilities are combined with chemical and manufacturing control (CMC) probabilities to estimate the overall likelihood of success.

v. Random Forest (RF)

Random Forest (RF) is an ensemble learning method that combines several Decision Trees [125]. This integration lowers classification and regression errors by employing bootstrap aggregation, also known as bagging. Random Forest is a supervised and straightforward technique (A group of Decision Trees combined) that is efficient and resilient to noise in the target dataset; RF generally achieves high precision and avoids overfitting in most cases, as it mitigates bias by averaging across all predictions [132]. The core principle of Random Forest is to improve predictive accuracy taking into account the collective Decision Trees within the forest along with the correlation among their outputs. The diagram below **(**Error! Reference source not found. depicts a random f orest classifier's structure constructed from multiple Decision Trees [129].

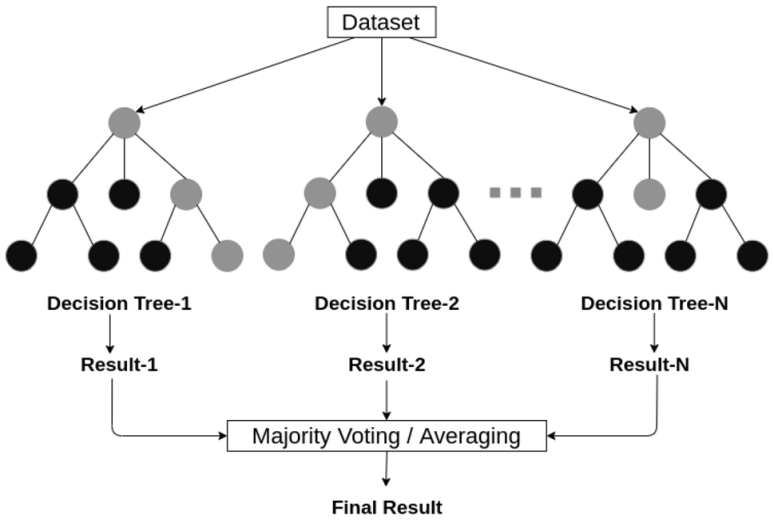


Figure 2.9: Random Forest Decision Tree example

vi. Naïve Bayes

The Naive Bayes (NB) algorithm is developed from Bayes' theorem, its core assumption being that all features used for prediction are independent [133]. It performs effectively in classification tasks that is both binary and multi-class categories across diverse applications, including text

categorization, spam detection, and related domains. The NB algorithm is particularly useful in classifying noisy data and in building reliable prediction models. Its main advantage resides in the fact that it just needs a relatively small dataset to quickly estimate the necessary parameters [133]. However, NB is highly sensitive to feature selection, which can adversely affect its predictive accuracy [134].

vii. Gradient Boosting Machine (GBM)

This ensemble technique called the Gradient Boosting Machine (GBM) uses Decision Tree models for both classification and regression problems. The final prediction model of GBM is created by combining the outputs of several trees, with the accuracy of each tree being improved through the boosting process. Gradient boosting further refines this process by simplifying tree boosting, thereby enhancing interpretability and computational efficiency [135].

viii. AdaBoost classifier

Adaptive Boosting (AdaBoost) is a method of ensembling that frequently uses a basic Decision Tree, sometimes known as a decision stump. The process starts by building a stump and giving the data points weights. Subsequently, while the weights of correctly categorized points are decreased, those of misclassified occurrences are increased. Data points with higher weights are given greater influence in subsequent models, allowing AdaBoost to iteratively refine and enhance the predictive performance of the initial stump [136].

Freund and Schapire (1997) [137] established the process for the AdaBoost as shown below:

1. *Initialize the observation weights* $\mathrm{w_i} = {}^{1}\!/_{n}\ , i = 1,2,\ldots,n$

2. $for\ m = 1\ to\ M$

$(a)\ Fit\ a\ classifier\ T^{(m)}(x)\ to\ the\ traning\ data\ using\ weights\ \mathrm{w_i}$

$(b)\ Compute$

$$err^{(m)} = \sum_{i=1}^{n} \mathrm{w_i}\, \Pi(\mathrm{c_i} \neq T^{(m)}(x_i))/\sum_{i=1}^{n} \mathrm{w_i}$$

$(c)\ Compute$

$$\alpha^{(m)} = \log\frac{1-err^{(m)}}{err^{(m)}}$$

$(d)\ Set$

$$\mathrm{w_i} \leftarrow \mathrm{w_i}.\, exp\big(\alpha^{(m)}.\, \Pi(\mathrm{c_i} \neq T^{(m)}(x_i))\big)$$

$for\ \ i = 1,2,\ldots,n$

$(e)\ Re-normalize$

3. $Output$

$$C(x) = argmax \sum_{m=1}^{M} \alpha^{(m)}.\, \Pi(T^{(m)}(x_i) = k) \qquad \text{Equation (2.6)}$$

2.10. Machine learning model evaluations

Several metrics may be applied when assessing the effectiveness of machine learning models based on the specific issue and the desired outcome. Common evaluation measures such as accuracy, precision, recall, F1 score, and the area under the receiver operating characteristic curve (AUC-

ROC) can be used to evaluate how well various models work and to determine what needs to be improved.

Accuracy is one of the most frequently adopted performance measures in ML models, because it is intuitive, easy to compute, and often sufficient when the dataset's distribution across classes is uniform, meaning every class has a similar number of examples. It describes the proportion of cases that were accurately anticipated compared to all the predictions. It performs effectively only when classes contain roughly the same number of samples. Precision represents the proportion of accurately anticipated positive cases to all expected positives (false positives plus genuine positives). Precision becomes vital in situations where false positives have a significant impact (e.g., spam detection or fraud detection). Recall, also known as sensitivity, show the share of accurately predicted positives among all actual positive cases and is important when minimising false negatives is a priority.

The F1 Score serves as a recall and harmonic mean of precision. It serves as a metric of the predicted precision of a model. The score goes from 0 to 1, where 1 denotes excellent recall and precision and 0 denotes total failure to detect positive instances. It provides insight into both the precision of a classifier (how many predictions are correct) and its robustness (its ability to avoid missing a substantial number of positive cases).

For binary classification tasks, the formulas for accuracy, precision, recall, and F1 Score are expressed as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$ Equation (2.7)

$$Precision = \frac{TP}{TP+FP}$$ Equation (2.8)

$$Recall = \frac{TP}{TP+FN}$$ Equation (2.9)

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$ Equation (2.10)

where:

TP (True Positive) = accurately predicted positive cases.

TN (True Negative) = accurately forecasted negative occurrences

FP (False Positive) = inaccurately anticipated positive cases

FN (False Negative) = instances of negative predictions that are not accurate.

Selecting the appropriate metric depends on the problem context. For example, in fraud detection, precision may be prioritised to avoid falsely accusing users. In cancer screening, recall may be prioritised to ensure that no actual cancer cases are missed. In balanced classification problems, accuracy or F1 score may provide a good summary of performance.

2.11. Limitations of Model Validation

Although the machine learning models developed in this study demonstrated strong predictive performance during internal validation, several limitations must be acknowledged when interpreting these results. The predictive reliability of the models is limited to the chemical space defined by the training dataset. Compounds that fall outside this applicability domain, particularly structurally novel phytochemicals, may not be predicted with the same level of accuracy. This is an inherent limitation in ligand-based virtual screening approaches, where model generalisability depends on the diversity and representativeness of the training data [138].

Extreme care taken to minimise data leakage during model development; however, there remains a potential risk when feature preprocessing, scaling, or fingerprint generation is not strictly separated between training and test sets. Data leakage can lead to overly optimistic performance estimates, thereby inflating model accuracy and reducing reliability in real-world applications.

The models were validated using internal cross-validation strategies, which provide an estimate of predictive performance within the dataset. However, external validation using completely independent datasets was not performed. This limits the ability to fully assess the generalisability of the model to unseen chemical spaces, such as novel natural products or clinically tested compounds. Therefore, these limitations highlight that while the models are suitable for prioritising compounds within the current virtual screening framework, further validation using external datasets and experimental confirmation is essential before clinical or translational applications.

2.12. Ensemble machine learning techniques: Stacking and bagging

Ensembles are effective machine learning techniques in improving the model predictive performance since they utilise the ability of more than one classifier [139]. Ensemble machine learning utilises multiple individual models to achieve better predictive performance than any single model; it mainly follows two strategies: bagging and stacking [140]. Bagging (bootstrap aggregating) is used to train multiple versions of the same type of model on different bootstrap samples of the original dataset; the predictions of these individual models are then combined to produce the final prediction, using averaging for regression tasks and voting for classification tasks. These models typically employ different classifiers, feature sets, and hyperparameters [141]. Ensemble methods, including bagging, boosting, and stacking [142], outperform single learners by leveraging diversity among models, thereby benefiting from the variance in Decision Trees [143].

Stacking differs from bagging by combining predictions from diverse (heterogeneous) base models using a meta-model. Stacking represents an ensemble learning strategy in which a separate ML algorithm is trained to combine multiple models' predictions. These ML algorithms are known as base learners or base models. It consists of training several base-level models (level-0) which make predictions and then apply a meta-learning model (level-1) that combines their predictions into a single final prediction. This approach improves overall performance by utilizing the advantages of multiple models [140]. According to Schaduangrat et al. (2022), the downside of this technique is that the selection of the base-learner level and the meta-learner models affects the classification's performance [144].

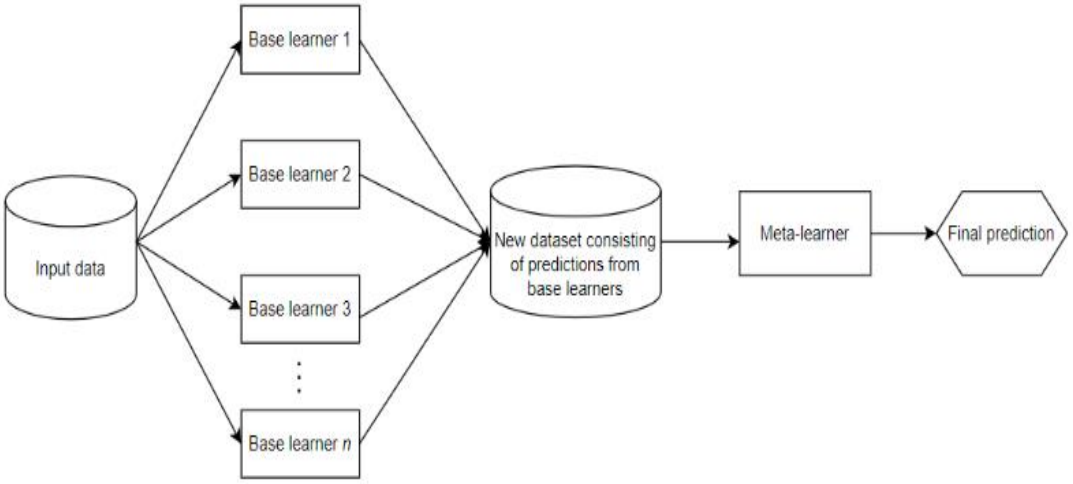


Figure 2.10: Block diagram of the stacking framework

Stacking ensemble learning was selected in this study due to its ability to integrate multiple heterogeneous base learners and improve predictive performance, particularly in complex and imbalanced bioactivity datasets [145]. Unlike bagging, which reduces variance through homogeneous models, or boosting, which sequentially focuses on misclassified samples, stacking combines diverse algorithms and leverages a meta-learner to optimise final predictions. This is particularly advantageous in bioactivity classification tasks, where class imbalance between active and inactive compounds can bias single-model performance.

The stacking approach enhances generalisation by allowing different base learners (e.g., tree-based models, linear models, and distance-based classifiers) to capture complementary patterns in molecular descriptor space. The meta-learner then learns how to best combine these outputs, resulting in improved stability and predictive accuracy compared to individual models or simpler ensemble methods.

2.13. Research gap in literature

Despite extensive research on natural product phytochemistry and the development of machine learning-based virtual screening approaches, these two fields have largely evolved independently. Most phytochemical studies on *Ceratonia siliqua L.* and similar medicinal plants have focused on compound identification and biological activity validation, without integrating computational prediction models for systematic drug prioritisation. Similarly, existing machine learning-based virtual screening studies have predominantly relied on publicly available chemical databases such as ChEMBL or ZINC, with limited incorporation of experimentally derived LC–MS/MS datasets from plant extracts. This creates a significant gap in linking experimentally identified natural compounds directly to predictive bioactivity modelling. Furthermore, few studies have applied ensemble machine learning approaches, such as stacking, in combination with LC–MS/MS-derived phytochemical datasets for the specific identification of dual-target inhibitors, particularly against EGFR and HER2 in colorectal cancer. As a result, the potential of *Ceratonia siliqua L.* as a source of computationally prioritised multi-target anti-cancer compounds remains underexplored. Therefore, this study addresses this gap by integrating LC–MS/MS phytochemical profiling with a stacking ensemble machine learning framework to enable the virtual screening and prioritisation of bioactive compounds with dual EGFR/HER2 inhibitory potential.

CHAPTER 3: METHODOLOGY

PREAMBLE

The present chapter offers a detailed outline of the methodology employed in this research project. The chapter is divided into three parts (experimental laboratory procedure, machine learning and computational method, and experimental validation), thereby highlighting the application of stacking ensemble models for virtual screening using data from liquid chromatography–tandem mass spectrometry (LC–MS/MS) analysis. It explains how LC–MS/MS analysis, machine learning (ML), and virtual screening (VS) were integrated to identify and validate HER2 and EGFR inhibitors for colorectal cancer treatment. It focuses on natural compounds from *Ceratonia siliqua* L. (carob tree), exploring their anti-cancer potential through both computational predictions and experimental validations. In addition, it discusses the anti-cancer and antioxidant properties of *Ceratonia siliqua* L. pod extracts, supporting their therapeutic relevance. This also represents a translational link between our computational framework and preliminary biological evidence, providing experimental support for the model's predictions. Molecular docking and computational ADMET (absorption, distribution, metabolism, excretion, and toxicity) analyses were performed to compare the predicted active compounds with standard FDA-approved drugs **(Figure 3.1)**.

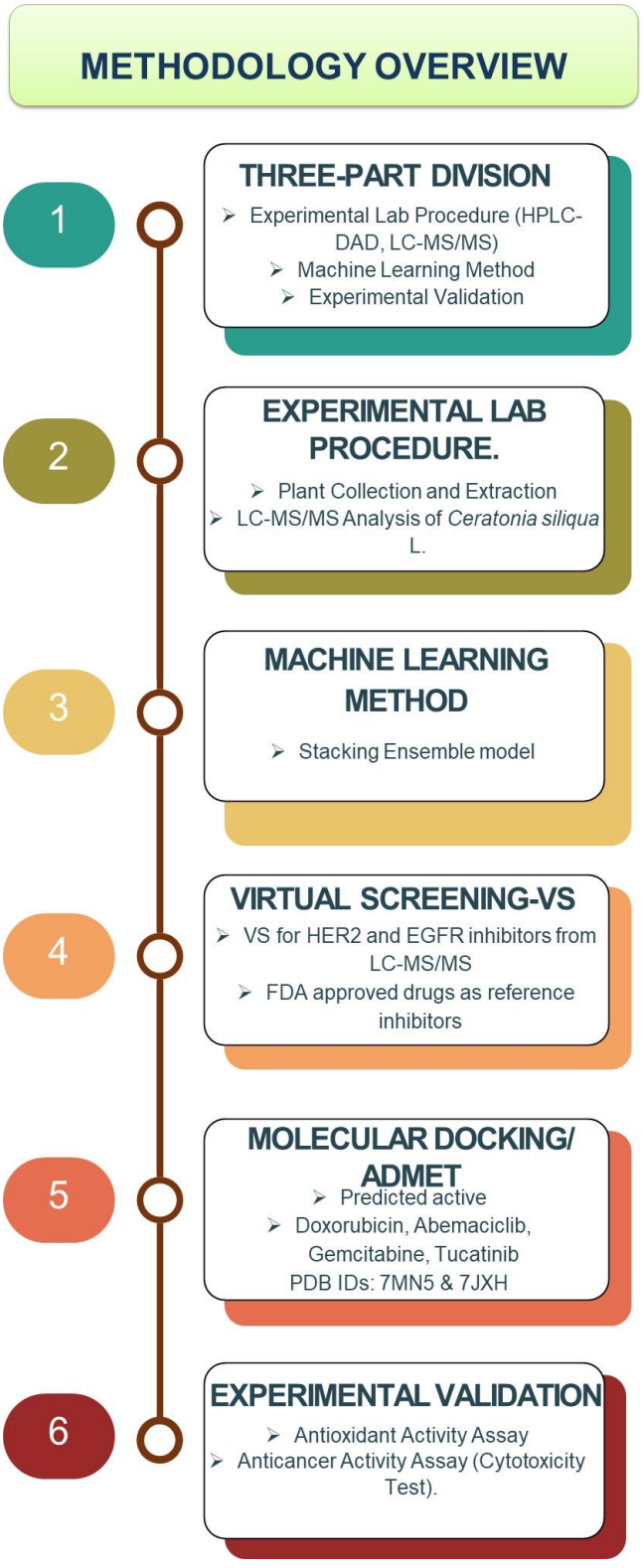


Figure 3.1: Flow diagram depicting the methodology followed in this study

3.1. Experimental procedure

3.1.1. Collection and identification of *Ceratonia siliqua* L. pods

Fresh samples of *Ceratonia siliqua* L. pods were collected in November 2024 from the Kazimingi Nursery Farm in Benoni (S26° 04′ 30.36″ | E28° 19′ 32.99″), located in the Gauteng Province of South Africa. Dr R Munyai at the College of Agriculture and Environmental Sciences Horticulture Research Centre, University of South Africa (UNISA), identified the specimen, and a voucher specimen bearing the number UNISA/CAES/2024/SB0054 has been deposited at the UNISA herbarium for reference. The plant specimens were allowed to naturally desiccate indoors for 14 days at an approximate ambient temperature of 20 °C. The dehydrated pods (seed and pulp) were subsequently milled into fine powder using an electric blender and kept at room temperature in clear, labelled polyethylene bags that were protected from light until required.

3.1.2. Preparation of Crude Extract from Ceratonia siliqua L. Pods

The crude extract of *Ceratonia siliqua* L. pods were obtained by cold maceration, whereby 250 g of the plant material (pods) in powder form was macerated in 1 000 mL (1:1 ethanol–water mixture; 500 mL of ethanol + 500 mL of water) for 72 h. Subsequently, the crude extracts were filtered through Whatman filter paper using a vacuum Buchner funnel, and the resulting filtrates were concentrated under reduced pressure with a rotary evaporator at 60–70 °C. After that, the concentrated extracts were put into glass vials that had been previously weighed and allowed to dry air at room temperature to maintain a consistent weight, then kept in the dark at 4 °C until they were needed again.

3.1.3. Instrumentation

3.1.3.1 High-performance liquid chromatography- diode array detector (HPLC-DAD)

Compounds in sample extracts were separated, identified, and quantified using an Agilent 1260 Infinity high-performance liquid chromatography (HPLC) setup from Agilent Technologies (Waldbronn, Germany), equipped with a a dual high-pressure pump, autosampler, temperature-controlled column chamber, diode array detector, and a fluorescence detector. Instrument control, data collection, and processing were managed using the Agilent OpenLab ChemStation software (version 3.5.77).

3.1.3.2 High performance liquid chromatography–tandem mass spectrometry (HPLC-MS)

The LC–MS/MS analysis was performed on a Thermo Scientific™ Q Exactive™ Plus Hybrid Quadrupole-Orbitrap™ Mass Spectrometer, coupled with a Thermo Scientific™ Dionex UltiMate™3000 UHPLC system (Thermo Fisher Scientific, Waltham, MA, USA). The system was equipped with a heated electrospray ionisation (HESI) source for efficient ion generation. The instrument was operated in an untargeted acquisition mode, using a combination of full MS/MS, single ion monitoring (SIM), and all-ion fragmentation (AIF) to enable comprehensive metabolite profiling of the extract. Source parameters were optimised to suit the HPLC flow rate, including a capillary temperature of 285 °C, spray voltage of 3.5 kV, sheath gas flow of 50 arbitrary units, and auxiliary gas temperature of 400 °C. Positive ionisation mode was applied with a mass range of 100–1 500 m/z, selected to ensure adequate coverage of both low- and high-molecular-weight phytochemicals typically present in natural product extracts. The instrument operated at a

resolution of 70 000 full width at half maximum (FWHM), with an automatic gain control (AGC) target of 1.0×10^6 and a maximum injection time of 100 ms. Mass calibration was performed using Thermo Scientific™ Pierce™ calibration standards in both positive and negative ion modes. Data acquisition, instrument control, and processing were conducted using Thermo Scientific™ TraceFinder™ software (version 5.1).

3.1.4. HPLC-DAD method development

The concentrated extract of *Ceratonia siliqua* L. pods (1 mg) was initially dissolved in 100 mL of a water and acetonitrile mixture (1:1 v/v) and was then filtered using a 0.45 μm PVDF membrane filter prior to analysis. The development of a high-performance liquid chromatography with diode array detection (HPLC-DAD) method was undertaken to achieve optimal separation and quantification of phytochemicals present in the *Ceratonia siliqua* L. pod extracts. The compounds in the filtered crude pod extract were separated using an XTerra MS C18 analytical column (100 × 4.6 mm, 3.5 μm; Waters Corporation, Milford, MA, USA). The mobile phase consisted of solvent A (0.1% v/v formic acid in water) and solvent B (0.1% v/v formic acid in acetonitrile). Initially, the method was tested using isocratic elution with three consecutive injection sequences; however, this approach resulted in poor resolution and overlapping peaks. To improve separation efficiency, gradient elution was introduced and systematically optimised.

A gradient elution mode was used as follows: 0–3 min, 2–5% B; 3–10 min, 5–15% B; 10–16 min, 15–30% B; 16–22 min, 30–60% B; 22–26 min, 60–90% B; 26–28 min, 90–5% B; and 28–32 min, 5–5% B. The flow rate was set at 0.9 mL/min, the column temperature was maintained at 25 °C, and the injection volume was 5 μL. Detection was performed using a diode array detector (DAD),

monitoring primarily at 254 nm and 210 nm. The HPLC-DAD analysis was performed primarily as a method development and optimisation step to support the subsequent LC–MS/MS analysis. It was used to evaluate chromatographic separation efficiency, assess extract complexity, and guide gradient optimisation to ensure suitable resolution of phytochemical constituents prior to high-resolution mass spectrometric profiling.

3.1.5. LC–MS/MS analysis of the *Ceratonia siliqua* L. crude pod extract

3.1.5.1 LC–MS/MS analysis and data processing

The method developed using the HPLC-DAD was subsequently transferred to an LC–MS/MS system for further detailed analysis. The LC–MS/MS data obtained was processed and interpreted through the Global Natural Products Social Molecular Networking (GNPS 2.0) platform. The network configuration and library search conditions were adjusted according to the standard parameters of the natural product analysis workflow. This workflow incorporated blank data subtraction, feature detection, determination of elemental composition, spectral library comparison, and MS/MS fragmentation pattern analysis. Compound identification was mainly performed through MS/MS spectral alignment with the following reference libraries: MoNA, MSNLIB-positive, MSNLIB-negative, NEO-MSMS, Berkeley-Lab, Birmingham-UHPLC-MS-POS, MASSBANK. GNPS-NIST14-MATCHES and GNPS-NIH.

