ΠΑΝΕΠΙΣΤΗΜΙΟ ΔΥΤΙΚΗΣ ΑΤΤΙΚΗΣ

ΣΧΟΛΗ ΜΗΧΑΝΙΚΩΝ

ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΒΙΟΪΑΤΡΙΚΗΣ

Usage of Machine Learning and Computational Chemistry methodologies to predict the activity of potential antiviral compounds against the Zika virus protease

IOANNA ARBOUNIOTI
Registration number: 19388012

Supervising Professor
Dionisis Cavouras, Emeritus Professor

Athens 10/10/2024

Usage of Machine Learning and Computational chemistry methodologies to predict the activity of potential antiviral compounds against the Zika virus protease

Η Τριμελής Εξεταστική Επιτροπή

Επιβλέπων Καθηγητής

Διονύσιος Κάβουρας	Μίνως Μαστούκας	Ευτυχία Κρίτση
Καθηγητής Ομότιμος	Επίκουρος Καθηγητής	Επίκουρος Καθηγήτρια
Digitally signed by Dionysios Kavouras Date: 2024.10.11 18:55:19 +03'00'	MINOS TIMOTHEOS MATSOUKAS Digitally signed by MINOS TIMOTHEOS MATSOUKAS Date: 2024.10.13 18:33:05 +03'00'	Eftychia Kritsi Digitally signed by Eftychia Kritsi Date: 2024.10.11 19:53:55 +03'00'

ΔΗΛΩΣΗ ΣΥΓΓΡΑΦΕΑ ΔΙΠΛΩΜΑΤΙΚΗΣ ΕΡΓΑΣΙΑΣ

Η υπογράφουσα Ιωάννα Αρμπουνιώτη του Ιωάννη, με αριθμό μητρώου 19388012 φοιτήτρια του Τμήματος Μηχανικών Βιοϊατρικής της Σχολής Μηχανικών του Πανεπιστημίου Δυτικής Αττικής, δηλώνω υπεύθυνα ότι:

«Είμαι συγγραφέας αυτής της διπλωματικής εργασίας και κάθε βοήθεια την οποία είχα για την προετοιμασία της είναι πλήρως αναγνωρισμένη και αναφέρεται στην εργασία. Επίσης, οι όποιες πηγές από τις οποίες έκανα χρήση δεδομένων, ιδεών ή λέξεων, είτε ακριβώς είτε παραφρασμένες, αναφέρονται στο σύνολό τους, με πλήρη αναφορά στους συγγραφείς, τον εκδοτικό οίκο ή το περιοδικό, συμπεριλαμβανομένων και των πηγών που ενδεχομένως χρησιμοποιήθηκαν από το διαδίκτυο. Επίσης, βεβαιώνω ότι αυτή η εργασία έχει συγγραφεί από μένα αποκλειστικά και αποτελεί προϊόν πνευματικής ιδιοκτησίας τόσο δικής μου, όσο και του Ιδρύματος.

Παράβαση της ανωτέρω ακαδημαϊκής μου ευθύνης αποτελεί ουσιώδη λόγο για την ανάκληση του διπλώματός μου».