3.2. Machine learning and computational method

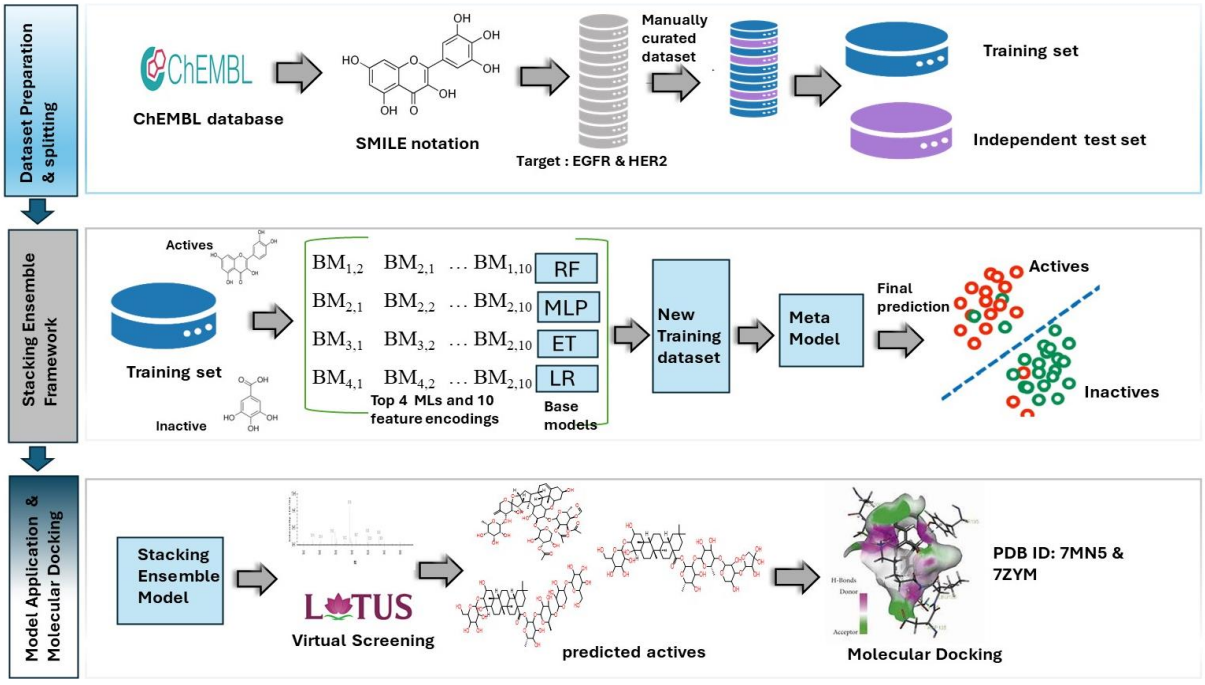


Figure 3.2: Diagram depicting the overall workflow for development of the stacking ensemble method comprising data preparation, splitting, model optimisation, construction, virtual screening and molecular docking.

3.2.1. Dataset compilation

A dataset of 21 991 unique compounds tested for activity against HER2 (Target ID: CHEMBL1871) and EGFR (Target ID: CHEMBL203) was obtained from the ChEMBL database. Of the total compounds, **7** 165 were identified as HER2 inhibitors, while the remaining molecules were classified as EGFR inhibitors. The datasets comprised compound IDs and SMILES notations and were subjected to preprocessing and curation using Python code. The HER2 and EGFR datasets were pre-processed and curated independently and later merged to construct a unified dataset for dual-target modelling.

3.2.2. Data preprocessing

The IC_{50} (half-maximal inhibitory concentration) values, recorded as 'standard_value', were standardised to micromolar (μM) units. Entries containing ambiguous symbols such as '<', '>', or '/' were excluded to ensure data consistency, as these indicate uncertain values. Only standard_values with the '=' symbol were retained. Additionally, any entries with missing data in the 'standard_value' or 'canonical_smiles' column were removed from the dataset.

3.2.3. Data curation

Only compounds reported with IC_{50} (half-maximal inhibitory concentration) measurements were retained to build the final dataset, which contained 17,287 unique entries. These molecules were divided into two activity classes: those with IC_{50} values ≤ 1 μM were designated as *active* (positive class), whereas compounds exhibiting IC_{50} values ≥ 10 μM were labeled *inactive* (negative class).

The intermediate values between 1-10 μM ($1 \leq IC_{50} \leq 10$) were excluded from the dataset to ensure a clear binary classification between active and inactive compounds. To assess how the final trained stacking model handles intermediate IC_{50} values (1–10 μM), a virtual screening of the excluded compounds labelled as 'class = 2' was performed, and the predicted probabilities were analysed. The pairwise Tanimoto similarity index was then computed for all entries using Molecular Access System (MACCS) molecular fingerprints within the RDKit package [146]. Compounds showing identical fingerprints (Tanimoto index = 1) were eliminated, and for duplicated structures, the average IC_{50} value was retained to represent that molecule

3.2.4. Chemical space analysis

Eight key physicochemical properties, critical for assessing molecular complexity and drug-likeness, were calculated, visualised and distinguished between compound groups that are active and inactive. These included molecular weight (MW), the Ghose-Crippen-Viswanadhan octanol–water partition coefficient (ALogP), number of hydrogen bond acceptors (nHBA), number of hydrogen bond donors (nHBD), aromatic ratio (ARR), number of rings (nCIC), number of rotatable bonds (RBN), and number of benzene-like rings (nBnz). These properties are commonly referenced in Lipinski's Rule of Five (Ro5).

This was followed by an assessment of the mean, median, minimum, and maximum values. The *p*-values derived from the Mann-Whitney U test (at the threshold of $p < 0.001$) were used to establish statistical significance.

3.2.5. Molecular fingerprint

The machine learning model was developed using various descriptors, including molecular fingerprint-based features, with fingerprint parameters selected according to established literature [147]. The molecular structures were represented in SMILES format, which served as a binary encoding scheme to extract the structural characteristics of each compound. The RDKit (https://www.rdkit.org/) was used to implement the SMILES notation and to calculate ten molecular fingerprint descriptors (AP2D, CDK, CDKExtended, CDKGraph, KR, MACCS, Circle, E-State, Hybrid, and PubChem). Features for virtual screening were encoded using these molecular fingerprints. For example, atom pairs at various topological distances can be identified using the 2D Atom Pair (AP2D) fingerprint, generating 780 distinct features. The Chemistry Development Kit (CDK) fingerprint, on the other hand, creates a 2 048-bit representation based on a search depth of eight. Building on this, the CDK Extended (CDKExt) version incorporates additional bits to capture ring-specific characteristics (with radius = 3 with nBits = 2048), while CDK Graph Only (CDKGraph) simplifies the representation by focusing fully on molecular connectivity while disregarding bond order [148-149]. The Circle fingerprint is another known technique that encodes circular molecular features using a 2 048-bit structure and radius = 2 [150]. Additionally, the E-State fingerprint has a distinct focus on atomic properties and uses 79 bits with each bit corresponding to the presence or absence of a particular Electrotopological State (E-State) atom type or fragment within a molecule. Hybrid fingerprints generate a 2 048-bit (with radius = 3) encoding by combining both hybridisation and data from CDK [151]. With 2 048 chemical substructures encoded, the Klekota–Roth (KR) fingerprint also provides a detailed representation [152]. While PubChem fingerprints incorporate 881 features generated from substructure definitions in the PubChem database, MACCS fingerprints use 166 binary features specified by

MACCS keys, which are smaller but no less relevant [153-154]. Each type of fingerprint assists the predictive models in virtual screening and is an essential tool for accurately representing molecular structures.

3.2.6. Construction of training and independent test datasets

The final dataset was derived from the integration of curated HER2 and EGFR bioactivity data to support dual-target modelling and comprised 6 325 active and 2 179 inactive compounds. To evaluate the model's generalization performance, 20% of the compounds were randomly sampled and reserved as an independent test set, designated H2EGFR-IND, which included 1 265 active and 436 inactive compounds. The remaining 80% of the data, consisting of 5 060 active and 1 743 inactive compounds, was used to construct the training set, referred to as H2EGFR-TRN. In order to address class imbalance, sampling strategies (Random Undersampling, SMOTE, and a combined SMOTE + Undersampling) were integrated into the model training pipeline. The independent test set (H2EGFR-IND) was strictly isolated after dataset splitting and was not subjected to any resampling or preprocessing steps related to class imbalance. All sampling strategies, including random undersampling, SMOTE, and combined approaches, were applied exclusively to the training set (H2EGFR-TRN) during cross-validation to avoid data leakage and ensure unbiased model evaluation.

3.2.7. Machine learning (ML) model development

The trained dataset was utilised to build machine learning models after molecular fingerprint descriptors had been calculated with the RDKit AllChem software. The baseline model development and stacking ensemble phases were the two primary stages of this study. In the baseline model development stage, 11 ML algorithms and 10 molecular fingerprint descriptors were employed to develop these baseline models. The following popular ML algorithms, including Random Forest (RF), AdaBoost (ADA), *k*-Nearest Neighbour (*k*NN), Light Gradient Boosting Machine (LGBM), Decision Tree (DT), Multilayer Perceptron (MLP), Extremely Randomised Trees (ET), logistic regression (LR), Extreme Gradient Boosting (XGB) and Support Vector Machine (SVM) with linear kernel (SVM LN) and with radial basis function (RBF) kernel (SVM RBF) were considered for the construction of base models. Among all the models evaluated, those with the best cross-validation Matthews Correlation Coefficient (MCC) scores were selected to build the baseline models. Specifically, four (MLP, RF, ET, LR) out of the 11 ML algorithms were selected based on their performance. Each of these four algorithms was trained on each fingerprint type, resulting in a total of 40 baseline models.

3.2.8. Stacking ensemble method

A stacking learning technique was employed to create a stacking ensemble at this stage. The method for the selection of the ML models required to construct the 40 baselines was based on the top-performing single-feature-based models from the independent test data (H2EGFR-IND) and the training dataset H2EGFR-TRN. After obtaining 40 baselines, the 40 baseline models were then used to calculate a feature vector for each compound, which served as an input to construct the

meta-model in the stacking ensemble. For any given compound *Q*, each of the 40 baseline models outputs a predicted probability (*PF*), which ranges from 0 to 1. These individual probabilities are combined to form a 40-dimensional feature vector, denoted as *FV(Q)*.

This feature vector can be represented as: $FV(Q) = \{PF_{BM1,1}, PF_{BM1,1}, \ldots, PF_{BMi,\,j,} \ldots PF_{BM4,10}\}$

where: $PF_{BMi,\,j}$ is the predicted probability (*PF*) from the baseline model (BM) that was trained using the i^{th} machine learning algorithm and the j^{th} fingerprint descriptor, $i \in \{1,2,3,4\}$ represents the four selected ML models (RF, MLP, ET, LR), and $j \in \{1, 2...,10\}$ corresponds to the 10 molecular fingerprinting methods.

The output from our baseline model development was utilised as the new training data for the ensemble to construct a stacking method. The stacking ensemble was created by combining the predictions of these baseline models using the logistic regression model as a final estimate. The stacking process used 5-fold cross-validation to prevent overfitting and offer dependable training. The model was then trained on both H2EGFR-TRN and H2EGFR-IND datasets using parallel processing to speed it up. The model generated predictions for both the training and test datasets, including its class labels and probabilities. Its performance was then evaluated using key metrics such as accuracy (ACC), F1 score, sensitivity (Sn), specificity (Sp), Matthews Correlation Coefficient (MCC), and area under the curve (AUC). The stacking ensemble was constructed using predictions from single feature-based models trained on the balanced subsets obtained by applying three resampling techniques, random undersampling, synthetic minority over-sampling technique (SMOTE), and combined sampling, thus enabling improved detection of active compounds without introducing excessive variance.

3.2.9. Dual-target dataset integration and model evaluation

The two different datasets of bioactive compounds, one targeting EGFR and the other targeting HER2, were collected and curated separately. After standardising and preprocessing each dataset independently, they were merged to create a combined dataset for model training. This unified dataset was used to train the stacking ensemble model that learned patterns associated with both EGFR and HER2 inhibition. By integrating both targets into a unified stacking framework, the capability of the model to identify compounds with potential dual-inhibitory activity was enhanced. The stacking ensemble model was validated on the H2EGFR-IND dataset. In order to prevent overfitting and data leakage, the test set (H2EGFR-IND) was strictly isolated during model training and feature selection. The generation of fingerprint and data scaling were first carried out using only the H2EGFR-TRN dataset, and the identical preprocessing parameters were subsequently applied to the H2EGFR-IND dataset to maintain uniformity. This demonstrated superior predictive performance across all evaluation metrics on the test dataset.

3.2.10. Virtual screening of compounds identified by LC–MS/MS using the developed stacking ensemble model for dual targeting of EGFR and HER2 in colorectal cancer

The compounds identified in the *Ceratonia siliqua* L. pod extract by LC–MS/MS were subjected to virtual screening using the developed stacking ensemble model. The SMILES notations for all the identified compounds were obtained from the PubChem database (https://pubchem.ncbi.nlm.nih.gov) and paired with their corresponding PubChem IDs to generate the dataset used for virtual screening. The stacking ensemble model was then employed to predict and estimate the activity probabilities of these compounds against both HER2 and EGFR targets

associated with colorectal cancer. Additionally, the model's performance was evaluated using four FDA-approved reference drugs (Doxorubicin, Abemaciclib, Gemcitabine, and Tucatinib).

3.2.11. Molecular docking studies of ligands predicted by the developed stacking ensemble model and reference FDA-approved drugs

The identification of potent ligands against both HER2 and EGFR is critical for advancing targeted cancer therapies. This study employed the developed stacking ensemble model to predict promising ligands. The 3D structures of the ligands predicted by the developed stacking ensemble model were generated using Chem3D, and then the 3D structures of these ligands and of the four FDA-approved drugs underwent energy minimisation by applying the MM2 force field in Chem3D (PerkinElmer Informatics, 2020). These energy-minimised 3D ligand structures were then converted to and saved in Protein Data Bank (PDB) format for molecular docking simulations. The molecular docking simulations assessed the binding affinity and molecular interactions of HER2 (7MN5 (resolution 2.93 Å)) [155] and EGFR (7ZYM (resolution 2.50 Å)) [156], with their crystal structures retrieved from the Research Collaboratory for Structural Bioinformatics Protein Data Bank (RCSB PDB [157]. The downloaded protein crystal structures were prepared using Discovery Studio Visualizer (v19.1.0.18287) by removing heteroatoms and water molecules, followed by energy minimisation before docking. Autodock Vina, incorporated within PyRx, was utilised for docking simulation. Grid parameters were centred within the reported receptors' binding pocket in the literature. Meanwhile, the docking outputs were examined through the root-mean-square deviation (RMSD) analysis calculation and were visually analysed to obtain the atomic interactions using Discovery Studio for 2D interaction maps, and ChimeraX 1.9 for the 3D representations.

3.2.12. In silico ADME analysis

The pharmacokinetic characteristics (absorption, distribution, metabolism, and excretion (ADME)) of the top hit compounds from virtual screening and selected FDA-approved reference drugs were evaluated using Swiss ADME (http://www.swissadme.ch) and absorption, distribution, metabolism, excretion, and toxicity (ADMET) using ADMETlab 2.0 (https://admetmesh.scbdd.com) web platforms. The physicochemical profile was assessed based on parameters such as molecular weight (MW), number of hydrogen bond acceptors (nHBA), number of hydrogen bond donors (nHBD), number of rotatable bonds (RBN), lipophilicity (logP), solubility, and topological polar surface area (TPSA).

Lipinski's Rule of Five was applied to assess the drug-likeness of the selected compounds. Important absorption indicators include Caco-2 cell permeability and human intestinal absorption (HIA). Excretion characteristics were estimated using predicted clearance rates and biological half-lives. The SMILES representations of the reference compounds and active compounds from the LC–MS/MS analysis were obtained from the PubChem (https://pubchem.ncbi.nlm.nih.gov) database, and the pharmacokinetic characteristics were determined by submitting the SMILES strings to Swiss ADME and ADMETlab 2.0 web servers.

3.3. Experimental validation

3.3.1. Antioxidant activity of *Ceratonia siliqua* L. pod extract

The antioxidant activity of the *Ceratonia siliqua* L. pod extract was determined using the 2,2-diphenyl-1-picrylhydrazyl (DPPH) assay to measure the radical scavenging activity (RSA), following the method established by More and Makola in 2020 [158]. In summary, under low light conditions, 100 µL samples were combined with 100 µL of methanol and 100 µL of DPPH at a concentration of 0.1 mM. After incubating for 30 min at 25 °C in the dark, the absorbance of the resulting mixture was recorded at 517 nm using a microplate reader (Varioskan Flash, Thermo Fisher Scientific, Finland). Ascorbic acid and methanol served as respective positive and negative controls. Three duplicates of the experiment were carried out. Each extract's %RSA or positive control's percentage RSA was calculated as follows:

$$RSA\ (\%) = 100\ (1 - AE/AD) \qquad \text{Equation (3.1)}$$

where *AE* denotes the absorbance of the mixture of the plant extract or ascorbic acid, while *AD* represents the absorbance of DPPH without the extract or ascorbic acid.

3.3.2. Anti-cancer activity of *Ceratonia siliqua* L. pod extract

The anti-cancer property of *Ceratonia siliqua* L. pod extract against cancer cells was evaluated using the MTT cytotoxicity assay. This colorimetric method, based on 3-(4,5-dimethylthiazol-2-yl)-2,5-diphenyl tetrazolium bromide (MTT), measures cell metabolic activity by the enzymatic conversion of MTT into insoluble formazan crystals in viable cells. The process depends on the activity of mitochondrial succinate dehydrogenase, which serves as an indicator of cell viability

and proliferation. The assay was conducted following a slightly modified version of the method described by Johan van Meerloo et al (2011) [159]. The macrophage cells (Raw 264.7), HCT116 colorectal cancer cells and Vero normal kidney cells were cultured in 96-well plates at a density of 1×10^4 cells per well and incubated for 24 h at 37 °C under 5% CO_2 to allow proper cell attachment. The cells were then exposed to varying concentrations of the extract (3.125–100 µg/mL), while doxorubicin served as the positive control at concentrations ranging from 0.003125 to 0.2 mM. Following 48 h of incubation, MTT solution (5 mg/mL) was added, and the plates were further incubated for one hour. Dimethyl sulfoxide (DMSO) was subsequently used to dissolve the formed formazan crystals. The absorbance of each well was read at 570 nm using a Varioskan Flash microplate reader (Thermo Fisher Scientific). The percentage of viable cells was calculated using the formula given in Equation (3.2) below:

$$\text{Cell viability (\%)} = \left(\frac{A_T}{A_C}\right) \times 100 \qquad \text{Equation (3.2)}$$

where A_T and A_C denote the absorbance of treated and control cells, respectively.

3.4. Statistical Analysis

All experimental and computational data were analysed using appropriate statistical methods to determine significance and ensure reliability of the results. Quantitative data were expressed as mean ± standard deviation (SD) from at least three independent replicates. Statistical comparisons between groups were performed using the Mann–Whitney U test for non-parametric data, with a significance threshold set at $p < 0.05$. Where applicable, highly significant differences were considered at $p < 0.001$.

For chemical space analysis, distributional differences in physicochemical properties between active and inactive compounds were also assessed using the Mann–Whitney U test. In the machine learning framework, model performance was evaluated using cross-validation metrics, and results were reported using standard performance measures including accuracy, F1-score, sensitivity, specificity, Matthews correlation coefficient (MCC), and area under the curve (AUC). All analyses and visualisations were performed using Python-based scientific libraries.

CHAPTER 4: RESULTS AND DISCUSSION

PREAMBLE

In this chapter, the results of the experiments outlined in the previous chapter are analysed, evaluated, and discussed. The subsequent subsections cover: HPLC-DAD; liquid chromatography–tandem mass spectrometry (LC–MS/MS) analysis and virtual screening; data distribution and exploratory data analysis; prediction outcomes across various machine learning algorithms; performance evaluation of the stacking ensemble model; cytotoxic activity and antioxidant properties of *Ceratonia siliqua* L. pod extract; as well as molecular docking, and absorption, distribution, metabolism, excretion, and toxicity (ADMET) properties of the predicted active compounds and reference Food and Drug Administration (FDA)-approved drugs.

4.1. HPLC-DAD chromatographic separation of compounds in *Ceratonia siliqua* L. pod extract

A reversed-phase column was employed for the separation of compounds in the filtered crude extract of *Ceratonia siliqua* L. pods due to its high efficiency and suitability for separating moderately polar to nonpolar secondary metabolites such as phenolic acids and flavonoids [160]. Development of the method commenced with an isocratic elution mode to assess the polarity of the constituents in the filtered crude pod extract. At 50:50 A (0.1% v/v formic acid in water): B (0.1% v/v formic acid in acetonitrile), it was observed that most compounds eluted within a narrow retention window, resulting in poor resolution of overlapping peaks **(Figure 4.1A).** Thereafter, the polarity of the mobile phase was decreased by increasing the amount of acetonitrile to 80%, and earlier elution of compounds was observed, but peak broadening and co-elution still occurred,

indicating inadequate separation under isocratic conditions **(Figure 4.1B).** It was not possible to achieve a satisfactory separation under isocratic conditions and hence a gradient elution mode was explored. The final optimised separation was achieved using the following gradient programme: 0–3 min, 2–5% B; 3–10 min, 5–15% B; 10–16 min, 15–40% B; 16–20 min, 40–80% B; and 20–25 min, 80–90% B, at a flow rate of 1.0 mL/min with UV detection at 254 nm **(Figure 4.1C).** The representative chromatogram recorded at 254 nm is presented in **Figure 4.1**. Several peaks were observed, indicating the presence of multiple phytochemical constituents eluting at distinct retention times **(Figure 4.1C).**

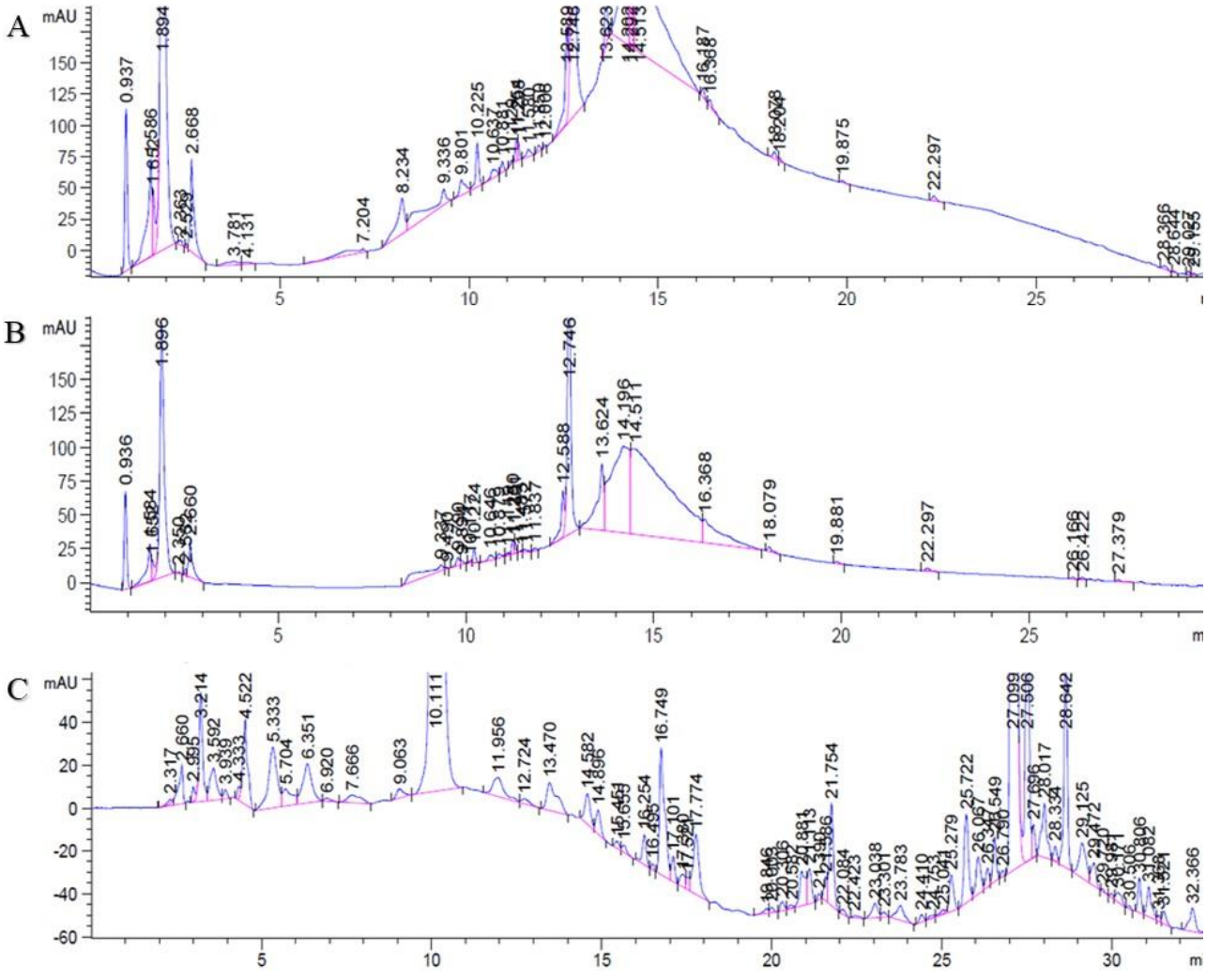


Figure 4.1: Representative HPLC-DAD chromatograms of *Ceratonia siliqua L.* pod extract

The chromatogram revealed early eluting peaks between 0.9 and 4.0 min, corresponding to highly polar constituents such as phenolic acids, which exhibit strong affinity for the aqueous phase and elute quickly under reversed-phase conditions. A prominent sharp peak was observed at ~1.8 min, suggesting the presence of an abundant polar metabolite.

Between 9.0 and 14.5 min, several medium-intensity peaks were detected (retention times: 9.06, 9.86, 10.64, 11.87, 12.74, 13.62, 14.28 min). These peaks are characteristic of moderately polar flavonoid glycosides and related phenolics, whose retention is influenced by hydrogen bonding with the mobile phase [161], [162]. Later peaks, eluting between 16.0 and 22.2 min, correspond to less polar constituents such as flavonoids and other hydrophobic secondary metabolites. Their delayed elution is attributed to their stronger interactions with the C18 column, requiring higher acetonitrile concentration for desorption.

The use of 0.1% formic acid in the mobile phase improved peak symmetry by suppressing ionisation of phenolic hydroxyl groups, reducing peak tailing, and enhancing reproducibility. The baseline separation achieved for most peaks demonstrates that the applied gradient programme successfully balanced polarity differences across the constituents of the extracts.

Importantly, the chromatographic profile demonstrates the chemical diversity of the *C. siliqua* pod extract, ranging from highly polar to nonpolar compounds. These findings justify the subsequent transfer of the HPLC-DAD method to an LC–MS/MS system, where the resolved peaks could be matched with accurate mass spectra and fragmentation patterns for structural elucidation.

4.2. LC-MS/MS analysis and virtual screening of identified compounds in *C. siliqua*

The developed HPLC-DAD method was then transferred to an LC–MS/MS system for further analysis and identification of the separated compounds. **Table 4.1** provides a summary of the compounds identified in the *Ceratonia siliqua* L. pod extract through LC–MS/MS analysis. Identification was achieved using Global Natural Products Social Molecular Networking (GNPS 2.0) by matching molecular (precursor) ions and MS/MS fragmentation patterns obtained from our LC–MS/MS data with reference spectra from publicly available mass spectral libraries. For each compound, the retention time (RT), mass-to-charge ratio (*m*/*z*), molecular formula, and mass errors are reported.

A total of 21 compounds were characterised in the filtered crude extract of *Ceratonia siliqua* L. pods as listed in **Table 4.1**, comprising 11 detected in positive ion mode and 10 in negative ion mode **(Figure 4.2 & Figure 4.3).** The pod extract contained diverse classes of bioactive molecules, including flavonoids, phenolics, terpenoids, alkaloids, and cinnamic acid derivatives. Among these, two flavonoids, one phenolic compound, and five terpenoids have previously been highlighted as major contributors to the antioxidant potential and pharmacological effects of *Ceratonia siliqua* L. [163]. Additional groups such as fatty acids, alkaloids, amino acids, and peptides may also contribute to its medicinal properties, as supported by earlier reports [164], [165].

Table 4.1: LC–MS/MS data (positive and negative ionisation modes) for compounds identified in *Ceratonia siliqua* L. pod extract

Class	**Compound name**	**Formula**	**R_t (min)**	**Precursor *m*/*z***	**Mass error (ppm)**
Fatty acyl	2-methylidene-4-[(2R,3R,4S,5S,6R)-3,4,5-trihydroxy-6-(hydroxymethyl) oxan-2-yl] oxybutanoic acid	$C_{11}H_{18}O_8$	1.92	277.093	1.833
Carbohydrate	Melezitose	$C_{18}H_{32}O_{16}$	17.68	522.203	0.002
Terpenoids	**NCGC00385704-01**	**$C_{35}H_{54}O_{13}$**	**8.97**	**415.148**	**2.286**
Carbohydrate	Sucrose	$C_{12}H_{22}O_{11}$	17.72	387.114	0.002
Alkaloids	Phenazine-1-carboxylic acid	$C_{13}H_8N_2O_2$	14.48	224.06	0.711
Amino acids and peptides	(2S)-2-[[(2S,3R)-2-amino-3-hydroxybutanoyl]amino]-3-phenylpropanoic acid	$C_{13}H_{18}N_2O_4$	0.87	267.134	1.045

Terpenoids	6-O-Isobutyryl-α-D-glucopyranosyl alpha-D-glucopyranoside	$C_{16}H_{28}O_{12}$	0.45	430.192	0.001
Fatty acyl	[6-O-(beta-D-Glucopyranosyl)-β-D-glucopyranosyl]oxy}-2 phenylacetamide	$C_{20}H_{29}NO_{12}$	1.50	520.167	1.008
Cinnamic acids (phenylpropanoids)	2R,3S,4S,5R,6R)-2-(hydroxymethyl)-6-[[(2R,3S,4S,5R,6S)-3,4,5-trihydroxy-6-(2-hydroxy-4-prop-2-enylphenoxy) oxan-2-yl]methoxy] oxane-3,4,5-triol	$C_{21}H_{30}O_{12}$	28.94	492.208	0.014
Alkaloid	4-N-(2-morpholin-4-ylethyl)-6-thiophen-2-ylpyrimidine-2,4-diamine	$C_{14}H_{19}N_5OS$	0.78	299.062	1.719
Flavonol	Gossypetin	$C_{15}H_{10}O_8$	1.85	357.105	1.864
Phenolic	1-Caffeoyl-β-D-glucose	$C_{15}H_{18}O_9$	17.53	341.081	1.022

Carbohydrate	(2R,3S,4S,5R,6S)-2-[[(2R,3R,4R)-3,4-dihydroxy-4-(hydroxymethyl)oxolan-2-yl]oxymethyl]-6-[4-hydroxy-3-(3-methylbut-2-enyl)phenoxy]oxane-3,4,5-triol	$C_{22}H_{32}O_{11}$	29.15	471.187	1.965
Amino acids and peptides	(-)-jasmonoyl-L-isoleucine	$C_{18}H_{29}NO_4$	0.60	322.202	1.128
Flavonol	Chrysin	$C_{15}H_{10}O_4$	0.68	254.29	1.971
Alkaloid	11-oxo-N-(3,3,5-trimethylcyclohexyl)-1-azatricyclo[6.3.1.04,12]dodeca-4,6,8(12)-triene-6-sulfonamide	$C_{20}H_{28}N_2O_3S$	1.69	375.175	1.600
Nucleoside	Guanosine	$C_{10}H_{13}N_5O_5$	6.58	283.05	1.221
Terpenoid	(5E)-4,9-dihydroxy-6-methyl-3,10-dimethylidene-4,7,8,9,11,11a-hexahydro-3aH-cyclodeca[b]furan-2-one	$C_{15}H_{20}O_4$	17.83	293.123	1.606

Fatty acyl	(3R)-4,4-dimethyl-3-[(2S,3R,4S,5S,6R)-3,4,5-trihydroxy-6-(hydroxymethyl)oxan-2-yl]oxyoxolan-2-one	$C_{12}H_{20}O_8$	0.34	297.123	1.982
Terpenoid	Gentiopicroside	$C_{16}H_{20}O_9$	17.50	401.109	0.647
Terpenoids	Mussaendoside S	$C_{42}H_{66}O_{15}$	11.81	275.259	0.001

NCGC00385704-01: *3-[(3S,8R,10S,13R,14S,17R)-3-[(2R,3R,4R,5R,6S)-4,5-dihydroxy-6-methyl-3-[(2S,3R,4S,5S,6R)-3,4,5-trihydroxy-6-(hydroxymethyl)oxan-2-yl]oxyoxan-2-yl]oxy-14-hydroxy-10,13-dimethyl-1,2,3,4,5,6,7,8,9,11,12,15,16,17-tetradecahydrocyclopenta[a]phenanthren-17-yl]-2H-furan-5-one*

The relatively low number of annotated compounds (21 features) is attributed to the strict GNPS filtering and annotation criteria applied in this study, where only high-confidence spectral matches were retained for downstream analysis. In addition, the identified compounds represent features that were consistently detected across both positive and negative ionisation modes under the applied LC–MS/MS conditions.

The predominance of primary metabolites such as sugars and amino acids is consistent with the ethanol–water (1:1) extraction system used, which preferentially extracts polar, low-molecular-weight constituents. The limited representation of certain secondary metabolite classes, including some polyphenols and terpenoids, may be associated with extraction selectivity, ionisation efficiency, and limitations in GNPS spectral library coverage.

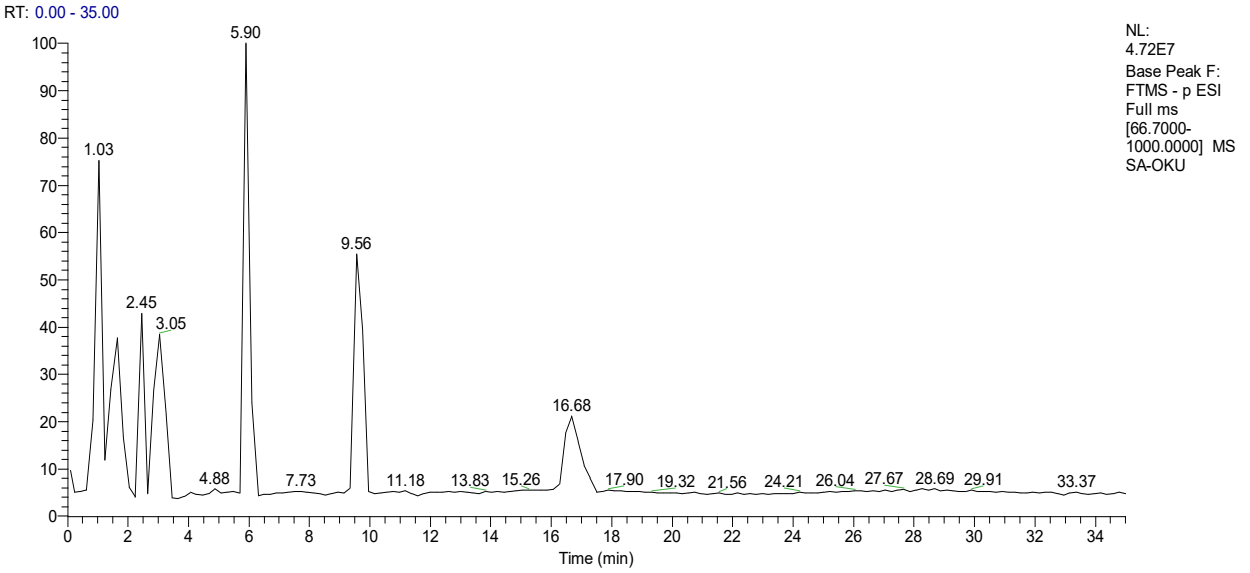


Figure 4.2: LC–MS/MS chromatograms of *Ceratonia siliqua* L. pod extract detected in negative ion mode

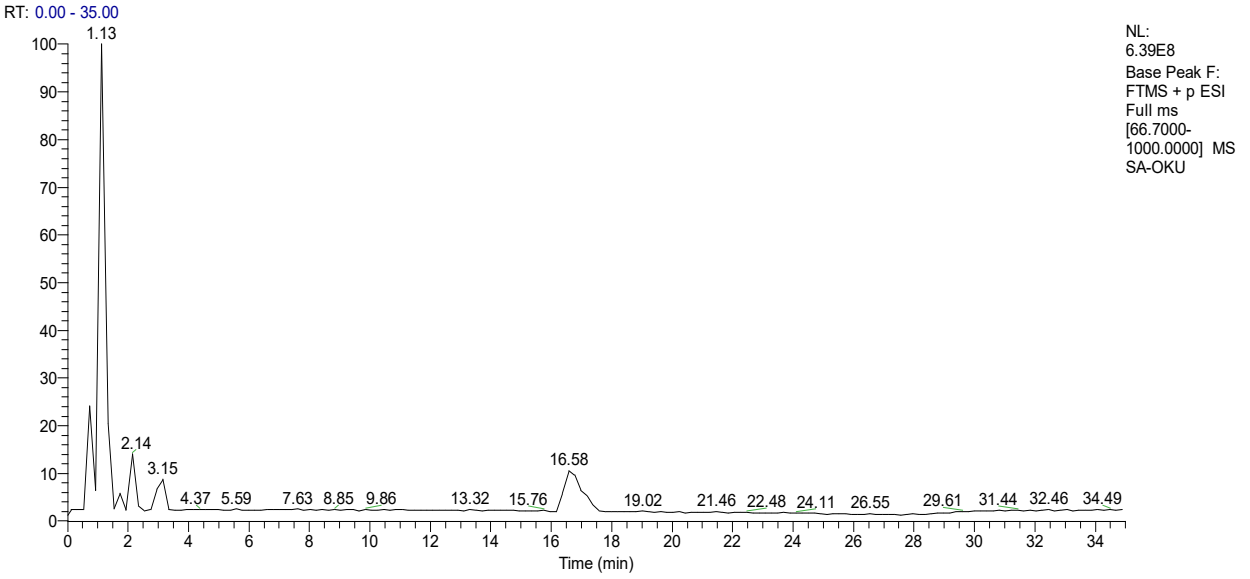


Figure 4.3: LC–MS/MS chromatograms of *Ceratonia siliqua* L. pod extract detected in positive ion mode

4.3. Data distribution analysis

Following the application of the IC_{50} cutoff and the removal of duplicates using a Tanimoto coefficient of 1, a total of 17 287 unique compounds were identified. After averaging IC_{50} values for the remaining duplicates, 6 325 active and 2 179 inactive compounds were retained, resulting in 8 504 confirmed inhibitors targeting both EGFR and HER2 receptors **(Table 4.2).**

Compounds with intermediate potency (IC_{50} between 1 and 10 μM) were excluded from model training to maintain a clear binary classification. However, these 1 626 intermediate compounds were subsequently evaluated to test the generalisability of the stacking model. Predictions showed that 42.8% were classified as active and 57.2% as inactive, with predicted probabilities broadly distributed (mean = 0.462, standard deviation (SD) = 0.123). This distribution indicates that the

model was able to distinguish varying levels of activity even among compounds outside its training set. A negative Pearson correlation ($r = -0.165$) was observed between IC_{50} values and predicted probabilities, suggesting that within the intermediate range, lower IC_{50} values were generally associated with higher predicted activity. This finding highlights the potential of the model to provide biologically relevant insights for borderline compounds, offering a valuable tool for prioritisation in virtual screening pipelines. The distribution patterns and prediction outcomes are illustrated in Appendix A 3. The finalised dataset was divided into 6 803 compounds for training (H2EGFR-TRN) and 1 701 compounds for independent testing (H2EGFR-IND).

Table 4.2: Stepwise dataset construction and curation

Step	Description	Number of Compounds
Initial dataset (ChEMBL)	Total compounds retrieved (EGFR + HER2)	21,991
After preprocessing	Removal of ambiguous and missing values	17,287
After activity filtering	$IC_{50} \leq 1$ μM (active), ≥ 10 μM (inactive)	8,504
Final dataset	Active + inactive compounds	8,504 (6,325 active + 2,179 inactive)
Training set (80%) (H2EGFR-TRN)	Model development	6, 812 (5 060 active and 1 743 inactive)
Test set (20%) (H2EGFR-IND)	Independent validation	1, 701(1 265 active and 436 inactive)

4.4. Exploratory data analysis

An analysis of the chemical space study was to look at patterns and relationships between active and inactive molecules. To achieve a comprehensive view, the distribution of compounds was

examined in relation to their molecular weight (MW) and the logarithm of the octanol–water partition coefficient (ALogP). The evaluation incorporated Lipinski's Rule of Five (Ro5) descriptors to predict whether a compound possesses the physicochemical properties required for potential oral drug-likeness. The Ro5 criteria, widely recognised for assessing drug-likeness, include MW < 500 Da, ALogP less than five, hydrogen bond acceptors (HBAs) fewer than 10, and hydrogen bond donors (HBDs) fewer than five. The visualisation of the chemical space using a scatter plot of MW vs. ALogP **(Figure 4.4)** showed that 48.42% of the compounds were distributed within the MW range of 200–500 Da and ALogP values between 1 and 5, suggesting a favourable profile for biological activity. Conversely, 77.8% of the compounds failed to meet either the MW (< 500 Da) or ALogP (< 5) requirement, while 22.20% failed to meet both, pointing to potential areas for optimisation to enhance physicochemical characteristics and drug potential. These findings emphasise the inherent complexity of the chemical space and the necessity for a nuanced approach in identifying promising drug candidates.

Further analysis of the active and inactive sets against Ro5 parameters indicated that most compounds adhered to the guidelines, typically displaying molecular weights below 500 Da, lipophilicity (ALogP) under five, and fewer than 10 hydrogen bond donors and acceptors.

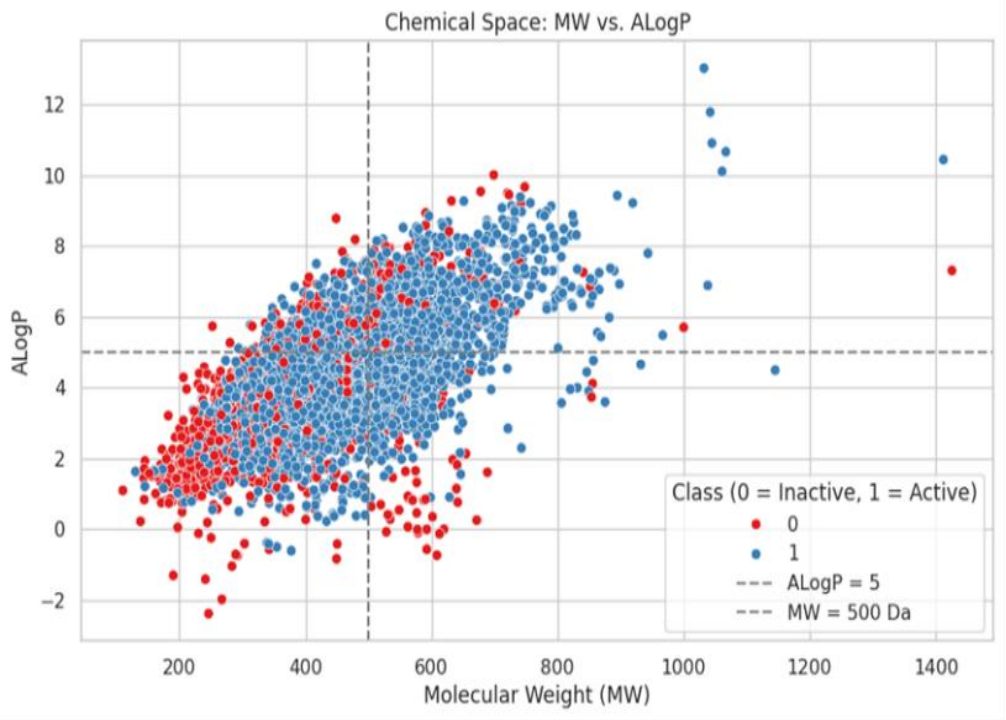


Figure 4.4: A scatter plot of molecular weight (MW) vs. Ghose-Crippen-Viswanadhan octanol–water partition coefficient (ALogP) for compounds in the curated dataset

Significant differences ($p < 0.001$) in molecular weight (MW) between active and inactive substances were found by statistical comparison using the Mann-Whitney U test. On average, inactive compounds exhibited a lower MW (394.27 ± 274.10 Da) than active compounds (490.51 ± 267.12 Da), as reflected in the box plots presented in **Figure 4.5.** When considering the Rule of Five (Ro5) threshold for hydrogen bond acceptors (HBAs) the number of active and inactive compounds appeared to be similar, with statistical significance arising mainly from differences in their mean values, as shown in **Figure 4.5** below.