Ημερομηνία

10/10/2024

Η Δηλούσα

ΠΕΡΙΛΗΨΗ

Ο ιός Ζίκα είναι ένας αρβοϊός, μεταδίδεται μέσω αιματοφάγων αρθροπόδων και μπορεί να προκαλέσει συμπτώματα που κυμαίνονται από πυρετό και αίσθημα κόπωσης έως σοβαρότερες νευρολογικές και ανοσολογικές επιπλοκές. Παρά την εκτεταμένη έρευνα, δεν υπάρχουν εγκεκριμένα εμβόλια ή αντιϊικά φάρμακα από διεθνείς οργανισμούς υγείας για αυτή τη λοίμωξη. Το γονιδίωμά του ιού είναι ένα μονόκλωνο ριβοζονουκλεϊνικό οξύ. Κωδικοποιεί μία πολυπρωτεΐνη, η οποία διασπάται από πρωτεάσες, με το σύμπλεγμα NS2B-NS3 να είναι το πιο κρίσιμο για τον πολλαπλασιασμό του ιού. Στόχος της παρούσας μελέτης είναι η πρόβλεψη πιθανών αντιϊικών ενώσεων που αναστέλλουν αυτή την πρωτεάση, χρησιμοποιώντας μια πολύπλευρη προσέγγιση που περιλαμβάνει στατιστική ανάλυση, μηχανική μάθηση και τεχνικές υπολογιστικής χημείας. Μια βάση δεδομένων με ενώσεις που προήλθαν από το ChEMBL αναλύθηκε για την αναγνώριση των σημαντικότερων χαρακτηριστικών. Το τεστ κατάταξης Wilcoxon αποκάλυψε ότι οκτώ από αυτά τα χαρακτηριστικά παρουσίασαν στατιστικά σημαντικές διαφορές. Στη συνέχεια, ένα μοντέλο μηχανικής μάθησης, που αναπτύχθηκε χρησιμοποιώντας τη μέθοδο εξαντλητικής αναζήτησης με ταξινομητή το Random Forest, εντόπισε τον βέλτιστο συνδυασμό επτά χαρακτηριστικών, επιτυγχάνοντας ακρίβεια 95,46%. Στην υπολογιστική χημεία, επιλέχθηκε η κατάλληλη κρυσταλλική δομή της πρωτεάσης του ιού, καθώς και οι ενώσεις που θα δοκιμαστούν για την ανασταλτική τους δράση. Στη συνέχεια, πραγματοποιήθηκαν πειράματα μοριακής πρόσδεσης χρησιμοποιώντας το Webina και το Maestro. Στο τέλος, πέντε ενώσεις αναδείχθηκαν ως πιθανοί προσδέτες και όλες ταξινομήθηκαν ως ενεργές με πιθανότητα 70%. Αυτά τα ευρήματα αναδεικνύουν την αποτελεσματική ενσωμάτωση αυτών των προσεγγίσεων στον εντοπισμό ενώσεων που θα μπορούσαν δυνητικά να αναστείλουν την πρωτεάση του ιού Ζίκα, παρέχοντας πολύτιμες πληροφορίες για μελλοντική πειραματική επικύρωση.

ABSTRACT

Zika virus is an arbovirus, it is transmitted through blood-feeding arthropods and can cause symptoms ranging from fever and malaise to more severe neurological and immunological complications. Despite the extensive research, there are no approved vaccines or antiviral drugs from health organizations for this infection. Its genome is a single-stranded ribonucleic acid. It encodes a polyprotein that is cleaved by proteases, with the NS2B-NS3 complex being the most crucial for viral replication. The aim of this study is to predict potential antiviral compounds that inhibit this protease using an integrative approach involving statistical analysis, machine learning and computational chemistry techniques. A compound database from ChEMBL was analyzed to identify the most significant features. The Wilcoxon rank-sum test revealed that eight of these features showed statistically significant differences. Subsequently, a machine learning model, developed using the exhaustive search method with a Random Forest classifier, identified the optimal combination of seven features, achieving an accuracy of 95.46%. In computational chemistry, initially, the appropriate crystal structure of the virus protease complex was selected, along with the compounds to be tested for their inhibitory potential. Molecular docking experiments were then conducted using Webina and Maestro. Five compounds in the end emerged as promising candidates, and all were classified as active with a probability of 70%. These findings highlight the effective integration of these approaches in identifying compounds that could potentially inhibit the Zika virus protease, providing valuable insights for future experimental validation.

Keywords: *zika virus; machine learning; compounds; statistical analysis; molecular docking;*

ACKNOWLEDGEMENTS:

I am immensely grateful to Mr. Dionisis Cavouras, Mr. Minos Matsoukas, Mr. Vasilis Panagiotopoulos, Mrs. Eftichia Kritsi and Mr. Paraschos Christodoulou for their help and support.

TABLE OF CONTENTS

INTRODUCTION

The Zika virus outbreak, in recent years, has underscored the urgent need for effective antiviral treatments. Despite, extensive research efforts, creating targeted therapies against this virus has proven challenging. The aim is to make a contribution to this effort by employing various approaches that integrates statistical analysis, machine learning and molecular docking techniques, to predict potential antiviral compounds targeting to inhibit the activity of the Zika virus protease, a pivotal enzyme in viral replication.