The ALogP values indicated a modest but notable difference between orally active (4.71 ± 3.41) and inactive (3.92 ± 3.10) compounds. In contrast, both groups displayed comparable mean values

for hydrogen bond donors (HBDs), with no statistically significant difference observed. Molecular complexity plays a critical role in determining a compound's clinical performance, as features such as aromaticity, fused rings, ring count, rotatable bonds (RBN), functional groups and chiral centres directly influence toxicity, oral bioavailability and solubility. These factors ultimately affect the therapeutic potential of a compound [166]

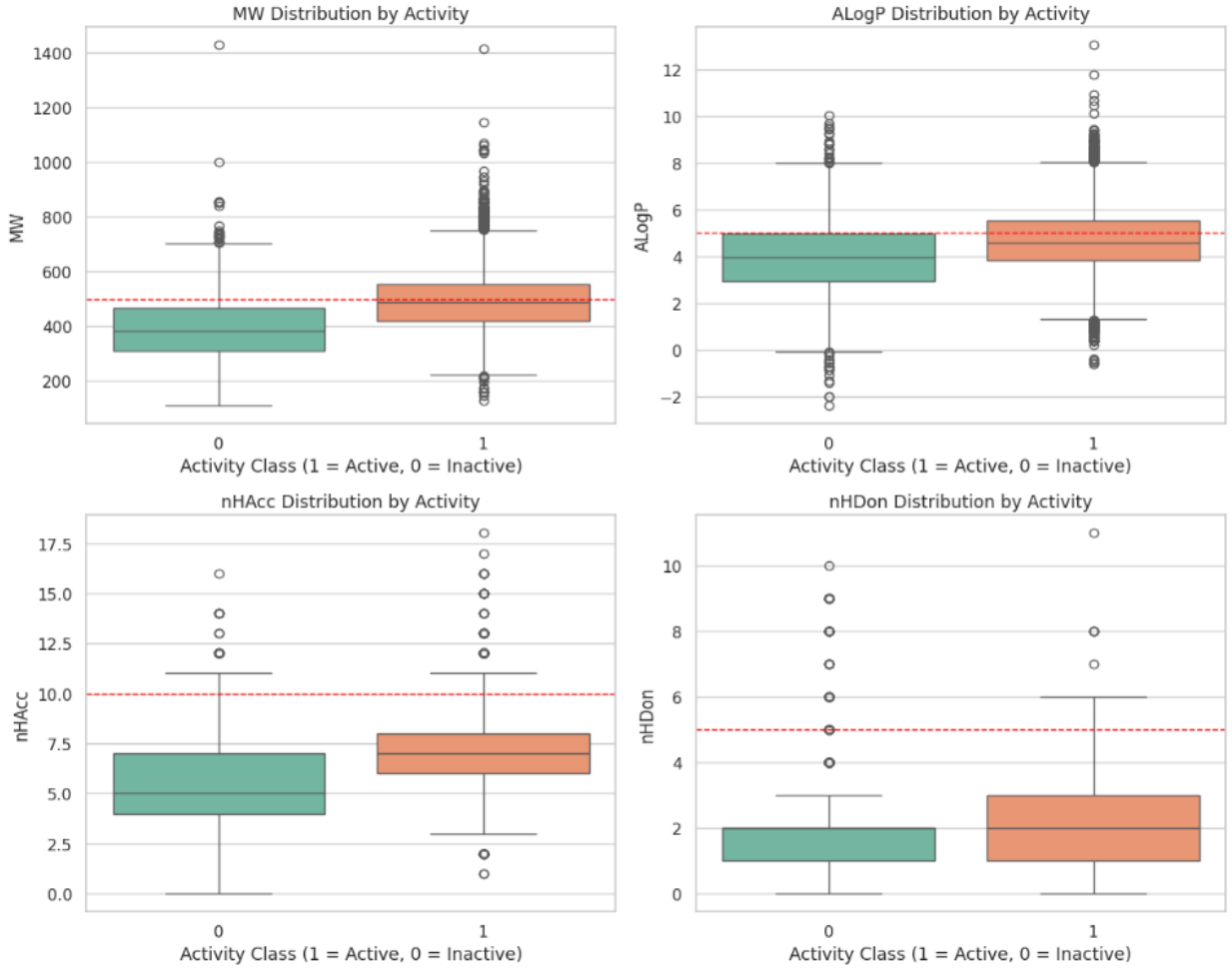


Figure 4.5: Box plots of Lipinski's Ro5 descriptors. The four Ro5 descriptors illustrated are molecular weight (MW), Ghose-Crippen-Viswanadhan octanol–water partition coefficient (ALogP), number of hydrogen bond donors (nHBD) and number of hydrogen bond acceptors (nHBA)

To further evaluate molecular complexity, four structural descriptors were analysed: number of rotatable bonds (RBN), number of rings (nCIC), number of benzene-like rings (nBnz), and aromatic ratio (ARR). The box plots in **Figure 4.6** show that active compounds generally possess a lower ARR, a higher number of rotatable bonds, and more benzene-like rings than inactive compounds. The statistical significance of these differences between the active and inactive molecules was established at $p < 0.001$.

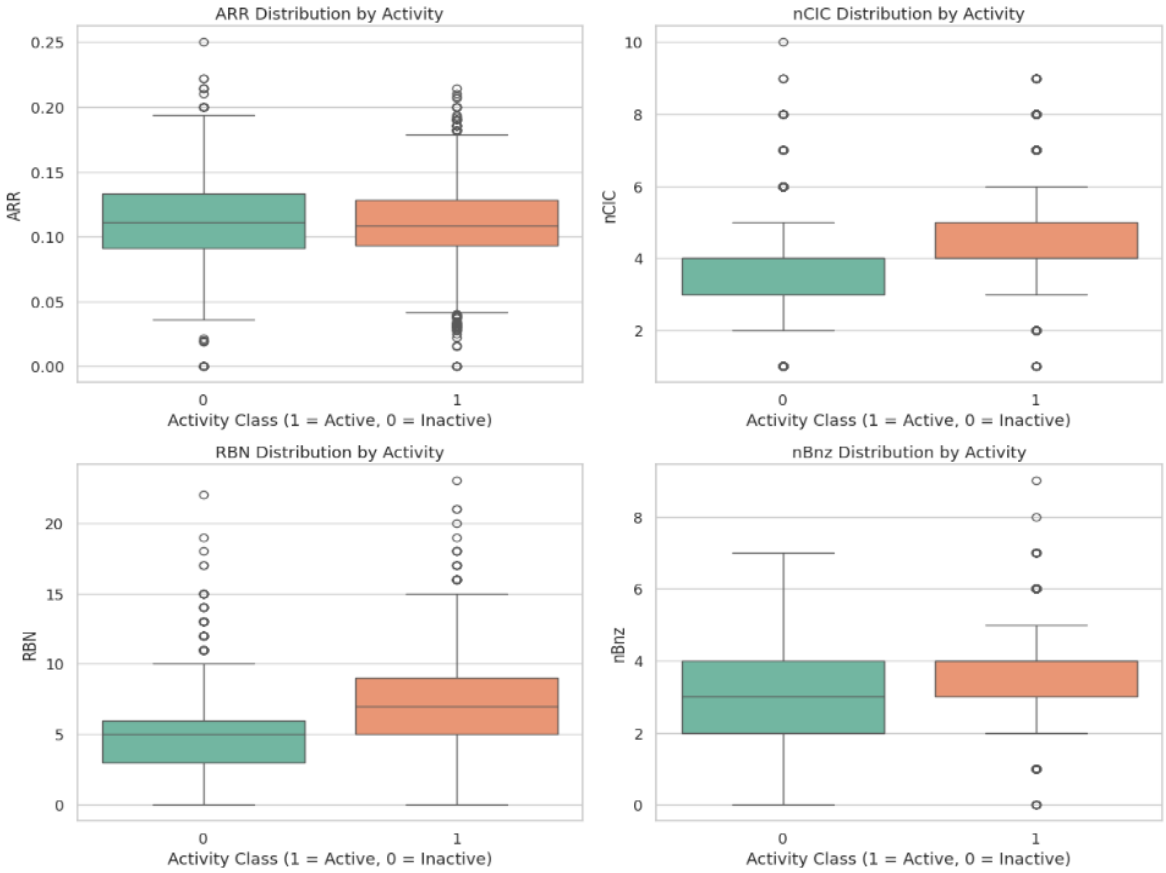


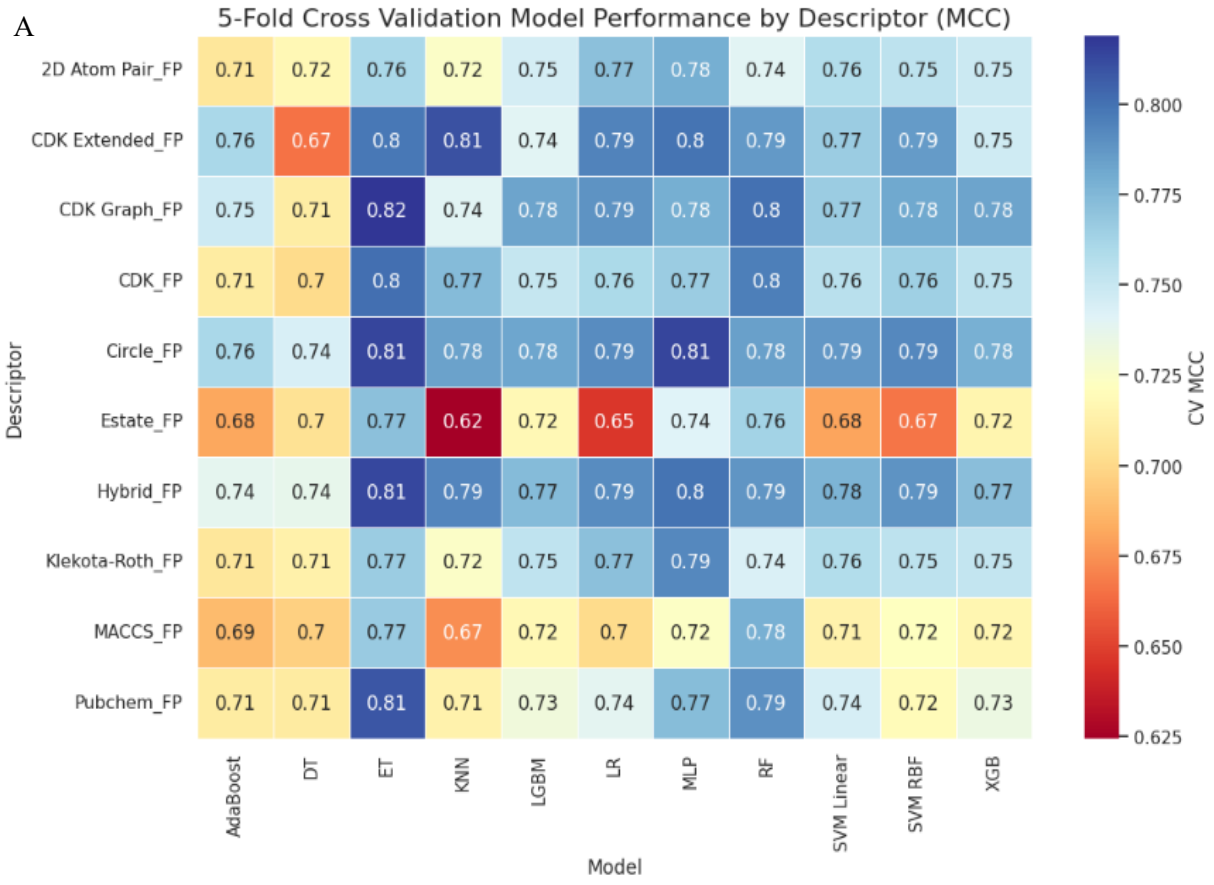
Figure 4.6: Box plots of molecular complexity descriptors. The four descriptors shown in this figure represent aromatic ratio (ARR), number of rings (nCIC), number of rotatable bonds (RBN), and number of benzene-like rings (nBnz)

4.5. Prediction outcomes across various machine learning algorithms and molecular descriptors

A total of 110 machine learning (ML) classifiers were constructed by combining 10 different molecular descriptors with various algorithms. In this section, the performance of 11 ML algorithms is compared. Evaluation was carried out using both an independent test set and 5-fold cross-validation on the training data. The heatmaps in **Figure 4.7** summarise these results, highlighting that certain models particularly Decision Tree (DT), Extra Trees (ET), Random Forest (RF), and Multilayer Perceptron (MLP) consistently achieved Matthews Correlation Coefficient (MCC) values greater than 0.985 **(Figure 4.7A)**, which reflect robust performance across multiple descriptor types.

Among the descriptors tested, CDK, Extended_FP, Circle_FP, and Hybrid_FP demonstrated the strongest predictive capabilities across most algorithms **(Figure 4.7B)**. The model with the highest MCC score in cross-validation was designated as the best performer. The five top-ranking classifiers, along with their MCC values, were: ET-CDK Graph (0.819), ET-Circle (0.814), MLP-Circle (0.814), ET-Hybrid (0.813), and *k*NN-CDK Extended (0.810).

A


5-Fold Cross Validation Model Performance by Descriptor (MCC)
Descriptor
2D Atom Pair_FP 0.71 0.72 0.76 0.72 0.75 0.77 0.78 0.74 0.76 0.75 0.75
CDK Extended_FP 0.76 0.67 0.8 0.81 0.74 0.79 0.8 0.79 0.77 0.79 0.75
CDK Graph_FP 0.75 0.71 0.82 0.74 0.78 0.79 0.78 0.8 0.77 0.78 0.78
CDK_FP 0.71 0.7 0.8 0.77 0.75 0.76 0.77 0.8 0.76 0.76 0.75
Circle_FP 0.76 0.74 0.81 0.78 0.78 0.79 0.81 0.78 0.79 0.79 0.78
Estate_FP 0.68 0.7 0.77 0.62 0.72 0.65 0.74 0.76 0.68 0.67 0.72
Hybrid_FP 0.74 0.74 0.81 0.79 0.77 0.79 0.8 0.79 0.78 0.79 0.77
Klekota-Roth_FP 0.71 0.71 0.77 0.72 0.75 0.77 0.79 0.74 0.76 0.75 0.75
MACCS_FP 0.69 0.7 0.77 0.67 0.72 0.7 0.72 0.78 0.71 0.72 0.72
Pubchem_FP 0.71 0.71 0.81 0.71 0.73 0.74 0.77 0.79 0.74 0.72 0.73
AdaBoost DT ET KNN LGBM LR MLP RF SVM Linear SVM RBF XGB
Model
CV MCC
0.800
0.775
0.750
0.725
0.700
0.675
0.650
0.625

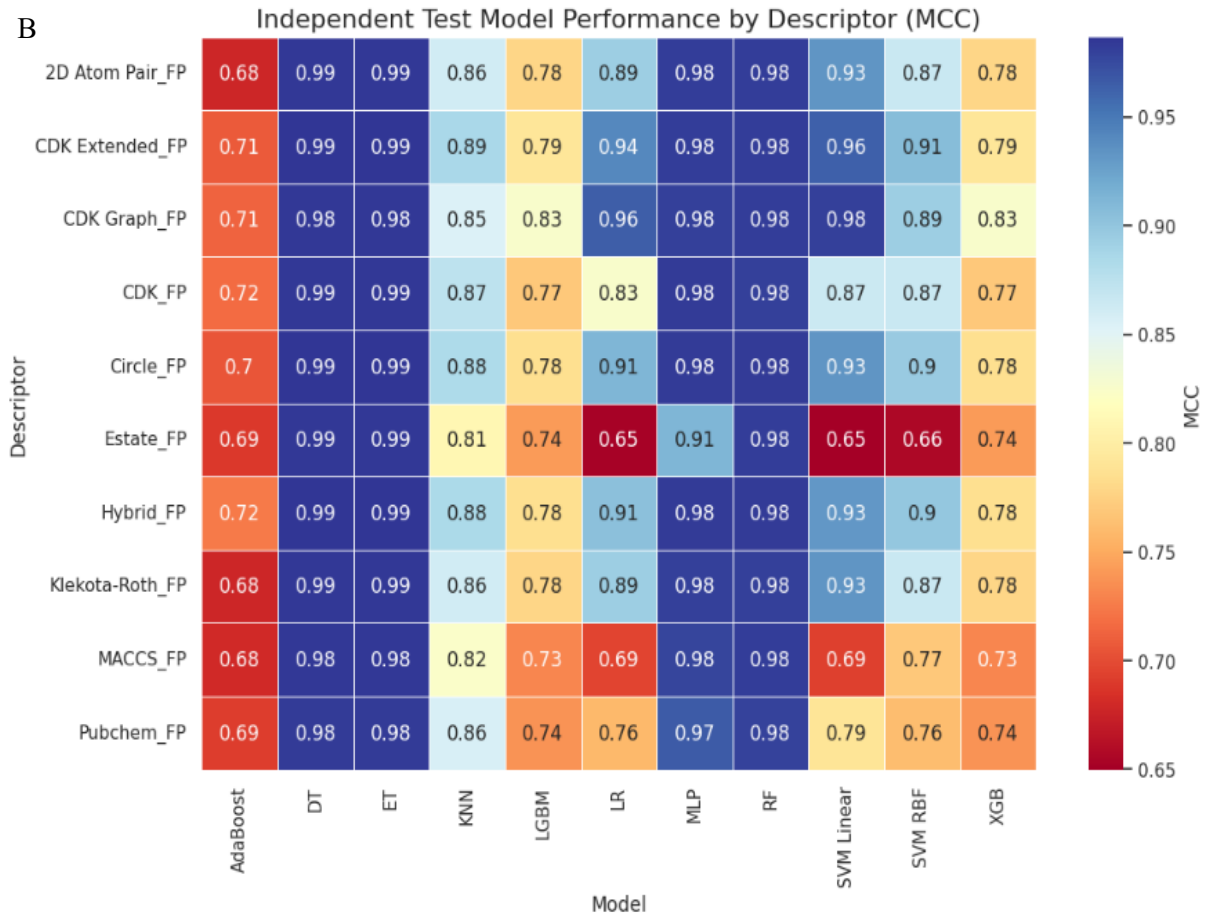


Figure 4.7: (A) MCC values of 110 baseline models in terms of 5-fold cross-validation training; and (B) MCC values of 110 baseline models in terms of independent tests

Based on cross-validation outcomes, the ET-CDK Graph model emerged as the best-performing individual classifier, achieving a Matthews Correlation Coefficient (MCC) of 0.819, an accuracy (ACC) of 0.930, and an area under the curve (AUC) of 0.70 on the independent test set **(Table 4.3)**. In comparison, the DT-Hybrid, ET-Hybrid, and ET-CDK Extended models each attained an MCC of 0.987, with corresponding ACC values of 0.993 and AUC scores of 0.999 on the independent test. Although the Decision Tree (DT) model exhibited exceptional performance on the independent test set, its markedly lower results during training indicated possible overfitting.

To mitigate this, the DT model was replaced with logistic regression (LR) in the stacking ensemble design. Despite not ranking among the top five classifiers in the training phase **(Table 4.3)**, the LR model provided more balanced and consistent performance across both training and independent testing. This stability is evident from the heatmaps presented in **Figure 4.7.** Full numerical results for both individual models and ensemble configurations are provided in **Table 4.3** and **Table 4.4**.

Table 4.3: Top five models based on cross-validation training metrics

Descriptor	Model	ACC	F1 Score	Sn	Sp	MCC	AUC
CDK Graph	ET	0.930	0.953	0.956	0.855	0.819	0.970
Circle	ET	0.930	0.952	0.957	0.847	0.814	0.974
Circle	MLP	0.929	0.953	0.966	0.826	0.814	0.964
Hybrid	ET	0.928	0.952	0.960	0.839	0.813	0.972
CDK Extended	*k*NN	0.927	0.951	0.960	0.834	0.810	0.955

Table 4.4: Top five models based on independent test metrics

Descriptor	Model	ACC	F1 Score	Sn	Sp	MCC	AUC
Hybrid	DT	0.994	0.996	0.991	1.0	0.987	0.999

Hybrid	ET	0.994	0.996	0.991	1.0	0.987	0.999
CDK Extended	DT	0.994	0.995	0.991	1.0	0.986	0.999
CDK Extended	ET	0.994	0.995	0.991	1.0	0.986	0.999
Circle	DT	0.994	0.995	0.991	1.0	0.986	0.999

The ET-CDK Graph model was identified as the top performer in cross-validation and also ranked among the highest-scoring models on the independent test set. Its consistent results across both the H2EGFR-TRN and H2EGFR-IND datasets suggest strong stability, even when relying on a single descriptor. Nevertheless, the focus of this study extended beyond single-feature models to develop a stacking ensemble by integrating the best-performing ML models (Extra Trees (ET), logistic regression (LR), Multilayer Perceptron (MLP), and Random Forest (RF)) across all 10 molecular descriptors. For each algorithm, baseline models were trained using combined descriptors, and their outputs served as inputs for the final ensemble. This multi-descriptor stacking approach outperformed single-feature models, producing a more robust and accurate predictive framework.

4.6. Performance evaluation of stacking-based ensemble model compared to the single-feature-based model

To ensure stable and reliable predictive performance, multiple machine learning classifiers were combined into a meta-model using a stacking strategy. Unlike models trained on single feature sets, this ensemble approach has consistently demonstrated superior outcomes [167]. In many machine learning studies, performance is enhanced by aggregating the predictions from multiple models rather than relying on a single classifier. In this work, the ensemble consisted of several

base learners, with their predictions integrated through logistic regression (LR), which acted as the final estimator. As summarised in **Table 4.5**, the stacking classifier achieved outstanding results, with a Matthews Correlation Coefficient (MCC) of 0.988 on the H2EGFR-TRN and 1.0 on the H2EGFR-IND dataset. On the training data, it recorded the highest performance across all metrics, including an accuracy of 0.995, an F1 score of 0.994, and an area under the curve (AUC) of 1.0. Sensitivity (Sn) was perfect (1.0), confirming the model's capacity to accurately detect every positive instance, while specificity (Sp) remaining high (98.2%), minimising false positives. On the independent test dataset (H2EGFR-IND), the stacking classifier achieved perfect scores across all evaluation metrics: accuracy, F1 score, sensitivity (Sn), specificity (Sp), MCC, and AUC **(Table 4.6).** Overfitting and data leakage were avoided by maintaining strict separation of the test set during training and feature selection. This indicates that the model generalises effectively to unseen data. When compared to single-feature models, RF and ET emerged as the strongest performers. Nonetheless, the stacking strategy, which integrates the strengths of multiple classifiers, proved to be the most robust and reliable model for this dataset, achieving perfect predictive performance on the independent test set.

Table 4.5: Cross-validation performance metrics of models together with the stacking classifier on the training dataset (H2EGFR-TRN)

Descriptor	Model	ACC	F1 Score	Sn	Sp	MCC	AUC
All	**Stacking Ensemble**	**0.995**	**0.994**	**1.0**	**0.982**	**0.988**	**1.0**
All	MLP	0.929	0.953	0.964	0.831	0.815	0.970
	ET	0.926	0.950	0.958	0.834	0.807	0.974

	LR	0.923	0.949	0.958	0.827	0.803	0.969
	SVM Linear	0.921	0.947	0.954	0.826	0.792	0.962
	RF	0.920	0.947	0.965	0.792	0.787	0.973
	SVM RBF	0.917	0.943	0.966	0.776	0.777	0.961
	XGB	0.913	0.943	0.964	0.768	0.768	0.960
	LGBM	0.911	0.941	0.961	0.768	0.763	0.959
	AdaBoost	0.903	0.936	0.954	0.758	0.741	0.947
	*k*NN	0.901	0.935	0.966	0.716	0.734	0.942
	DT	0.882	0.920	0.924	0.760	0.691	0.842

Table 4.6: Classification metrics for performance evaluation of the stacking classifier and its base models on the independent test dataset (H2EGFR-IND)

Descriptor	**Model**	**Test ACC**	**Test F1 Score**	**Test Sn**	**Test Sp**	**Test MCC**	**Test AUC**
All	**Stacking Ensemble**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**
All	RF	0.987	0.992	0.988	0.985	0.966	0.998
All	ET	0.986	0.991	0.987	0.985	0.963	0.998
	MLP	0.981	0.987	0.981	0.980	0.947	0.995

LR	0.975	0.983	0.972	0.983	0.933	0.994
LGBM	0.973	0.982	0.984	0.938	0.925	0.994
XGB	0.973	0.982	0.985	0.935	0.925	0.995
DT	0.970	0.980	0.969	0.973	0.920	0.971
SVM Linear	0.965	0.976	0.959	0.983	0.908	0.994
AdaBoost	0.952	0.968	0.970	0.893	0.866	0.987
*k*NN	0.932	0.956	0.975	0.796	0.807	0.974
SVM RBF	0.931	0.955	0.966	0.816	0.804	0.985

4.7. Virtual screening of compounds identified by LC–MS/MS

Following identification by LC–MS/MS, the compounds were subjected to virtual screening against both HER2 and EGFR targets using the custom-built stacking ensemble model. This model predicted the activity status (active or inactive) of each compound and provided associated probability scores. Additionally, predictions were performed separately against HER2 and EGFR to further evaluate potential selectivity. This screening facilitated the assessment of the therapeutic relevance of the identified compounds in colorectal cancer treatment.

Table 4.7: LC–MS/MS data (positive and negative ionisation modes) for compounds identified in *Ceratonia siliqua* L. pod extract, along with their virtual screening (VS) results and predicted probabilities. Compounds are screened based on their predicted activity status, where **1 = Active** and **0 = Inactive**.

Compound name	**Virtual screening prediction (EGFR&HER2)**	**Probability**	**VS single target (EGFR or HER)**	**Probability**
2-methylidene-4-[(2R,3R,4S,5S,6R)-3,4,5-trihydroxy-6-(hydroxymethyl) oxan-2-yl] oxybutanoic acid	0	0.269	0	0.31
Melezitose	0	0.317	0	0.32
NCGC00385704-01	**1**	**0.513**	**1 (BOTH)**	**0.612 & 0.596**
Sucrose	0	0.249	0	0.223
phenazine-1-carboxylic acid	0	0.214	1 (EGFR)	0.630

(2S)-2-[[(2S,3R)-2-amino-3-hydroxybutanoyl] amino]-3-phenylpropanoic acid	0	0.353	0	0.271
6-O-Isobutyryl-α-D-glucopyranosyl,alpha-D-glucopyranoside	0	0.330	0	0.30
[6-O-(beta-D-Glucopyranosyl)-β-D-glucopyranosyl]oxy}-2 phenylacetamide	0	0.369	0	0.35
2R,3S,4S,5R,6R)-2-(hydroxymethyl)-6-[[(2R,3S,4S,5R,6S)-3,4,5-trihydroxy-6-(2-hydroxy-4-prop-2-enylphenoxy) oxan-2-yl]methoxy] oxane-3,4,5-triol	0	0.267	0	0.482
4-N-(2-morpholin-4-ylethyl)-6-thiophen-2-ylpyrimidine-2,4-diamine	0	0.321	0	0.311
Gossypetin	0	0.280	1 (EGFR)	0.681
1-Caffeoyl-β-D-glucose	0	0.271	1 (HER2)	0.65

(2R,3S,4S,5R,6S)-2-[[(2R,3R,4R)-3,4-dihydroxy-4-(hydroxymethyl)oxolan-2-yl]oxymethyl]-6-[4-hydroxy-3-(3-methylbut-2-enyl)phenoxy]oxane-3,4,5-triol	0	0.452	0	0.450
(-)-jasmonoyl-L-isoleucine	0	0.272	0	0.38
Chrysin	0	0.246	1 (HER2)	0.735
11-oxo-N-(3,3,5-trimethylcyclohexyl)-1-azatricyclo[6.3.1.04,12]dodeca-4,6,8(12)-triene-6-sulfonamide	0	0.346	0	0.412
Guanosine	0	0.334	0	0.361
(5E)-4,9-dihydroxy-6-methyl-3,10-dimethylidene-4,7,8,9,11,11a-hexahydro-3aH-cyclodeca[b]furan-2-one	0	0.249	0	0.330

(3R)-4,4-dimethyl-3-[(2S,3R,4S,5S,6R)-3,4,5-trihydroxy-6-(hydroxymethyl)oxan-2-yl]oxyoxolan-2-one	0	0.383	0	0.315
Gentiopicroside	0	0.424	0	0.264
Mussaendoside S	0	0.490	0	0.214

Table 4.8: Virtual screening (VS) predictions of reference anti-cancer drugs against EGFR, HER2, and dual targets using the stacking ensemble model

Compound Name	**Virtual screening prediction (EGFR & HER2)**	**Probability**	**VS single target (HER2)**	**Probability**	**VS single target (EGFR)**	**Probability**
Abemaciclib	1	0.5109	1	0.712	1	0.746
Gemcitabine	0	0.3623	0	0.395	1	0.663
Doxorubicin	1	0.5510	1	0.621	1	0.694
Tucatinib	1	0.5507	1	0.665	1	0.652

1: Active; and 0: Inactive

To contextualise the predictive performance of our virtual screening, a set of FDA-approved anti-cancer drugs (Doxorubicin, Abemaciclib, Gemcitabine, and Tucatinib) were screened against HER2 and EGFR using the stacking ensemble model **(Table 4.8)**. All reference compounds, including Abemaciclib (dual-inhibition probability = 0.5109; HER2 = 0.712; EGFR = 0.746), Doxorubicin (0.5510; HER2 = 0.621; EGFR = 0.694), and Tucatinib (0.5507; HER2 = 0.665; EGFR = 0.652), demonstrated strong predicted dual-inhibitory activity consistent with their established clinical roles in cancer therapy. Gemcitabine, however, showed no dual-inhibitory activity (0.3623), with weak activity against HER2 (0.395) but moderate inhibition of EGFR (0.663). This validation step confirmed the capacity of the model to identify known dual inhibitors. When compared to these benchmarks, one phytochemical *(3-[(3S,8R,10S,13R,14S,17R)-3-[(2R,3R,4R,5R,6S)-4,5-dihydroxy-6-methyl-3-[(2S,3R,4S,5S,6R)-3,4,5-trihydroxy-6-(hydroxymethyl)oxan-2-yl]oxyoxan-2-yl]oxy-14-hydroxy-10,13-dimethyl-1,2,3,4,5,6,7,8,9,11,12,15,16,17-tetradecahydrocyclopenta[a]phenanthren-17-yl]-2H-furan-5-one)* from *Ceratonia siliqua* L., given the code name NCGC00385704-01, was predicted to possess moderate dual EGFR/HER2 inhibitory activity (probability = 0.513) **(Table 4.7)**. The compound NCGC00385704-01 was predicted to inhibit both proteins individually, with probabilities of 0.612 for EGFR and 0.596 for HER2. Although weaker than FDA-approved drugs, this compound may serve as a potential lead scaffold for dual-target inhibitor design, particularly relevant given the overexpression of both HER2 and EGFR in up to 85% of colorectal cancer cases.

In addition, several compounds that were identified in this study showed selective inhibition of either HER2 or EGFR. For example, Chrysin and 1-caffeoyl-β-D-glucose exhibited HER2 inhibition with probabilities of 73.5% and 65.0%, respectively, while Gossypetin and Phenazine-

1-carboxylic acid demonstrated activity against EGFR with predicted probabilities of 68.1% and 63.0%. **(Table 4.7).** These findings align with prior studies that have reported their inhibitory activity toward individual receptors but not dual-target inhibition [168], [169]. Importantly, many phytochemicals exert multi-targeted anti-cancer effects beyond HER2/EGFR blockade. For instance, Chrysin and Gossypetin are known modulators of oxidative stress and inflammatory pathways such as NF-κB and COX-2 [170], processes that play pivotal roles in colorectal carcinogenesis. Although classified as inactive by the dual-inhibition model, their receptor-selective activity and pleiotropic mechanisms suggest that they may contribute to synergistic or complementary therapeutic effects. The significant cytotoxicity observed in HCT116 cells (IC_{50} = 13.32 ± 1.09 μg/mL) may therefore be attributable to these multifactorial interactions rather than direct HER2/EGFR dual blockade.

Beyond polyphenolic constituents, carbohydrate-based compounds, including melezitose and sucrose were also detected **(Table 4.7)**, reflecting the presence of dietary fibres and oligosaccharides in the pod extract. Previous studies indicate that carob pod fibres are composed primarily of insoluble polysaccharides and can interact with phenolic compounds, altering their bioavailability and enhancing their bioactivity in the colon. These interactions contribute to gut health and offer chemopreventive benefits in colorectal [34], [83-84].

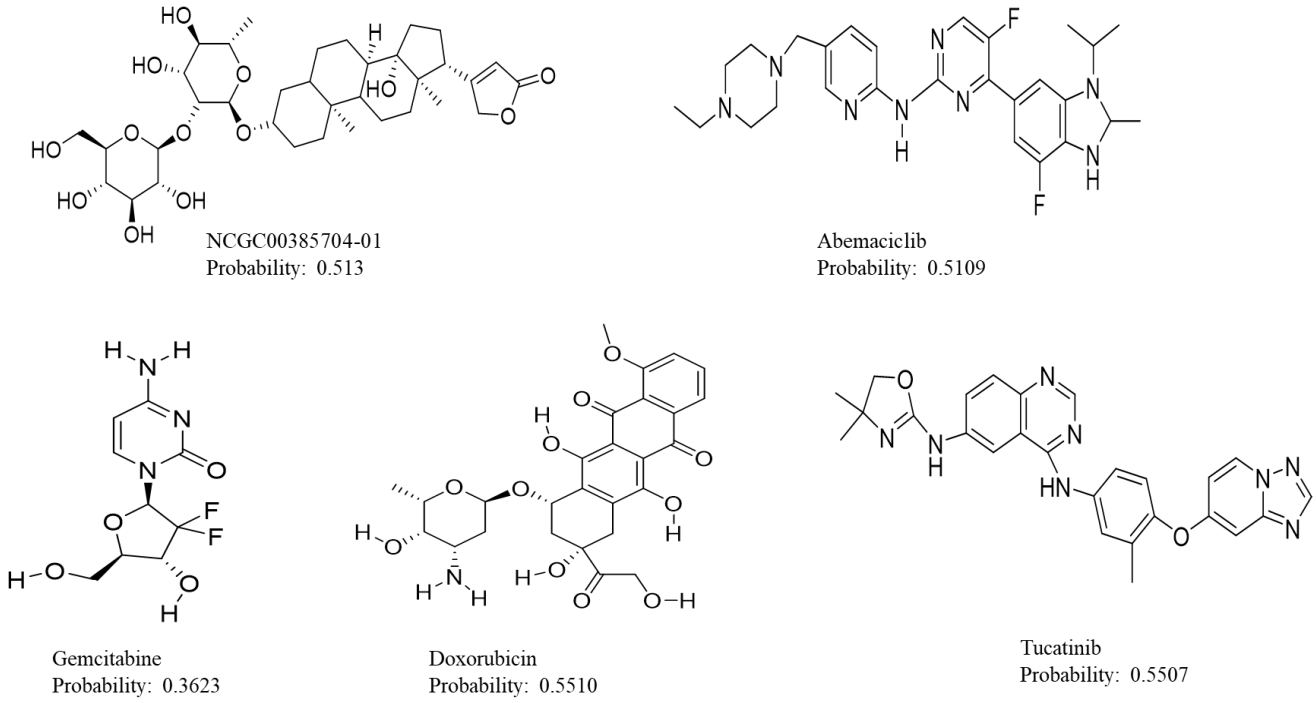


Figure 4.8: Chemical structures of the active compound identified from the LC–MS/MS analysis of *Ceratonia siliqua* L. (CS) and the reference FDA-approved drugs

4.8. Molecular docking studies of NCGC00385704-01 and reference FDA-approved drugs against EGFR and HER2

Molecular docking is a fundamental computational approach for investigating ligand–receptor interactions at the atomic scale, providing valuable insights into underlying biochemical mechanisms [171]. In this study, docking analysis was performed to evaluate the binding activity of the compounds illustrated in **Figure 4.8** against human epidermal growth factor receptor 2 (HER2) and epidermal growth factor receptor (EGFR). It is well-established that HER2, a member of the ErbB receptor tyrosine kinase (RTK) family, is a critical driver of oncogenesis through aberrant activation of PI3K/AKT and MAPK/ERK1/2 signalling pathways, commonly resulting from gene amplification or activating mutations. Such dysregulation, observed in approximately

20% of breast cancers, is also associated with poor clinical outcomes in oesophagogastric, ovarian, bladder, and colorectal cancers [172]. Doxorubicin, abemaciclib, and gemcitabine were included as anticancer reference compounds to provide a comparative baseline for binding affinity within the EGFR and HER2 active sites. Their inclusion was intended to benchmark the predicted ligands against clinically established cytotoxic agents rather than EGFR/HER2-specific inhibitors.

Docking results, summarised in **Table 4.9**, report binding affinities as Gibbs free energy ($\Delta G°$, kcal/mol), where lower values indicate stronger and more stable interactions between ligands and their target proteins. The inhibition constant (K_i) was subsequently calculated from the binding energy using Equation (4.1), with a correction of +0.01 kcal/mol. This parameter reflects the drug-like potential of the compounds, as lower K_i values correspond to stronger binding affinity and higher inhibitory potential.

$$K_i = \exp(-\Delta G°/RT) \qquad \text{Equation (4.1)}$$

where R is the gas constant (0.001987 kcal/K·mol) and T is the temperature (298.15 K). Collectively, Gibbs free energy values provide insight into the stability of protein–ligand complexes, while the calculated K_i values highlight their therapeutic relevance. Thus, molecular docking served as a key predictive tool in this study, guiding the evaluation of predicted compound and reference drugs against HER2 and EGFR [171].

The molecular docking analysis **(Table 4.9)** provides insights into the inhibiting constants and binding affinities of the only performing compound in the *Ceratonia siliqua* L. pod extract predicted by the stacking ensemble model and four FDA-approved drugs against HER2 (PDB ID: 7MN5) and EGFR (PDB ID: 7ZYM). Among the standard drugs, Tucatinib demonstrated the

strongest binding affinity to both HER2 (–10.5 kcal/mol, K_i = 0.0204 μM) and EGFR (–10.1 kcal/mol, K_i = 0.0401 μM), confirming its established role as a highly potent HER2/EGFR inhibitor. Abemaciclib also showed strong dual-target affinity (–9.9 kcal/mol and –9.2 kcal/mol), while Doxorubicin performed moderately well against both receptors, and Gemcitabine exhibited the weakest binding, consistent with its non-specific mechanism of action as a nucleoside analogue.

The machine learning-predicted compound NCGC00385704-01 in the *Ceratonia siliqua* L. pod extract demonstrated binding energies of –7.2 kcal/mol (HER2) and –8.0 kcal/mol (EGFR), corresponding to K_i values of 5.3598 μM and 1.3890 μM, respectively **(Table 4.9)**. Although its affinity was lower than that of Tucatinib and Abemaciclib, NCGC00385704-01 showed more favourable binding than Gemcitabine across both receptors. This indicates potential selective activity with moderate inhibitory strength, which could serve as a scaffold for optimisation in future drug development efforts.

Overall, the docking results reinforce the reliability of the machine learning model in identifying novel compounds with reasonable binding profiles compared to standard FDA-approved drugs. While Tucatinib remains the most potent inhibitor, the predicted compound NCGC00385704-01 shows promise for further refinement, highlighting the synergy of machine learning-guided virtual screening and molecular docking in accelerating drug discovery**.**

Table 4.9: Molecular docking results for the predicted compound and FDA-approved drugs against HER2 and EGFR targets

Ligand	Binding energy (kcal/mol) (7MN5)	Inhibition constant (μM)	Binding energy (kcal/mol) (7ZYM)	Inhibition constant (μM)
Abemaciclib	-9.9	0.0562	-9.2	0.1832
Doxorubicin	-8.4	0.7071	-9.5	0.1104
Gemcitabine	-6.5	17.4706	-6.1	34.3194
NCGC00385704-01	-7.2	5.3598	-8.0	1.3890
Tucatinib	-10.5	0.0204	-10.1	0.0401

Figure 4.9 illustrates the molecular interactions of NCGC00385704-01 and the four FDA-approved standard drugs within the HER2 binding pocket. The predicted compound NCGC00385704-01 **(Figure 4.9A)** formed multiple conventional hydrogen bonds with residues CYS311, LEU256, and ARG103, along with a π-alkyl interaction with PHE258. These interactions, while stabilising, were fewer in number compared to the standard drugs, aligning with its moderate binding energy and higher inhibition constant observed in molecular docking results **(Table 4.9)**.

Among the reference drugs, Abemaciclib **(Figure 4.9B)** established an extensive hydrogen-bonding network with GLN283, SER300, and THR273, supplemented by hydrophobic and π-alkyl contacts, supporting its strong dual-target affinity. Doxorubicin **(Figure 4.9C)** also formed multiple hydrogen bonds and π–π stacking interactions, contributing to its stable binding profile despite lower potency than Abemaciclib and Tucatinib. Gemcitabine **(Figure 4.9D)**, consistent with its weak docking score, exhibited fewer stabilising contacts, dominated by hydrogen bonds with ASN259 and CYS255. Tucatinib **(Figure 4.9E)** demonstrated the strongest binding among the standards, engaging residues such as VAL272, TYR274, and GLU265 through hydrogen bonding and π interactions, reflecting its high docking affinity and low K_i values. Although the ligand (NCGC00385704-01) did not come into interaction with the 95 residues in the active site of the receptor, their binding patterns suggest potential as partial agonists or allosteric modulators, warranting further pharmacological evaluation.

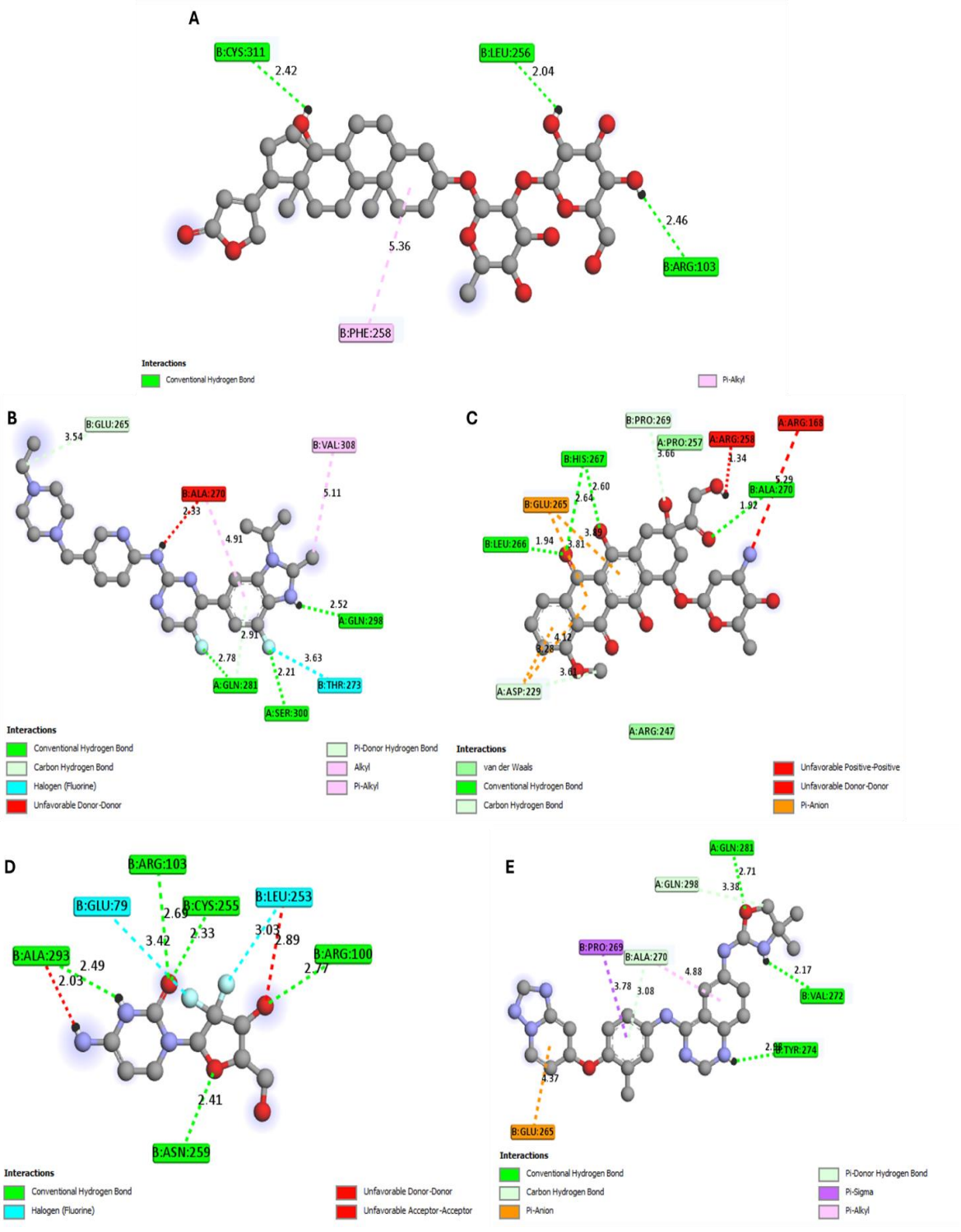


Figure 4.9: 2D representation of HER2 (7MN5) interaction with (A) NCGC00385704-01, and the standard drugs such as (B) Abemaciclib, (C) Doxorubicin, (D) Gemcitabine, and (E) Tucatinib

For the mutated EGFR (PDB ID: 7ZYM), Tucatinib exhibited the strongest binding affinity (–10.1 kcal/mol, K_i = 0.0401 μM), followed by Doxorubicin (–9.5 kcal/mol, K_i = 0.1104 μM) and Abemaciclib (–9.2 kcal/mol, K_i = 0.1832 μM). The predicted compound NCGC00385704-01 also demonstrated moderate inhibitory activity (–8.0 kcal/mol, K_i = 1.3890 μM), surpassing Gemcitabine (–6.1 kcal/mol, Ki = 34.3194 μM) (**Table 4.8**). Although standard drugs showed overall superior binding to EGFR, binding mode analyses **(Figure 4.10)** highlighted key insights into the selectivity of molecules targeting the EGFR-T790M/C797S binding site.

This binding site primarily mediates adenosine triphosphate (ATP) binding and phosphorylation, thereby activating PI3K/AKT and MAPK/ERK pathways to promote cell proliferation and survival [173]. The T790M mutation enhances ATP affinity, reducing the effectiveness of first-generation tyrosine kinase inhibitors (TKIs), while the C797S mutation hinders covalent binding of third-generation inhibitors such as Osimertinib, underscoring the need for novel agents to block aberrant signalling. Critical residues for mutant EGFR selectivity include LYS745, MET790, MET793, SER797, and ASP800, which contribute to conformational changes in the binding pocket that suppress tumour progression [174].

Among the studied compounds, NCGC00385704-01 interacted with catalytic residue LYS745 through its carbonyl oxygen, resembling binding modes observed in reference inhibitors such as Doxorubicin and Gemcitabine **(Figure 4.10)** [174]. Abemaciclib established hydrogen bonds with ASP800 and ASP855, along with a π–sigma interaction involving LEU718, whereas Gemcitabine showed an unfavourable donor–donor interaction with LYS745, likely explaining its weaker activity. The strong binding affinity of Tucatinib may be attributed to hydrogen bonding between its amine bridge and the DFG motif residue ASP855.