There are benefits and challenges associated with using these computational methods. They expedite the screening process by rapidly screening many compounds, reducing the time and cost required for traditional experimental approaches. Furthermore, these methods can predict the potential effectiveness of compounds with high accuracy, guiding researchers towards the most promising candidates for further experimentation. Once established, computational models can be readily applied to screen additional chemical libraries.

Difficulties arise with data availability and quality, because incomplete or biased datasets can lead to inaccurate predictions and flawed conclusions. Additionally, interpreting the results requires expertise in many scientific fields and ensuring model generalization to unseen data is crucial for reliable predictions in real world applications. It is worth mentioning, that although there are advancements in Zika virus research, gaps remain in our understanding of his transmission dynamics and long-term health effects, hindering the development of targeted therapies.

This thesis provides an overview of viruses, including their structure, function and classification, focusing on flaviviruses and more specifically on the Zika virus. The mechanisms of the immune system and how it responds to viral infections was, also, described. The aim was to identify compounds that would potentially inhibit the NS2B-NS3 protease complex. To achieve that, a database was created with a selection of compounds from ChEMBL that had been previously experimentally tested for their activity. Their features were extracted using the RDKit software and only the possibly most important ones, were kept for further analysis.

Statistical methods were used, and machine learning models were developed to extract the features that provided the best discriminative ability. Molecular docking experiments were then carried out in the freely accessible Webina software and in Maestro using two separate modes, GLIDE-SP and GLIDE-XP, to find potential ligands. Moreover, the compounds that were selected from the molecular docking experiments were extracted and their features were found using RDKit. From all the features produced the focus was on the most important ones that derived from the machine learning methodology. These compounds with the selected features were put into a model that was created in Metaboanalyst to determine if it would classify them as active or non-active. Finally, the results were evaluated.

TOPIC SELECTION & LITERATURE REVIEW

The subject of this thesis was chosen due to its critical relevance and innovative approach to addressing a significant public health threat. The Zika virus (ZIKV), known for causing severe congenital disabilities and neurological disorders, necessitates the urgent development of effective antiviral treatments. This research resides at the intersection of bioinformatics, cheminformatics and pharmacology, integrating computational techniques to enhance drug discovery. By leveraging machine learning to analyze vast datasets and predict compound activity, alongside molecular docking to elucidate drug-protein interactions, this study aims to streamline and optimize the identification of promising antiviral compounds.

Previous research on this topic has been conducted by a diverse group of scientists from various disciplines, including computational biologists, medicinal chemists, and pharmacologists. Notable contributions have come from researchers who specialize in machine learning applications in bioinformatics, as well as those focusing on molecular modeling and simulation techniques.

A. Jainul Fathima et al. (2018) used computational tools like Schrodinger’s Maestro and AutoDock, to study the interactions between Zika virus NS2B-NS3 protease (PDB: 5LC0) and various small molecule derivatives, including Isatin, Benzimidazole, Quinazoline, Indophenazine, and Indenoquinoxaline. The results showed that these derivatives exhibited favorable drug-like properties and strong binding interactions with the protease's active site, particularly for Benzimidazole derivative MBZ-SN and Isatin derivative SPIII-5CL-AC, indicating their potential as lead compounds for anti-ZIKV drug development [1].

A study by Dario Akaberi et al. (2020) utilized molecular docking techniques with AutoDock Vina to screen a library of Human Immunodeficiency Virus (HIV) protease inhibitors for their potential activity against the ZIKV NS2B-NS3 protease. Their molecular dynamics simulations, that were performed using the ZIKV protease crystal structure 5LC0, revealed that compound 9b, a C2-symmetric diol-based HIV protease inhibitor, exhibited notable binding stability and was further confirmed as an effective inhibitor in subsequent in vitro assays [2].

Additionally, Woon Yi Law et al. (2023) assessed the antiviral potential of Schiff base vanillin derivatives against Zika virus NS2B-NS3 protease, using ligand-based pharmacophore modeling with novobiocin, sofosbuvir, and azithromycin as training sets, followed by molecular docking simulations referencing the 5LC0 crystal structure. Several derivatives were identified with strong pharmacophore fit values, binding affinities, and key interactions within the protease's active site, suggesting their potential as promising ZIKV antiviral drugs [3].