Hydrogen bonding emerged as a crucial factor influencing binding strength. The compound NCGC00385704-01 formed three hydrogen bonds, comparable to standard inhibitors such as Abemaciclib, Doxorubicin, and Tucatinib **(Figure 4.10).** Furthermore, the sandwich motif observed between MET790 and LYS745 suggests promising strategies for enhancing inhibitory activity against T790M-mutated EGFR and overcoming C797S resistance, commonly associated with non-small cell lung cancer (NSCLC) [175]. Collectively, these results support the potential for designing reversible inhibitors against EGFR resistance mutations, offering valuable avenues for precision oncology.

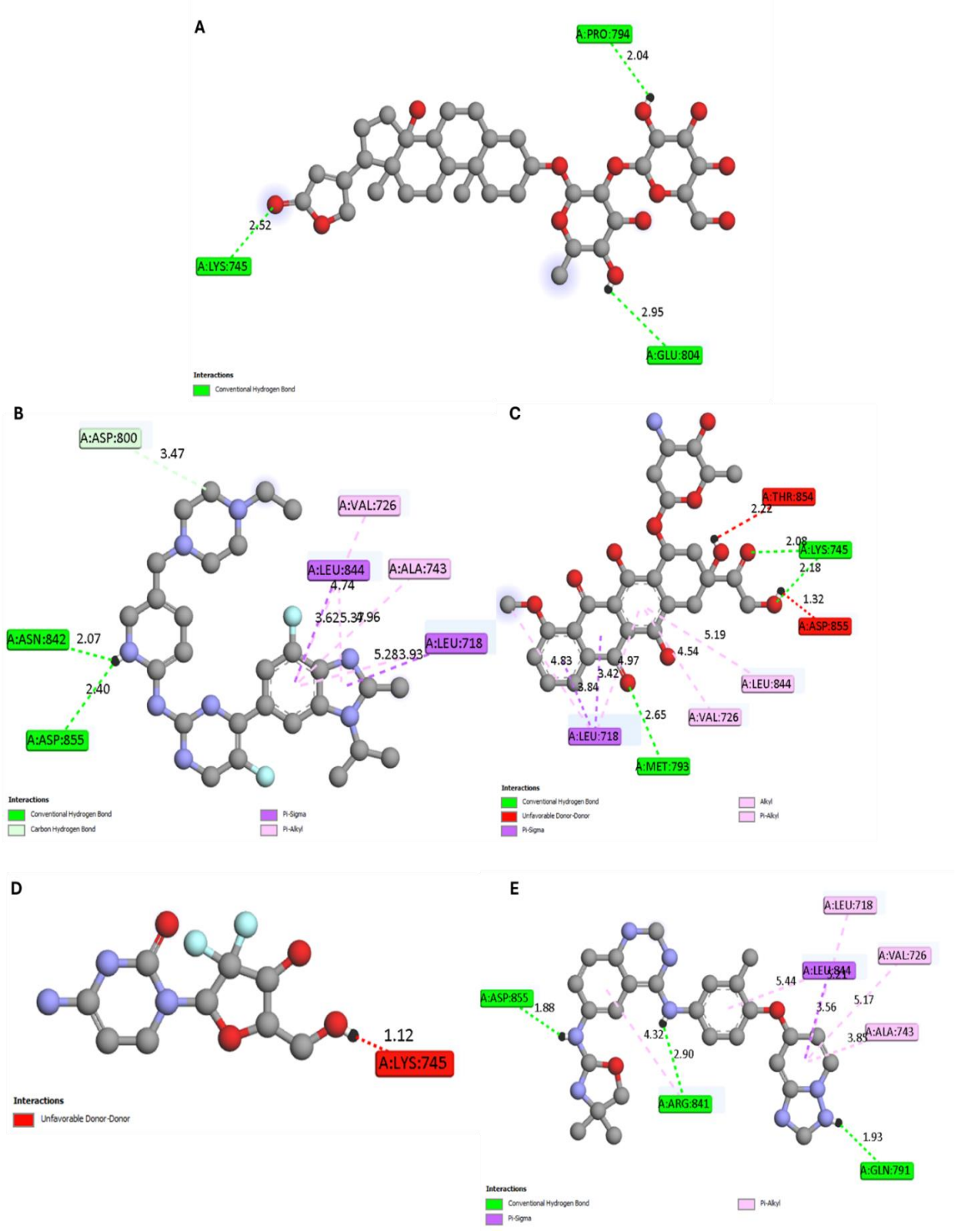


Figure 4.10: 2D representation of EGFR (7ZYM) interaction with (A) NCGC00385704-01, and the standard drugs such as (B) Abemaciclib, (C) Doxorubicin, (D) Gemcitabine, and (E) Tucatinib

4.9. Analysis of ADMET properties

The ADMET profiles of the predicted compound (NCGC00385704-01) obtained from virtual screening, along with four FDA-approved reference drugs, were evaluated using the Swiss ADME and ADMETlab 2.0 web servers. Interpretation of the results followed the standard criteria provided by ADMETlab. Oral bioavailability was assessed in accordance with Lipinski's Rule of Five, which considers molecular weight (MW ≤ 500 Da), lipophilicity (logP ≤ 5), hydrogen bond donors (HBDs ≤ 5), and hydrogen bond acceptors (HBAs ≤ 10). Compounds that violate more than one of these criteria are generally associated with reduced absorption. The absorption-related parameters analysed included human intestinal absorption (HIA), solubility classification, and Caco-2 cell permeability (C2P). Excretion-related factors such as clearance rate and biological half-life were also examined. A detailed summary of the pharmacokinetic characteristics of the compounds is presented in **Table 4.10.**

Table 4.10: ADMET properties of the predicted and reference compounds from molecular docking study

Compounds	Physicochemical						Medicinal	Absorption range		Excretion		Solubility	
	MW	**nHBA**	**nHBD**	**RBN**	**cLogP**	**TPSA**	**Lipinski**	**C2P**	**HIA**	**Clearance**	**T1/2**	**ESOL (LogS)**	**Solubility Class**
Abemaciclib	492.26	8	1	6	3.45	75.00	Accepted	-4.90	High	9.640	0.072	-5.12	Moderately soluble
Doxorubicin	543.17	12	7	5	2.012	206.07	Rejected	-6.07	low	17.278	0.046	-3.91	Soluble
Gemcitabine	264.23	19	10	9	1.54	305.67	Rejected	-6.21	low	9.899	0.051	-0.67	Very soluble
NCGC00385704-01	682.36	13	8	6	1.90	211.90	Rejected	-6.31	low	0.779	0.617	-4.00	Soluble
Tucatinib	480.20	10	2	6	5.20	110.85	Accepted	-4.86	High	6.104	0.321	-5.45	Moderately soluble

MW: molecular weight; nHBA: number of hydrogen bond acceptors; nHBD: number of hydrogen bond donors; RBN: number of rotatable bonds; cLogP: calculated logarithm of the partition coefficient; a measure of the lipophilicity of a molecule; C2P: Caco-2 cell permeability; HIA: human intestinal absorption; T1/2: half-life; ESOL: Estimated solubility.

All the virtually screened hit compounds were predicted to be non-toxic including NCGC00385704-01, showing favourable physicochemical properties with low clogP values (< 3) and high topological polar surface area (TPSA > 75). These features align with Pfizer's rule, which states that compounds with clogP > 3 and TPSA < 75 are more likely to exhibit toxicity. Among the reference drugs, Abemaciclib and Tucatinib complied with Lipinski's Rule of Five, demonstrating moderate water solubility, and showed high human intestinal absorption, suggesting good oral bioavailability. In contrast, Doxorubicin and Gemcitabine violated more than one of Lipinski's parameters, displaying high solubility but poor predicted intestinal absorption. The virtually screened hit (NCGC00385704-01) also violated multiple Lipinski rules, with predictions of low solubility and poor absorption. Abemaciclib and Tucatinib further displayed acceptable Caco-2 permeability values, supporting their improved absorption characteristics. With respect to excretion, Doxorubicin showed higher clearance rates, while Gemcitabine, Abemaciclib, and Tucatinib exhibited moderate clearance.

4.10. Antioxidant activity of *Ceratonia siliqua* L. pod extract

The antioxidant activity of the *Ceratonia siliqua* L. (CB) pod extract was evaluated using the DPPH assay to measure the radical scavenging activity (RSA). The results in **Figure 4.11** demonstrated that the *Ceratonia siliqua* L. (CB) extract showed a significant antioxidant activity, with a dose-dependent increase in DPPH inhibition, thereby suppressing the DPPH radical scavengers. At the highest measured concentration (10 µg/mL), the extract was found to have an assessed DPPH radical scavenging activity (RSA) of 92-95%, indicating strong free radical neutralisation capacity. Although the performance was slightly lower than that of the positive

control (ascorbic acid), it still indicates that the extract contains significant antioxidant components. As shown, it almost attained near-complete inhibition (100%) at a lower dose **(Figure 4.11)**. This is evidenced by the calculated IC_{50} values that are the concentration required to inhibit 50% of DPPH radicals, which are 1.74 ± 0.25 µg/mL for the *Ceratonia siliqua* L. (CB) extract, compared to that of ascorbic acid, which exhibited the IC_{50} value of 0.20 ± 0.07 µg/mL, highlighting its notable antioxidant strength **(Table 4.11).**

These results are clearly illustrated in **Table 4.10** and further represented in the dose-response curve shown in **Figure 4.11.** The superior antioxidant capacity of the *Ceratonia siliqua* L. (CB) extract is linked to the rich content of polyphenolic compounds identified in the phytochemical profile of *C. siliqua*.

Table 4.11: Antioxidant activity of *Ceratonia siliqua* L. (CB) pod extracts and of positive control (ascorbic acid) against DPPH radicals.

Samples	**IC_{50} (µg/mL ± SD)**
	DPPH
Ceratonia siliqua L. (CB)	1.74 ± 0.25
Ascorbic acid	0.20 ± 0.07

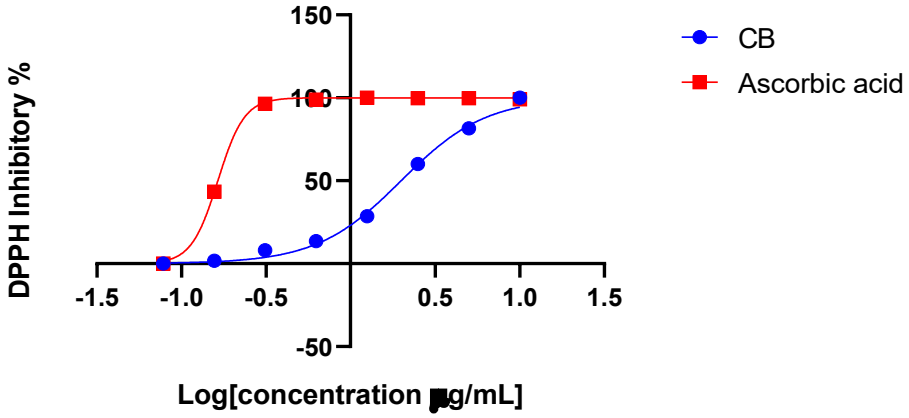


Figure 4.11: A dose-response curve generated using a sigmoid regression model to determine the IC_{50} values associated with antioxidant activity

Polyphenolic compounds function as potent antioxidants, primarily through their ability to donate hydrogen atoms or electrons, thereby neutralising free radicals and reducing oxidative stress. This is extremely relevant in cancer chemoprevention, since oxidative stress caused by these free radicals plays a critical role in DNA damage, mutation, and subsequent tumour development through different mechanisms such as radical scavenging and chelating divalent cations involved in Fenton chemistry [81]. Diets rich in polyphenols, found in plant-based foods like fruits, vegetables, tea, coffee, and chocolate, can improve health and reduce the risk of chronic diseases, including the development of cancer. These dietary polyphenols can also prevent injury caused by free radicals and block the initiation step of cancer [82]. Therefore, by mitigating oxidative damage induced by free radicals, natural antioxidants present in *Ceratonia siliqua* L, in conjunction with its fibre content, may contribute to a reduced risk of colorectal cancer (CRC) and offer therapeutic support.

4.11. Cytotoxic activity of *Ceratonia siliqua* L. pod extract on colorectal cancer and normal cell line

The pods (comprising both seed and pulp) of *Ceratonia siliqua* L. are rich in bioactive compounds, particularly polyphenols, and have been extensively reported for their chemopreventive and therapeutic effects against cancer [18]. This highlights the likelihood of anti-cancer constituents within the plant pods. The therapeutic action of *Ceratonia siliqua* L. in colorectal cancer has been attributed to the interaction between its polyphenols and dietary fibre, which not only safeguards polyphenols during digestion but also enhances their bioactivity in the gut and colon [34], [83], [84]. The pod extract of *Ceratonia siliqua* L., evaluated in this study, showed selective cytotoxicity against HCT116 colorectal cancer cells, with minimal effects on Vero normal kidney epithelial cells, according to MTT assay results. In accordance with the American National Cancer Institute (NCI), a crude extract is considered to possess notable cytotoxic potential when its IC_{50} value falls below 20 µg/mL [176]. As presented in **Table 4.12**, the *Ceratonia siliqua* L. pod extract recorded an IC_{50} value of 13.32 ± 1.09 µg/mL against HCT116 cells, placing it within the NCI threshold and confirming strong anti-cancer potential. Conversely, the pod extract showed reduced toxicity in Vero cells, with an IC_{50} of 21.39 ± 1.30 µg/mL. For comparison, the FDA-approved anti-cancer drug Doxorubicin, used as a positive control, demonstrated IC_{50} values of 1.85 ± 0.21 µg/mL for HCT116 and 1.61 ± 0.26 µg/mL for Vero cells. These results collectively suggest that the *Ceratonia siliqua* L. pod extract preferentially inhibits cancerous cells while showing minimal effects on normal cells.

Table 4.12: Calculated IC_{50} values of Ceratonia siliqua L. pod extracts and of Doxorubicin indicating cytotoxic activity against Vero and HCT116 cell lines

Samples	IC_{50} (µg/mL ± SD)	
	Vero	**HCT116**
Ceratonia siliqua L. (CB)	21.39 ± 1.30	13.32 ± 1.09
Doxorubicin	1.61 ± 0.26	1.85 ± 021

The dose–response curves **(Table 4.13)** illustrate a comparison between the *Ceratonia siliqua* L. pod extract and Doxorubicin, a standard chemotherapeutic drug. As depicted, both treatments caused a progressive reduction in cell viability with increasing concentrations, confirming a dose-dependent cytotoxic effect. Doxorubicin demonstrated pronounced cytotoxicity at lower concentrations, whereas the *C. siliqua* pod extract showed moderate cytotoxicity against HCT116 cells, highlighting its notable anti-cancer potential. Doxorubicin was used as a positive control in the MTT assay as a standard reference anticancer agent to benchmark overall cytotoxic activity of the extract against cancer cells. Its inclusion was intended for comparison of general antiproliferative effects rather than target-specific EGFR/HER2 inhibition, which was assessed separately through the computational analysis. Overall, these MTT assay findings indicate that the *C. siliqua* pod extract possesses anti-cancer activity, most likely mediated by bioactive phytochemicals that interfere with cancer cell metabolism [177-178]. Statistical analysis (one-way

ANOVA followed by appropriate post-hoc comparison) was applied to compare treated groups. In addition, the selectivity index (SI) was calculated as the ratio of IC_{50} (Vero normal cells) to IC_{50} (HCT116 cancer cells) to assess preferential cytotoxicity of the extract toward cancer cells relative to normal cells.

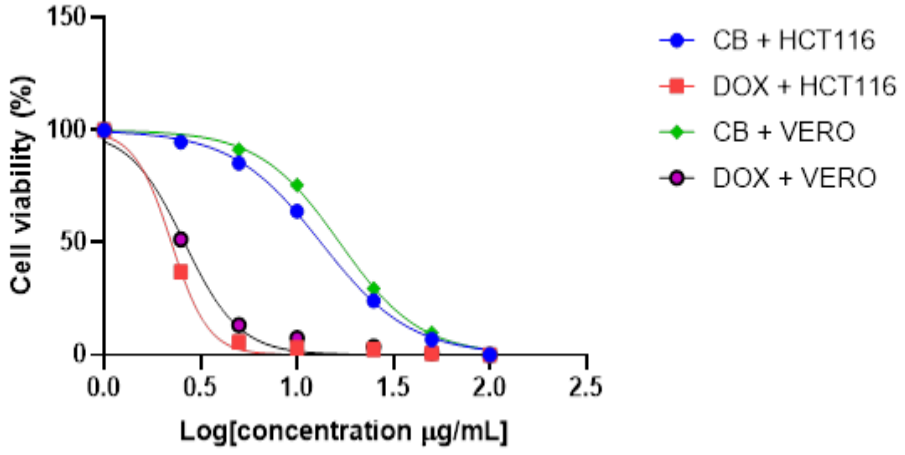


Figure 4.12: A dose-response curve generated using a sigmoid regression model to determine the IC_{50} values associated with cytotoxic activity.

These results align with earlier reports demonstrating the dose-dependent antiproliferative effects of the *Ceratonia siliqua* L. extract on different cancer cell lines. For instance, Hsouna et al. (2011) documented comparable cytotoxic activities of *C. siliqua* and other Mediterranean plant extracts against cancer cells [177-178]. Similarly, Amessis-Ouchemoukh et al. (2017) identified bioactive metabolites in *Ceratonia siliqua* L. extracts that contribute to their antioxidant, anti-cancer, and anti-calpain properties [20].

4.12. Limitations

The study is subject to certain limitations. The number of compounds identified through LC–MS/MS was relatively low, which may be attributed to extraction selectivity, ionisation efficiency, and spectral library coverage constraints. In addition, compound annotations were largely based on GNPS spectral matching, which provides putative identifications and therefore carries inherent limitations in confirmation confidence in the absence of authentic standards. Furthermore, molecular docking results are predictive in nature and are limited by the static representation of protein structures, scoring function approximations, and the accuracy of available crystal structures used in the analysis.

CHAPTER 5: OVERALL CONCLUSION AND RECOMMENDATIONS FOR FUTURE RESEARCH

5.1. Overall conclusion

This study successfully combined LC–MS/MS phytochemical profiling with machine learning-based virtual screening to identify potential anticancer compounds from *Ceratonia siliqua* L., targeting EGFR and HER2 receptors in colorectal cancer therapy.

A curated dataset of known anti-cancer and anti-inflammatory compounds was compiled from ChEMBL, focusing on inhibitors of EGFR and HER2. Using these data, multiple machine learning models were developed and optimised, with a stacking ensemble achieving exceptional predictive performance, recording 99.5% accuracy on the training data and 100% accuracy, F1 score, sensitivity, specificity, Matthews Correlation Coefficient (MCC), and area under the curve (AUC) on the test dataset. By integrating outputs from 40 baseline models, the ensemble method significantly outperformed individual models such as the Multilayer Perceptron (MLP) and Extra Trees (ET), which attained 92.9% and 92.6% accuracy, respectively. These findings underscore the superiority of ensemble learning for identifying dual inhibitors of critical therapeutic targets.

Phytochemical profiling of *Ceratonia siliqua* L. pods by LC–MS/MS revealed diverse bioactive constituents, which were subsequently screened using the built stacking ensemble model. Virtual screening identified a single promising candidate, namely *3-[(3S,8R,10S,13R,14S,17R)-3-[(2R,3R,4R,5R,6S)-4,5-dihydroxy-6-methyl-3-[(2S,3R,4S,5S,6R)-3,4,5-trihydroxy-6-(hydroxymethyl)oxan-2-yl]oxyoxan-2-yl]oxy-14-hydroxy-10,13-dimethyl-1,2,3,4,5,6,7,8,9,11,12,15,16,17-tetradecahydrocyclopenta[a]phenanthren-17-yl]-2H-furan-5-one* (NCGC00385704-01) with potential dual inhibitory activity against EGFR and HER2. Molecular docking supported the virtual screening results, with selected compounds

demonstrating favourable binding interactions within the target active sites when compared to reference compounds.

Finally, in vitro assays supported the computational findings. The MTT assay demonstrated cytotoxic activity of *C. siliqua* pod extracts against HCT116 colorectal cancer cells, with preferential effects relative to Vero normal cells as indicated by the calculated selectivity index (SI). Antioxidant activity further supports the biological relevance of the extract.

Collectively, these results demonstrate that integrating LC–MS/MS analysis with machine learning provides a framework for identifying potential EGFR and HER2 inhibitors from natural sources. These findings position *Ceratonia siliqua* L. as a promising source of bioactive compounds for anticancer drug discovery and highlight the value of combining computational and experimental approaches.

5.2. Recommendations for future research

The dual inhibition strategy presented in this work enhances therapeutic potential while offering a foundation for more personalised approaches to colorectal cancer treatment. By combining advanced computational tools with the chemical diversity of natural products, this research opens avenues for the development of effective and sustainable therapeutic options.

Future studies should focus on validating the predicted mechanisms of action through targeted biological assays. Specifically, apoptosis induction can be assessed using flow cytometry, while cell cycle arrest may be evaluated through cell cycle analysis. In addition, intracellular reactive oxygen species (ROS) generation can be quantified using fluorescence-based assays to determine

oxidative stress involvement. These targeted experiments would provide mechanistic insight into the cytotoxic effects observed in this study.

Furthermore, NCGC00385704-01 was tentatively annotated as a terpenoid-like compound based on GNPS spectral library matching (Level 2 confidence). As no authentic or MS/MS reference standard was available, future work should prioritise structural confirmation using purified isolates and complementary analytical techniques such as NMR spectroscopy to improve annotation confidence.

REFERENCES

[1] J. Patarra, "Evaluation of the in vitro biological activities of extracts from carob tree and mediterranean oaks," Master's thesis, 2009, University of the Algarve, Faro, Portugal, sapiential.ualg.pt/handle/10400.1/336

[2] J. Liang, Y. Zheng, X. Tong, N. Yang, and S. Dai, "In Silico Identification of Anti-SARS-CoV-2 Medicinal Plants Using Cheminformatics and Machine Learning," *Molecules*, 2023, doi: 10.3390/molecules28010208.

[3] A. G. Atanasov *et al.*, "Natural products in drug discovery: advances and opportunities," *Nature Reviews Drug Discovery 2021*, doi: 10.1038/s41573-020-00114-z.

[4] D. J. Newman and G. M. Cragg, "Natural Products as Sources of New Drugs from 1981 to 2014," *J. Nat. Prod.*, 2016, doi: 10.1021/ACS.JNATPROD.5B01055.

[5] V. Steenkamp and M. C. Gouws, "Cytotoxicity of six South African medicinal plant extracts used in the treatment of cancer," *South African Journal of Botany*, 2006, doi: 10.1016/j.sajb.2006.02.004.

[6] S. Agajanian, "Development of Integrated Machine Learning and Data Science Approaches for the Prediction of Cancer Mutation and Autonomous Drug Discovery of Anti-Cancer

Therapeutic Agents," PhD dissertation, Chapman University, Orange, CA, USA, 2021, doi.org/10.36837/chapman.000220

[7] E. N. Feinberg, A. B. Farimani, R. Uprety, A. Hunkele, G. W. Pasternak, S. Majumdar and V. S. Pande, "Machine learning harnesses molecular dynamics to discover new μ opioid chemotypes," *Eprint arXiv*:1803.04479, 2018, doi: arxiv.o rg/abs/1803.04479.

[8] R. Burbidge, M. Trotter, B. Buxton, and S. Holden, "Drug design by machine learning: support vector machines for pharmaceutical data analysis," *Comput. Chem.*, 2001, doi: 10.1016/S0097-8485(01)00094-8.

[9] P. J. Ballester and J. B. O. Mitchell, "A machine learning approach to predicting protein–ligand binding affinity with applications to molecular docking," *Bioinformatics*, 2010, doi: 10.1093/BIOINFORMATICS/BTQ112.

[10] P. M. Kasson and S. Jha, "Adaptive ensemble simulations of biomolecules," *Curr. Opin. Struct. Biol.*, 2018, doi: 10.1016/J.SBI.2018.09.005.

[11] K. A. Carpenter and X. Huang, "Machine Learning-based Virtual Screening and Its Applications to Alzheimer's Drug Discovery: A Review," *Curr. Pharm. Des.*, 2018, doi: 10.2174/1381612824666180607124038.

[12] S. Gezici and N. Şekeroğlu, "Current Perspectives in the Application of Medicinal Plants Against Cancer: Novel Therapeutic Agents," *Anticancer Agents Med. Chem.*, 2019, doi: 10.2174/1871520619666181224121004.

[13] Z. Nazarian-Samani, R. D. E. Sewell, Z. Lorigooini, and M. Rafieian-Kopaei, "Medicinal Plants with Multiple Effects on Diabetes Mellitus and Its Complications: a Systematic Review," *Curr. Diab. Rep.*, 2018, doi: 10.1007/S11892-018-1042-0/TABLES/2.

[14] S. Tasneem, B. Liu, B. Li, M. I. Choudhary, and W. Wang, "Molecular pharmacology of inflammation: Medicinal plants as anti-inflammatory agents," *Pharmacol. Res.*, 2019, doi: 10.1016/J.PHRS.2018.11.001.

[15] A. G. Desai *et al.*, "Medicinal Plants and Cancer Chemoprevention," *Curr. Drug Metab.*, 2008, doi: 10.2174/138920008785821657.

[16] M. Greenwell and P. K. S. M. Rahman, “Medicinal Plants: Their Use in Anticancer Treatment.,” *Int. J. Pharm. Sci. Res.*, 2015, doi: 10.13040/IJPSR.0975-8232.6(10).4103-12.

[17] H. Yuan, Q. Ma, L. Ye, and G. Piao, “The Traditional Medicine and Modern Medicine from Natural Products,” *Molecules*, 2016, doi: 10.3390/MOLECULES21050559.

[18] A. S. Azmi, S. H. Bhat, S. Hanif, and S. M. Hadi, “Plant polyphenols mobilize endogenous copper in human peripheral lymphocytes leading to oxidative DNA breakage: A putative mechanism for anticancer properties,” *FEBS Lett.*, 2006, doi: 10.1016/J.FEBSLET.2005.12.059.

[19] B. B. Aggarwal and P. Gehlot, “Inflammation and cancer: how friendly is the relationship for cancer patients?,” *Curr. Opin. Pharmacol.*, 2009, doi: 10.1016/j.coph.2009.06.020.

[20] N. Amessis-Ouchemoukh *et al.*, “Bioactive metabolites involved in the antioxidant, anticancer and anticalpain activities of Ficus carica L., Ceratonia siliqua L. and Quercus ilex L. extracts,” *Ind. Crops Prod.*, 2017, doi: 10.1016/J.INDCROP.2016.10.007.

[21] N. Takegawa and K. Yonesaka, “HER2 as an Emerging Oncotarget for Colorectal Cancer Treatment After Failure of Anti-Epidermal Growth Factor Receptor Therapy,” *Clin. Colorectal Cancer*, 2017, doi: 10.1016/j.clcc.2017.03.001.

[22] A. Gupta *et al.*, “EGFR-directed antibodies promote HER2 ADC internalization and efficacy,” *Cell Rep. Med.*, 2024, doi: 10.1016/J.XCRM.2024.101792.

[23] M. Zhu *et al.*, “Advancements in the application of artificial intelligence in the field of colorectal cancer,” *Front. Oncol.*, 2025, doi: 10.3389/FONC.2025.1499223/TEXT.

[24] I. J. Sagbo and W. Otang-Mbeng, “Plants Used for the Traditional Management of Cancer in the Eastern Cape Province of South Africa: A Review of Ethnobotanical Surveys, Ethnopharmacological Studies and Active Phytochemicals,” *Molecules*, 2021, doi: 10.3390/MOLECULES26154639.

[25] J. S. Macdonald, “Toxicity of 5-fluorouracil” *Oncology (Williston Park)*, 1999 Jul;13(7 Suppl 3):33-4. PMID: 10442356.

[26] M. Barary *et al.*, "The effect of propolis on 5-fluorouracil-induced cardiac toxicity in rats," *Scientific Reports,* 2022, doi: 10.1038/s41598-022-12735-y.

[27] S. Kilickap, E. Akgul, S. Aksoy, K. Aytemir, and I. Barista, "Doxorubicin-induced second degree and complete atrioventricular block," *Europace*, 2005, doi: 10.1016/J.EUPC.2004.12.012.

[28] L. Manil, P. Mahieu, and P. Couvreur, "Acute renal toxicity of doxorubicin (adriamycin)-loaded cyanoacrylate nanoparticles," *Pharm. Res.*, 1995, doi: 10.1023/A:1016290704772.

[29] S. Gibaud, J. P. Andreux, C. Weingarten, M. Renard, and P. Couvreur, "Increased bone marrow toxicity of doxorubicin bound to nanoparticles," *Eur. J. Cancer*, 1994, doi: 10.1016/0959-8049(94)90299-2.

[30] I. Y. R. Adamson, "Pulmonary toxicity of bleomycin," *Environ. Health Perspect.*, 1976, doi: 10.2307/3428592.

[31] O. J. Wouters, M. McKee, and J. Luyten, "Estimated Research and Development Investment Needed to Bring a New Medicine to Market, 2009-2018," *JAMA*, 2020, doi: 10.1001/JAMA.2020.1166.

[32] N. Lachkar, H. El, H. Sidi, M. Ben, and M. Lachkar, "Anti-inflammatory and antioxidant effect of Ceratonia siliqua L. Methanol barks extract," *Journal of Chemical and Pharmaceutical Research*, 2016.

[33] K. K. Mak and M. R. Pichika, "Artificial intelligence in drug development: present status and future prospects," *Drug Discov. Today*, 2019, doi: 10.1016/J.DRUDIS.2018.11.014.

[34] A. Ikram *et al.*, "Nutritional, biochemical, and clinical applications of carob: A review," *Food Sci. Nutr.*, 2023, doi: 10.1002/FSN3.3367.

[35] E. Papaefstathiou, A. Agapiou, S. Giannopoulos, and R. Kokkinofta, "Nutritional characterization of carobs and traditional carob products," Food Science and Nutrition, 2018, doi: 10.1002/fsn3.776.

[36] Encyclopaedia *Britannica*, "Carob | Health Benefits, Nutritional Value & Culinary Uses," *Encyclopaedia Britannica*, 2024. URL: https://britannica.com/plant/carob. (Accessed, May 05, 2024).

[37] P. A. Dakia, "Carob (Ceratonia siliqua L.) Seeds, Endosperm and Germ Composition, and Application to Health," *Nuts and Seeds in Health and Disease Prevention*, 2011, doi: 10.1016/B978-0-12-375688-6.10035-0.

[38] S. S. Azimova and A. I. Glushenkova, "Ceratonia siliqua L.," *Lipids, Lipophilic Components and Essential Oils from Plant Sources*, 2012, doi: 10.1007/978-0-85729-323-7_1829.

[39] M. E. Brassesco, T. R. S. Brandão, C. L. M. Silva, and M. Pintado, "Carob bean (Ceratonia siliqua L.): A new perspective for functional food," *Trends Food Sci. Technol.*, 2021, doi: 10.1016/J.TIFS.2021.05.037.

[40] E. Batal *et al.*, "Assessment of nutritional composition of Carob pulp (Ceratonia Siliqua L.) collected from various locations in Morocco," *J. Mater. Environ. Sci., 2016. URL: https://www.jmaterenvironsci.com/Document/vol7/vol7_N9/339-JMES-2341-El%20Batal.pdf*. (Accessed, May 05, 2024).

[41] I. Batlle and J. Tous, "Carob Tree (*Ceratonia siliqua* L.) Promoting the Conservation and Use of Under-Utilised and Neglected Crops," Institute of Plant Genetics and Crop Plant Research, Gatersleben/International Plant Genetic Resources Institute, Rome, Italy. URL: https://www.scirp.org/reference/referencespapers?referenceid=2824410, (Accessed, May 05, 2024)

[42] W. Dahmani *et al.*, "Exploring Carob (Ceratonia siliqua L.): A Comprehensive Assessment of Its Characteristics, Ethnomedicinal Uses, Phytochemical Aspects, and Pharmacological Activities," *Plants*, 2023, doi: 10.3390/PLANTS12183303.

[43] A. Azab, "Carob Antioxidants in Human Health: From Traditional Uses to Modern Pharmacology," *J. Biomed. Res. Environ. Sci.,* 2022, doi: 10.37871/jbres1538.

[44] W. Dahmani *et al.*, "Exploring Carob (Ceratonia siliqua L.): A Comprehensive Assessment of Its Characteristics, Ethnomedicinal Uses, Phytochemical Aspects, and Pharmacological Activities," *Plants*, 2023, doi: 10.3390/PLANTS12183303.

[45] W. S. Darwish *et al.*, "Phytochemical characterization and evaluation of biological activities of egyptian carob pods (Ceratonia siliqua l.) aqueous extract: In vitro study," *Plants*, 2021, doi: 10.3390/plants10122626.

[46] E. M. Al-Olayan *et al.*, "Ceratonia siliqua pod extract ameliorates Schistosoma mansoni-induced liver fibrosis and oxidative stress," *BMC Complement. Altern. Med.*, 2016, doi: 10.1186/S12906-016-1389-1/FIGURES/7.

[47] Fadel, F. & Fattouch, Sami & Tahrouch, Saïda & Lahmar, R. & Benddou, A. & Hatimi, Abdelhakim. (2011). The phenolic compounds of Ceratonia siliqua pulps and seeds. *Journal of Materials and Environmental Science*, 2011 2. 285-292.

[48] R. Avallone, M. Plessi, M. Baraldi, and A. Monzani, "Determination of chemical composition of carob (*Ceratonia siliqua*): Protein, fat, carbohydrates, and tannins," *J. Food Compos. Anal.,* 1997, doi: 10.1006/jfca.1997.

[49] R. W. Owen *et al.*, "Isolation and structure elucidation of the major individual polyphenols in carob fibre," *Food and Chemical Toxicology*, 2003, doi: 10.1016/S0278-6915(03)00200-X.

[50] K. Rtibi *et al.*, "Ceratonia siliqua leaves exert a strong ROS-scavenging effect in human neutrophils, inhibit myeloperoxydase in vitro and protect against intestinal fluid and electrolytes secretion in rats," *RSC Adv.*, 2016, doi: 10.1039/c6ra11297h.

[51] H. El Hajaji *et al.*, "Antioxidant activity, phytochemical screening, and total phenolic content of extracts from three genders of carob tree barks growing in Morocco," *Arabian Journal of Chemistry*, 2011, doi: 10.1016/J.ARABJC.2010.06.053.

[52] I. J. Stavrou, A. Christou, and C. P. Kapnissi-Christodoulou, "Polyphenols in carobs: A review on their composition, antioxidant capacity and cytotoxic effects, and health impact," *Food Chem.*, 2018, doi: 10.1016/j.foodchem.2018.06.152.

[53] A. Benyaich, A. Nouayti, O. El Mekki, R. Errafia, F. Bahtat, and M. Aksissou, "Phytochemical and pharmacological properties of Ceratonia siliqua L.: A comparative review of Moroccan and Mediterranean varieties," *J. Appl. Pharm. Sci.*, 2025, doi: 10.7324/JAPS.2025.237923.

[54] F. Z. GHANEMI and M. BELARBI, "Phytochemistry and Pharmacology of Ceratonia siliqua L. leaves," *Journal of Natural Product Research and Applications*, 2021, doi: 10.46325/JNPRA.V1I01.7.

[55] B. de Falco, L. Grauso, A. Fiore, G. Bonanomi, and V. Lanzotti, "Metabolomics and chemometrics of seven aromatic plants: Carob, eucalyptus, laurel, mint, myrtle, rosemary and strawberry tree," *Phytochemical Analysis*, 2022, doi: 10.1002/pca.3121.

[56] M. A. Farag *et al.*, "Variation in Ceratonia siliqua pod metabolome in context of its different geographical origin, ripening stage and roasting process," *Food Chem.*, 2019, doi: 10.1016/J.FOODCHEM.2018.12.118.

[57] O. Samia, M. Nesrine, Z. K. Feten, and K. Riadh, "Tunisian Ceratonia siliqua: Phytochemical Analysis, Antioxidant Activity, Preparation and Characterization of Carob Emulsion System," *Eur. J. Nutr. Food Saf.*, 2022, doi: 10.9734/EJNFS/2022/V14I230479.

[58] Z. Basharat *et al.*, "Nutritional and functional profile of carob bean (Ceratonia siliqua): a comprehensive review," *Int. J. Food Prop.*, 2023, doi: 10.1080/10942912.2022.2164590.

[59] S. Abidar *et al.*, "The Aqueous Extract from Ceratonia siliqua Leaves Protects against 6-Hydroxydopamine in Zebrafish: Understanding the Underlying Mechanism," *Antioxidants,* 2020, doi: 10.3390/ANTIOX9040304.

[60] U. G. Spizzirri *et al.*, "Kefir Enriched with Carob (Ceratonia siliqua L.) Leaves Extract as a New Ingredient during a Gluten-Free Bread-Making Process," *Fermentation*, 2022, doi: 10.3390/FERMENTATION8070305/S1.

[61] D. M. Rasheed, D. M. El-Kersh, and M. A. Farag, "Ceratonia siliqua (Carob-Locust Bean) outgoing and potential trends of phytochemical, economic and medicinal merits," *Wild Fruits: Composition, Nutritional Value and Products*, 2019, doi: 10.1007/978-3-030-31885-7_36.

[62] M. E. Brassesco, T. R. S. Brandão, C. L. M. Silva, and M. Pintado, "Carob bean (Ceratonia siliqua L.): a new perspective for functional food," *Trends Food Sci. Technol.*, 2021, doi: 10.1016/J.TIFS.2021.05.037.

[63] K. Yahiaoui *et al.*, "Characterization and assessment of the antimicrobial function of total polyphenol extracts from pulps, leaves and seeds of two Ceratonia siliqua L. varieties," *Algerian Journal of Environmental Science and Technology*, 2024. [Online]. URL: https://www.aljest.net/index.php/aljest/article/view/395. (Accessed, May 05, 2024).

[64] I. Karmous *et al.*, "Phytosynthesis of Zinc Oxide Nanoparticles Using Ceratonia siliqua L. and Evidence of Antimicrobial Activity," *Plants*, 2022, doi: 10.3390/PLANTS11223079/S1.

[65] M. H. Shahrajabian and W. Sun, "Carob (Ceratonia siliqua L.), Pharmacological and Phytochemical Activities of Neglected Legume of the Mediterranean Basin, as Functional Food," *Rev. Recent Clin. Trials*, 2024, doi: 10.2174/0115748871278128240109074506.

[66] K. Rtibi *et al.*, "Chemical constituents and pharmacological actions of carob pods and leaves (Ceratonia siliqua L.) on the gastrointestinal tract: A review," *Biomedicine and Pharmacotherapy*, 2017, doi: 10.1016/j.biopha.2017.06.088.

[67] I. Abulyazid, S. A. Abd Elhalim, H. M. Sharada, W. M. Aboulthana, and S. T. A. Abd Elhalim, "Hepatoprotective Effect of Carob Pods Extract (Ceratonia siliqua L.) against Cyclophosphamide Induced Alterations in Rats," *International Journal of Current Pharmaceutical Review and Research*, 2017, doi: 10.25258/ijcprr.v8i02.9199.

[68] D. Rico *et al.*, "In vitro approach for evaluation of carob by-products as source bioactive ingredients with potential to attenuate metabolic syndrome (MetS)," *Heliyon*, 2019, doi: 10.1016/J.HELIYON.2019.E01175.

[69] R. Rodríguez-Solana, N. Coelho, A. Santos-Rufo, S. Gonçalves, E. Pérez-Santín, and A. Romano, "The Influence of In Vitro Gastrointestinal Digestion on the Chemical Composition and Antioxidant and Enzyme Inhibitory Capacities of Carob Liqueurs Obtained with Different Elaboration Techniques," *Antioxidants (Basel)*, 2019, doi: 10.3390/ANTIOX8110563.

[70] C. Christou, E. Poulli, S. Yiannopoulos, and A. Agapiou, "GC-MS analysis of D-pinitol in carob: Syrup and fruit (flesh and seed)," *J. Chromatogr. B Analyt. Technol. Biomed. Life Sci.*, 2019, doi: 10.1016/J.JCHROMB.2019.04.008.

[71] G. O. Sigge, L. Iipumbu, and T. J. Britz, "South African Journal of Plant and Soil Proximate composition of carob cultivars growing in South Africa Proximate composition of carob cultivars growing in South Africa," *South African Journal of Plant and Soil*, 2011, doi: 10.1080/02571862.2011.10640008.

[72] K. Dhaouadi *et al.*, "Sucrose supplementation during traditional carob syrup processing affected its chemical characteristics and biological activities," *LWT - Food Science and Technology*, 2014, doi: 10.1016/J.LWT.2014.01.025.

[73] L. B. Roseiro *et al.*, "Supercritical, ultrasound and conventional extracts from carob (Ceratonia siliqua L.) biomass: Effect on the phenolic profile and antiproliferative activity," *Ind. Crops Prod.*, 2013, doi: 10.1016/J.INDCROP.2013.02.026.

[74] J. Xie *et al.*, "Effects of enzymatic hydrolysate of locust bean gum on digestibility, intestinal morphology and microflora of broilers," *J. Anim. Physiol. Anim. Nutr. (Berl).*, 2020, doi: 10.1111/jpn.13245.

[75] S. H. Abu Hafsa, S. A. Ibrahim, and A. A. Hassan, "Carob pods (Ceratonia siliqua L.) improve growth performance, antioxidant status and caecal characteristics in growing rabbits," *J. Anim. Physiol. Anim. Nutr. (Berl).*, 2017, doi: 10.1111/JPN.12651.

[76] S. Gruendel *et al.*, "Carob pulp preparation rich in insoluble dietary fibre and polyphenols increases plasma glucose and serum insulin responses in combination with a glucose load in humans," *Br. J. Nutr.*, 2007, doi: 10.1017/S0007114507701642.

[77] P. Libby, "Inflammatory mechanisms: the molecular basis of inflammation and disease," *Nutr. Rev.*, 2007, doi: 10.1111/J.1753-4887.2007.TB00352.X.

[78] H. Zhao *et al.*, "Inflammation and tumor progression: signaling pathways and targeted intervention," *Signal Transduct. Target. Ther.*, 2021, doi: 10.1038/S41392-021-00658-5.

[79] L. Â. M. Santiago *et al.*, "Flavonoids, alkaloids and saponins: are these plant-derived compounds an alternative to the treatment of rheumatoid arthritis? A literature review," *Clinical Phytoscience 2021 7:1*, 2021, doi: 10.1186/S40816-021-00291-3.

[80] A. O. Aremu and S. C. Pendota, "Medicinal Plants for Mitigating Pain and Inflammatory-Related Conditions: An Appraisal of Ethnobotanical Uses and Patterns in South Africa," *Front. Pharmacol.*, 2021, doi: 10.3389/FPHAR.2021.758583/BIBTEX.

[81] N. R. Perron and J. L. Brumaghim, "A review of the antioxidant mechanisms of polyphenol compounds related to iron binding," *Cell Biochem. Biophys.*, 2009, doi: 10.1007/S12013-009-9043-X/FIGURES/14.

[82] A. M. Mileo and S. Miccadei, "Polyphenols as Modulator of Oxidative Stress in Cancer Disease: New Therapeutic Strategies," *Oxid. Med. Cell. Longev.*, 2015, doi: 10.1155/2016/6475624.

[83] C. A. Edwards *et al.*, "Polyphenols and health: Interactions between fibre, plant polyphenols and the gut microbiota," *Nutr. Bull.*, 2017, doi: 10.1111/NBU.12296.

[84] B. J. Zhu, M. Z. Zayed, H. X. Zhu, J. Zhao, and S. P. Li, "Functional polysaccharides of carob fruit: a review," *Chin. Med.*, 2019, doi: 10.1186/S13020-019-0261-X.

[85] F. M. F. Roleira, C. L. Varela, S. C. Costa, and E. J. Tavares-da-Silva, *Phenolic Derivatives From Medicinal Herbs and Plant Extracts: Anticancer Effects and Synthetic Approaches to Modulate Biological Activity*, 2018 doi: 10.1016/B978-0-444-64057-4.00004-1.

[86] K. N. Rashed, "Medicinal Plants as a Safe Target for Treatment of Cancer," *Nat. Prod. Chem. Res.*, 2014, doi: 10.4172/2329-6836.1000E106.

[87] M. J. Balunas and A. D. Kinghorn, "Drug discovery from medicinal plants," *Life Sci.*, 2005, doi: 10.1016/j.lfs.2005.09.012.

[88] A. M. L. Seca and D. C. G. A. Pinto, "Plant Secondary Metabolites as Anticancer Agents: Successes in Clinical Trials and Therapeutic Application," *Int. J. Mol. Sci.*, 2018, doi: 10.3390/IJMS19010263.

[89] "Colorectal cancer yields to dual HER2 blockade," *Cancer Discov.*, 2016, doi: 10.1158/2159-8290.CD-NB2016-061.

[90] L. Yang *et al.*, "Depleting receptor tyrosine kinases EGFR and HER2 overcomes resistance to EGFR inhibitors in colorectal cancer," *Journal of Experimental and Clinical Cancer Research*, 2022, doi: 10.1186/s13046-022-02389-z.

[91] M. I. Ellina, P. Bouris, A. J. Aletras, A. D. Theocharis, D. Kletsas, and N. K. Karamanos, "EGFR and HER2 exert distinct roles on colon cancer cell functional properties and expression of matrix macromolecules," *Biochimica et Biophysica Acta (BBA) - General Subjects*, 2014, doi: 10.1016/J.BBAGEN.2014.04.019.

[92] S. A. Djaballah, F. Daniel, A. Milani, G. Ricagno, and S. Lonardi, "HER2 in Colorectal Cancer: The Long and Winding Road From Negative Predictive Factor to Positive Actionable Target," *American Society of Clinical Oncology Educational Book*, 2022, doi: 10.1200/EDBK_351354.

[93] M. Ivanova *et al.*, "HER2 in Metastatic Colorectal Cancer: Pathology, Somatic Alterations, and Perspectives for Novel Therapeutic Schemes," *Life (Basel)*, 2022, doi: 10.3390/LIFE12091403/S1.