Hisham N. Altayb and Hanan Ali Alatawi (2024) developed a machine learning based Quantitative Structure-Activity Relationship (QSAR) model to screen a library of 2.864 natural compounds for their potential to inhibit the Zika virus NS3 protease. The QSAR model used various molecular descriptors (features) to capture key physicochemical, topological, and electronic properties of the compounds. These descriptors were the input of the machine learning algorithms so they can predict the antiviral activity of the compounds.

The best-performing QSAR model was then used for virtual screening, followed by molecular docking studies, to evaluate the binding affinities of the prioritized compounds to the NS3 protease. Molecular dynamics simulations were conducted to assess the stability of the protein-ligand complexes over time, and binding free energy calculations were performed to identify the most promising inhibitors. Through this comprehensive approach, compound Streptomycin emerged as the top candidate showing strong and stable binding interactions with the NS3 protease, suggesting its potential as a lead compound for further development as a Zika virus inhibitor [4].

These studies underscore the effectiveness of computational methods in identifying novel antiviral compounds, highlighting the pivotal role of interdisciplinary research in advancing antiviral drug discovery efforts. The collective outcomes from these studies provide a robust foundation for the present research endeavor, aiming to further enhance the efficacy and efficiency of antiviral drug discovery against Zika and other related viruses.

This thesis addresses the need for innovative antiviral strategies and explores the efficacy of computational methods in predicting biological activity, thus advancing the field of antiviral drug discovery. There are various perspectives on the use of computational methods in drug discovery. Some researchers are highly optimistic about the potential of machine learning and molecular docking to revolutionize the field, citing successes in predicting drug efficacy and identifying new therapeutic targets. Others are more cautious, pointing out limitations such as the availability and quality of the data, the complexity of biological systems, and the need for experimental validation of computational predictions.

There is also ongoing debate about the best practices for integrating different computational techniques and the ethical considerations related to data privacy and the use of artificial intelligence in biomedical research. Overall, while the field is generally viewed as promising and rapidly evolving, it also faces challenges that require continued research and collaboration.

THEORETICAL BACKGROUND

2.1 Viral pandemics

The complex relationship between humans and nature has been the subject of intense interest and study recently, as it profoundly impacts human health, particularly concerning viral infections.

As humans progressively encroach upon natural habitats and disrupt ecosystems through deforestation and urbanization, natural barriers that once separated humans from wildlife are dismantled, increasing the risk of contact with new viruses originating from natural hosts. Activities such as hunting, trading, and consuming wild animals can expose humans to novel viruses carried by these animals, leading to the discovery of new infectious diseases. In areas where agriculture is crucial for survival, dependence on wildlife can further heighten the risk of zoonotic transmission. This phenomenon has led to the emergence of several infectious diseases with global epidemic potential, including Ebola, Zika, and the Coronavirus Disease 2019 pandemic.

One significant factor contributing to the discovery of new diseases caused by viruses, is the mutation of already known viruses. Ribonucleic acid (RNA) viruses often exhibit unusually high mutation rates because the errors made during the replication of their RNA genomes are not subjected to proofreading. Some mutations convert existing viruses into new genetic varieties with pathogenic effects even in individuals who were immune to the original virus. Influenza outbreaks for example, can be caused by new strains of the influenza virus that differ genetically from previous strains, resulting in minimal immunity among people.

A second process leading to the emergence of new viral diseases is the spread of a virus from a small, isolated population of humans. In the case of acquired immunodeficiency syndrome (AIDS), various technological and social factors, such as the ability to travel to foreign countries, blood transfusions, and intravenous drug abuse, allowed a previously not so common human disease to become a worldwide threat.

Furthermore, a source of new viral diseases in humans can also be the transmission of existing viruses from other animals. It is estimated that about most of the new diseases affecting humans come from animals. Animals that harbor and can transmit a particular virus but remain generally unaffected by it act as a natural reservoir for the virus. Influenza epidemics are a prime example of the potential consequences of viruses transitioning from one species to another. Generally, pandemics begin with the mutation of the virus during its transition from one host species to another. Strong suspicions of human-to-human transmission arise when the disease caused by the virus is observed in many members of the same family.