[94] T. Fraga *et al.*, "HER2 Status in RAS and BRAF Wild-Type Metastatic Colorectal Cancer: A Portuguese Study," *Cureus*, 2023, doi: 10.7759/CUREUS.42536.

[95] J. Vázquez, M. López, E. Gibert, E. Herrero, and F. Javier Luque, "Merging Ligand-Based and Structure-Based Methods in Drug Discovery: An Overview of Combined Virtual Screening Approaches," *Molecules*, 2020, doi: 10.3390/MOLECULES25204723.

[96] J. Li, H. Wang, J. Li, J. Bao, and C. Wu, "Discovery of a Potential HER2 Inhibitor from Natural Products for the Treatment of HER2-Positive Breast Cancer," *International Journal of Molecular Sciences,* 2016, doi: 10.3390/IJMS17071055.

[97] Y. L. Tang, D. D. Li, J. Y. Duan, L. M. Sheng, and X. Wang, "Resistance to targeted therapy in metastatic colorectal cancer: Current status and new developments," *World J. Gastroenterol.*, 2023, doi: 10.3748/WJG.V29.I6.926.

[98] N. Iqbal and N. Iqbal, “Human Epidermal Growth Factor Receptor 2 (HER2) in Cancers: Overexpression and Therapeutic Implications,” *Mol. Biol. Int.*, 2014, doi: 10.1155/2014/852748.

[99] Benli Y, Arıkan H, Akbulut-Çalışkan Ö. HER2-targeted therapy in colorectal cancer: a comprehensive review. Clin Transl Oncol. 2025 doi: 10.1007/s12094-025-03887-0.

[100] P. Laurent-Puig *et al.*, “ERBB2 alterations a new prognostic biomarker in stage III colon cancer from a FOLFOX based adjuvant trial (PETACC8),” *Annals of Oncology*, 2016, doi: 10.1093/ANNONC/MDW370.08.

[101] F. Bray *et al.*, “Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries,” *CA Cancer J. Clin.*, 2024, doi: 10.3322/CAAC.21834.

[102] T. Hashimoto *et al.*, “A comprehensive appraisal of HER2 heterogeneity in HER2-amplified and HER2-low colorectal cancer,” *Br. J. Cancer*, 2023, doi: 10.1038/S41416-023-02382-Z.

[103] C. Parseghian, M. Eluri, S. Kopetz, and K. Raghav, “Mechanisms of resistance to EGFR-targeted therapies in colorectal cancer: more than just genetics,” *Front. Cell Dev. Biol.*, 2023, doi: 10.3389/FCELL.2023.1176657.

[104] T. I. Oprea and H. Matter, “Integrating virtual screening in lead discovery,” *Curr. Opin. Chem. Biol.*, 2004, doi: 10.1016/j.cbpa.2004.06.008.

[105] C. G. Wermuth, B. Villoutreix, S. Grisoni, A. Olivier, and J. P. Rocher, “Strategies in the Search for New Lead Compounds or Original Working Hypotheses,” *The Practice of Medicinal Chemistry*, 2015, doi: 10.1016/B978-0-12-417205-0.00004-3.

[106] L. G. Ferreira, R. N. Dos Santos, G. Oliva, and A. D. Andricopulo, “Molecular docking and structure-based drug design strategies,” *Molecules*, 2015, doi: 10.3390/MOLECULES200713384.

[107] X.-Y. Meng, H.-X. Zhang, M. Mezei, and M. Cui, “Molecular Docking: A Powerful Approach for Structure-Based Drug Discovery,” *Current Computer Aided-Drug Design*, 2012, doi: 10.2174/157340911795677602.

[108] Y. C. Lo, S. E. Rensi, W. Torng, and R. B. Altman, “Machine learning in chemoinformatics and drug discovery,” *Drug Discov. Today*, 2018, doi: 10.1016/J.DRUDIS.2018.05.010.

[109] S. Badillo *et al.*, “An Introduction to Machine Learning,” *Clin. Pharmacol. Ther.*, 2020, doi: 10.1002/cpt.1796.

[110] M. Shehab *et al.*, “Machine learning in medical applications: A review of state-of-the-art methods,” *Comput. Biol. Med.*, 2022, doi: 10.1016/J.COMPBIOMED.2022.105458.

[111] J. Vamathevan *et al.*, “Applications of machine learning in drug discovery and development,” *Nat. Rev. Drug Discov.*, vol. 18, no. 6, pp. 463–477, 2019, doi: 10.1038/s41573-019-0024-5.

[112] M. Mohammed, M. B. Khan, and E. B. M. Bashie, “Machine learning: Algorithms and applications,” *Machine Learning: Algorithms and Applications*, 2016, doi: 10.1201/9781315371658

[113] Y. Reich and S. V. Barai, “Evaluating machine learning models for engineering problems,” *Artificial Intelligence in Engineering*, 1999, doi: 10.1016/S0954-1810(98)00021-1.

[114] J. Tolles and W. J. Meurer, “Logistic regression: Relating patient characteristics to outcomes,” *Journal of the American Medical Association*, 2016, doi: 10.1001/jama.2016.7653.

[115] L. T. Lindsey, B. R. Olin, and R. A. Hansen, “Systematic review and meta-analysis," in: Rajender R. Aparasu and John P. Bentley (Eds.), *Principles of Research Design and Drug Literature Evaluation*, *2e,* McGraw-Hill Education, 2020, URL: https://accesspharmacy.mhmedical.com/book.aspx?bookid=2733. (Accessed, May 05, 2024).

[116] R. B. Sesay, M. Kpangay, and S. Seppeh, “An Ordinal Logistic Regression Model to Identify Factors Influencing Students Academic Performance at Njala University,” *International Journal of Research and Scientific Innovation*, 2021, doi: 10.51244/IJRSI.2021.8104.

[117] S. Uddin, I. Haque, H. Lu, M. A. Moni, and E. Gide, "Comparative performance analysis of K-nearest neighbour (KNN) algorithm and its different variants for disease prediction," *Scientific Reports,* 2022, doi: 10.1038/s41598-022-10358-x.

[118] G. Guo, D. Neagu, and M. T. D. Cronin, "Using kNN model for automatic feature selection," *Lecture Notes in Computer Science*, 2005, doi: 10.1007/11551188_44.

[119] Y. Lu, Y. Zhang, F. Richter, and T. Seidl, "K-Nearest Neighbor based Clustering with Shape Alternation Adaptivity," *Proceedings of the International Joint Conference on Neural Networks*, 2020, doi: 10.1109/IJCNN48605.2020.9207321.

[120] C. C. Chang and C. J. Lin, "LIBSVM: A Library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, 2011, doi: 10.1145/1961189.1961199.

[121] T. Uda, M. Serizawa, T. Kumada, and K. Sakai, "A new model for predicting three-dimensional beach changes by expanding Hsu and Evans' equation," *Coastal Engineering*, 2010, doi: 10.1016/J.COASTALENG.2009.10.006.

[122] W. S. Noble, "A biologist's introduction to support vector machines," November 1, 2006. Department of Genome Sciences, Department of Computer Science and Engineering, University of Washington, Seattle, WA, USA. URL: https://noble.gs.washington.edu/papers/noble2006biologists.pdf, May 05, 2024

[123] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis, "Machine learning applications in cancer prognosis and prediction," *Comput. Struct. Biotechnol. J.*, 2015, doi: 10.1016/J.CSBJ.2014.11.005.

[124] G. Seddon *et al.*, "Drug design for ever, from hype to hope," *J. Comput. Aided. Mol. Des.*, 2012, doi: 10.1007/S10822-011-9519-9.

[125] D. A. Pisner and D. M. Schnyer, "Support vector machine," *Machine Learning: Methods and Applications to Brain Disorders*, 2019, doi: 10.1016/B978-0-12-815739-8.00006-7.

[126] D. C. Toledo-Pérez, J. Rodríguez-Reséndiz, R. A. Gómez-Loenzo, and J. C. Jauregui-Correa, "Support Vector Machine-based EMG signal classification techniques: A review," *Applied Sciences (Switzerland)*, 2019, doi: 10.3390/APP9204402.

[127] G. R. L. Kodikara and T. Woldai, “Spectral indices derived, non-parametric Decision Tree Classification approach to lithological mapping in the Lake Magadi area, Kenya,” *Int. J. Digit. Earth*, 2018, doi: 10.1080/17538947.2017.1372525.

[128] E. Afsaneh, A. Sharifdini, H. Ghazzaghi, and M. Z. Ghobadi, “Recent applications of machine learning and deep learning models in the prediction, diagnosis, and management of diabetes: a comprehensive review,” *Diabetol. Metab. Syndr.*, 2022, doi: 10.1186/S13098-022-00969-9.

[129] I. H. Sarker, “Machine Learning: Algorithms, Real-World Applications and Research Directions,” *SN Comput. Sci.*, 2021, doi: 10.1007/S42979-021-00592-X/FIGURES/11.

[130] G. Obaido *et al.*, “Supervised machine learning in drug discovery and development: Algorithms, applications, challenges, and prospects,” *Machine Learning with Applications*, 2024, doi: 10.1016/J.MLWA.2024.100576.

[131] B.-G. Kerstin and G. J. M., “View of Determining the value of drug development candidates and technology platforms,” The journal of commencial biotechnology. Accessed: Apr. 26, 2025. [Online]. Available: https://commercialbiotechnology.com/menuscript/index.php/jcb/article/view/113/112

[132] L. Breiman, “Random forests,” *Mach. Learn.*, 2001, doi: 10.1023/A:1010933404324/METRICS.

[133] H. Chen, O. Engkvist, Y. Wang, M. Olivecrona, and T. Blaschke, “The rise of deep learning in drug discovery,” *Drug Discov. Today*, 2018, doi: 10.1016/J.DRUDIS.2018.01.039.

[134] D. A. Putri, D. A. Kristiyanti, E. Indrayuni, A. Nurhadi, and D. R. Hadinata, “Comparison of Naive Bayes Algorithm and Support Vector Machine using PSO Feature Selection for Sentiment Analysis on E-Wallet Review,” *J. Phys. Conf. Ser.*, 2020, doi: 10.1088/1742-6596/1641/1/012085.

[135] A. Natekin and A. Knoll, “Gradient boosting machines, a tutorial,” *Front. Neurorobot.*, 2013, doi: 10.3389/FNBOT.2013.00021.

[136] Y. CAO, Q.-G. MIAO, J.-C. LIU, and L. GAO, “Advance and Prospects of AdaBoost Algorithm,” *Acta Automatica Sinica*, 2013, doi: 10.1016/S1874-1029(13)60052-X.

[137] Y. Freund and R. E. Schapire, "A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting," *J. Comput. Syst. Sci.*, 1997, doi: 10.1006/JCSS.1997.1504.

[138] N. Aniceto, A. A. Freitas, A. Bender, and T. Ghafourian, "A novel applicability domain technique for mapping predictive reliability across the chemical space of a QSAR: reliability-density neighbourhood," *J. Cheminform.*, 2016, doi: 10.1186/S13321-016-0182-Y.

[139] L. T. Afolabi, F. Saeed, H. Hashim, and O. O. Petinrin, "Ensemble learning method for the prediction of new bioactive molecules," *PLoS One*, 2018, doi: 10.1371/journal.pone.0189538.

[140] I. D. Mienye and Y. Sun, "A Survey of Ensemble Learning: Concepts, Algorithms, Applications, and Prospects," *IEEE Access*, 2022, doi: 10.1109/ACCESS.2022.3207287.

[141] B. Pavlyshenko, "Using Stacking Approaches for Machine Learning Models," *Proceedings of the 2018 IEEE 2nd International Conference on Data Stream Mining and Processing, DSMP,* 2018, doi: 10.1109/DSMP.2018.8478522.

[142] L. Breiman, "Stacked regressions," *Mach. Learn.*, 1996, doi: 10.1007/BF00117832/METRICS.

[143] A. Lavecchia, "Machine-learning approaches in drug discovery: Methods and applications," *Drug Discov. Today*, 2015, doi: 10.1016/j.drudis.2014.10.012.

[144] N. Schaduangrat, N. Anuwongcharoen, M. A. Moni, P. Lio', P. Charoenkwan, and W. Shoombuatong, "StackPR is a new computational approach for large-scale identification of progesterone receptor antagonists using the stacking strategy," *Scientific Reports*, 2022, doi: 10.1038/s41598-022-20143-5.

[145] D. B. N. T. Oku, D. D. Babatunde, Y. Nuapia, G. K. More, and R. C. Chokwe, "Harnessing Machine Learning for the Virtual Screening of Natural Compounds as Both EGFR and HER2 Inhibitors in Colorectal Cancer: A Novel Therapeutic Approach," *ACS Omega*, doi: 10.1021/acsomega.5c07683

[146] G. Landrum, "RDKit : A software suite for cheminformatics, computational chemistry, and predictive modeling," URL: https://www.rdkit.org/RDKit_Overview.pdf (Accessed, May 05, 2024).

[147] C. C. Melo-Filho *et al.*, "Discovery of new potent hits against intracellular Trypanosoma cruzi by QSAR-based virtual screening," *Eur. J. Med. Chem.*, 2019, doi: 10.1016/J.EJMECH.2018.11.062.

[148] D. H. Smith, R. E. Carhart, and R. Venkataraghavan, "Atom Pairs as Molecular Features in Structure-Activity Studies: Definition and Applications," *J. Chem. Inf. Comput. Sci.*, 1985, doi: 10.1021/ci00046a002.

[149] C. Steinbeck, Y. Han, S. Kuhn, O. Horlacher, E. Luttmann, and E. Willighagen, "The Chemistry Development Kit (CDK): An open-source Java library for chemo- and bioinformatics," *J. Chem. Inf. Comput. Sci.*, 2003, doi: 10.1021/CI025584Y/ASSET/IMAGES/LARGE/CI025584YF00005.JPEG.

[150] E. L. Willighagen *et al.*, "The Chemistry Development Kit (CDK) v2.0: atom typing, depiction, molecular formulas, and substructure searching," *J. Cheminform.*, 2017, doi: 10.1186/S13321-017-0220-4/FIGURES/8.

[151] L. H. Hall and L. B. Kier, "Electrotopological State Indices for Atom Types: A Novel Combination of Electronic, Topological, and Valence State Information," *J. Chem. Inf. Comput. Sci.*, 1995, doi: 10.1021/CI00028A014/ASSET/CI00028A014.FP.PNG_V03.

[152] J. Klekota and F. P. Roth, "Chemical substructures that enrich for biological activity," *Bioinformatics*, 2008, doi: 10.1093/BIOINFORMATICS/BTN479.

[153] S. Kim *et al.*, "PubChem Substance and Compound databases," *Nucleic Acids Res.*, 2016, doi: 10.1093/NAR/GKV951.

[154] J. L. Durant, B. A. Leland, D. R. Henry, and J. G. Nourse, "Reoptimization of MDL Keys for Use in Drug Discovery," *J. Chem. Inf. Comput. Sci.*, 2002, doi: 10.1021/CI010132R.

[155] K. A. Verba and N. Jura, "Structures of the HER2 – HER3 – NRG1β complex reveal a dynamic dimer interface," *Nature*, 2021, doi: 10.1038/s41586-021-04084-z.

[156] T. Grabe *et al.*, "Addressing the Osimertinib Resistance Mutation EGFR-L858R/C797S with Reversible Aminopyrimidines," *ACS Med Chem Lett.*, 2023, doi: 10.1021/acsmedchemlett.2c00514.

[157] Protein Data Bank, "RCSB PDB: Homepage," 2024.

[158] G. K. More and R. T. Makola, "In-vitro analysis of free radical scavenging activities and suppression of LPS-induced ROS production in macrophage cells by Solanum sisymbriifolium extracts," *Sci. Rep.*, 2020, doi: 10.1038/s41598-020-63491-w.

[159] J. van Meerloo, G. J. L. Kaspers, and J. Cloos, "Cell sensitivity assays: the MTT assay," *Methods Mol. Biol.*, 2011, doi: 10.1007/978-1-61779-080-5_20.

[160] L. R. Snyder and J. J. Kirkland, "Introduction to modern liquid chromatography." *John Wiley & Sons, Inc.*, 2009. doi:10.1002/9780470508183.

[161] I. Ignat, I. Volf, and V. I. Popa, "A critical review of methods for characterisation of polyphenolic compounds in fruits and vegetables," *Food Chem.*, 2011, doi: 10.1016/J.FOODCHEM.2010.12.026.

[162] E. de Rijke, P. Out, W. M. A. Niessen, F. Ariese, C. Gooijer, and U. A. T. Brinkman, "Analytical separation and detection methods for flavonoids," *J. Chromatogr. A*, 2006, doi: 10.1016/J.CHROMA.2006.01.019.

[163] I. Kayel *et al.*, "Three Mediterranean species from natural plant communities (Ceratonia siliqua, Pistacia lentiscus, and Olea europaea var. sylvestris): phenolic acids, flavonoids, and biological activities," *South African Journal of Botany*, 2024, doi: 10.1016/J.SAJB.2024.10.056.

[164] F. Sahnouni and F. Lehfa, "Phytochemical study and biological activities of Ceratonia siliqua L.," *Phytothérapie*, 2023, doi: 10.3166/PHYTO-2022-0350.

[165] M. H. Shahrajabian and W. Sun, "Carob (Ceratonia siliqua L.), Pharmacological and Phytochemical Activities of Neglected Legume of the Mediterranean Basin, as Functional Food," *Rev. Recent Clin. Trials*, 2024, doi: 10.2174/0115748871278128240109074506.

[166] T. J. Ritchie, S. J. F. MacDonald, R. J. Young, and S. D. Pickett, "The impact of aromatic ring count on compound developability: further insights by examining carbo- and hetero-aromatic and -aliphatic ring types," *Drug Discov. Today*, 2011, doi: 10.1016/J.DRUDIS.2010.11.014.

[167] N. Schaduangrat, N. Anuwongcharoen, P. Charoenkwan, and W. Shoombuatong, "DeepAR: a novel deep learning-based hybrid framework for the interpretable prediction of androgen receptor antagonists," *J. Cheminform.*, 2023, doi: 10.1186/s13321-023-00721-z.

[168] Y. Feng *et al.*, "EGF signalling pathway regulates colon cancer stem cell proliferation and apoptosis," *Cell Prolif.*, 2012, doi: 10.1111/J.1365-2184.2012.00837.X.

[169] R. L. Carpenter and H.-W. Lo, "Regulation of Apoptosis by HER2 in Breast Cancer," *J. Carcinog. Mutagen.*, 2013, doi: 10.4172/2157-2518.S7-003.

[170] K. Muniandy, S. Gothai, K. M. H. Badran, S. S. Kumar, N. M. Esa, and P. Arulselvan, "Suppression of Proinflammatory Cytokines and Mediators in LPS-Induced RAW 264.7 Macrophages by Stem Extract of Alternanthera sessilis via the Inhibition of the NF- κ B Pathway," *J. Immunol. Res.*, 2018, doi: 10.1155/2018/3430684.

[171] P. C. Agu *et al.*, "Molecular docking as a tool for the discovery of molecular targets of nutraceuticals in diseases management," *Sci. Rep.*, 2023, doi: 10.1038/S41598-023-40160-2.

[172] J. Son *et al.*, "A Novel HER2-Selective Kinase Inhibitor Is Effective in HER2 Mutant and Amplified Non–Small Cell Lung Cancer," *Cancer Res.*, 2022, doi: 10.1158/0008-5472.CAN-21-2693.

[173] Y. Yarden and M. X. Sliwkowski, "Untangling the ErbB signalling network," *Nature Reviews Molecular Cell Biology,* 2001, doi: 10.1038/35052073.

[174] T. Grabe *et al.*, "Addressing the Osimertinib Resistance Mutation EGFR-L858R/C797S with Reversible Aminopyrimidines," *ACS Med. Chem. Lett.*, 2023, doi: 10.1021/ACSMEDCHEMLETT.2C00514/SUPPL_FILE/ML2C00514_SI_001.PDF.

[175] K. S. Thress *et al.*, “Acquired EGFR C797S mutation mediates resistance to AZD9291 in non-small cell lung cancer harboring EGFR T790M,” *Nat. Med.*, 2015, doi: 10.1038/NM.3854.

[176] A. T. Borchers, “Natural Compounds in Cancer Therapy—Promising Nontoxic Antitumor Agents from Plants & Other Natural Sources: by John Boik, 2001, 521 pages, softcover, $32. Oregon Medical Press, LLC, Princeton, MN,” *Am. J. Clin. Nutr.*, 2002, doi: 10.1093/AJCN/75.5.955A.

[177] A. Ben Hsouna, M. Trigui, R. Ben Mansour, R. M. Jarraya, M. Damak, and S. Jaoua, “Chemical composition, cytotoxicity effect and antimicrobial activity of Ceratonia siliqua essential oil with preservative effects against Listeria inoculated in minced beef meat,” *Int. J. Food Microbiol.*, 2011, doi: 10.1016/J.IJFOODMICRO.2011.04.028.

[178] G. Gregoriou *et al.*, “Anti-Cancer Activity and Phenolic Content of Extracts Derived from Cypriot Carob (Ceratonia siliqua L.) Pods Using Different Solvents,” *Molecules,* 2021, doi: 10.3390/MOLECULES26165017.

APPENDICES

PREAMBLE

The appendices provide supplementary figures and tables that support and extend the results presented in the main text. A link to the GitHub repository is also included, which contains the Python codes, curated dataset, and final prediction data used in this study. Information regarding software availability is provided in Chapter 4.

APPENDIX A (Machine Learning)

APPENDIX A 1: Data and software availability

The Python codes, curated dataset, final predictions, and molecular structures used in this study are available at the following public repository:

https://github.com/dbntoku/JCIM_ML_EGFR_HER2_.git

The dataset is provided in CSV format, and the molecular structures are represented as Canonical SMILES. Additionally, the associated LOTUS ID, fingerprints, predictions, and probabilities for each compound from the LOTUS Database are included. The data are available in a machine-readable format and can be used for further analysis or model training. If necessary, scripts for data extraction and preprocessing, along with machine learning models used in this study, are also included in the repository.

APPENDIX A 2: Details on the sampling techniques, evaluation results, and comparative model performance are provided below. Stacking Sampling Results:

Sampling	Phase	Accuracy	Precision	F1-Score	ROC AUC	MCC	Sensitivity	Specificity
no_sampling	Train	1.000	1.000	1.000	1.000	1.000	1.000	1.000
no_sampling	Test	0.667	0.000	0.000	0.815	0.000	0.000	1.000
undersampling	Train	0.850	0.706	0.828	0.995	0.735	1.000	0.766
undersampling	Test	0.700	0.529	0.667	0.780	0.476	0.900	0.600
Smote	Train	1.000	1.000	1.000	1.000	1.0000	1.0000	1.000
Smote	Test	0.800	0.700	0.700	0.865	0.5500	0.7000	0.850
smote_undersampling	Train	1.0000	1.0000	1.000	1.000	1.0000	1.0000	1.000
smote_undersampling	Test	0.833	0.778	0.737	0.860	0.617	0.700	0.900

APPENDIX A 3: (A) Scatterplot of predicted probability of activity for intermediate compounds (IC_{50}: 1–10 μM), with compounds predicted as active (green) or inactive (red). (B) Histogram showing the distribution of predicted probabilities for intermediate compounds. The vertical red dashed line indicates the 0.5 classification threshold.