Emerging viruses are not new in general. They are existing viruses that mutate, spread within the host species they already infect more widely or to new host species. Changes of the environment or in host behavior can increase the mobility of viruses responsible for emerging diseases. The construction of new roads in areas that are remote can spread viruses to previously isolated human populations. Additionally, the destruction of the forest for agricultural expansion can make humans come into contact with animals that may host pathogenic viruses.

Beyond human-nature interactions, the spread of viruses is exacerbated in countries with poor economic conditions and inadequate hygiene practices. Socioeconomic disparities often result in densely populated areas, facilitating close contact among individuals and increasing the likelihood of virus transmission through respiratory droplets or direct physical contact. Inadequate or limited healthcare infrastructure and lack of access to clean water also create conditions conducive to virus transmission.

In such environments, viruses can spread rapidly, as effective prevention measures, such as vaccinations and personal protective equipment, are challenging to implement. This lack of damage control can prolong infectious periods and delay diagnosis and treatment, allowing viruses to spread unchecked within communities. Furthermore, economic instability may compel individuals to continue working even when ill, contributing to the spread of viruses in workplaces and public spaces.

Addressing these complex challenges requires comprehensive public health interventions that prioritize improving access to healthcare, promoting hygiene and sterilization practices, and implementing measures to reduce human-wildlife interaction.

While these are essential priorities for managing epidemics, developing new drugs and vaccines for viruses currently lacking treatments is equally crucial. This is because transmissible viruses can evolve into pandemics, threatening global health and economies. Developing new therapies and vaccines can protect populations from the effects of viruses and reduce disease spread. Therefore, the timely discovery and application of effective treatments and vaccines are vital for preventing future pandemics and safeguarding global public health [5].

2.2 Virus

2.2.1 Discovery

Viruses lack the metabolic mechanisms and structures found in cells and are primarily just genes encased in a protein shell. Therefore, they are considered either as the most complex assemblies of biological macromolecules or the simplest forms of life. Previously, they were thought to be chemical substances with biological activity. Their ability to cause a wide variety of diseases and spread from one organism to another led researchers in the late 19th century to believe that viruses were similar to bacteria. However, viruses cannot perform metabolic activities or reproduce outside their host cells. Most scientists studying viruses come to the agreement that they are not living organisms but rather exist in a state of "borrowed life."

Viruses can infect all forms of life, not only plants, bacteria and animals but also algae, archaea, fungi and other protists. The genome of a virus can show more similarities to the genome of its host than to the genome of a virus with a different host. Indeed, the sequence of some viral genomes matches the sequence of certain host genes to a significant degree.

The first experiments leading to the discovery of viruses were conducted on plants, specifically the tobacco plant. One disease affecting this plant is tobacco mosaic disease, which stunts plant growth and makes the leaves have a mottled appearance. In 1883, German scientist Adolf Mayer found that he could transmit this disease from one plant to another by rubbing healthy plants with an extract from diseased ones. He could not find any infectious microbe and assumed that the disease occurred because of bacteria.

Later, Dutch botanist Martinus Beijerinck conducted experiments, showing that the infectious agent in the filtered extract could reproduce. He found that the pathogen reproduced only within the host it infected and could not be cultivated in a nutrient medium or test tube. He hypothesized that it was a much smaller and simpler reproducing particle than bacteria, and he is considered by many to be the first scientist to articulate the concept of a virus. In 1935 his suspicions were confirmed when American scientist Wendell Stanley managed to crystallize the infectious particle [6].

2.2.2 Structure

The smallest viruses have a diameter of only 20 nm, smaller even than ribosomes. When examined in more detail, they consist of one or more nucleic acid molecules enclosed within a protein shell and sometimes a membranous envelope. Their genome can be composed of double-stranded deoxyribonucleic acid (DNA), single-stranded DNA, double-stranded RNA, or single-stranded RNA, depending on the type of virus. Based on the type of nucleic acid in their genome, viruses are classified as DNA viruses or RNA viruses. The genome of the smallest known viruses has only four genes, while the largest can have thousand genes.

Capsid is the protein shell surrounding the viral genome. Depending on the virus, the capsid can be polyhedral, rod-shaped, or have a more complex structure.

Capsids are composed of many protein subunits called capsomeres. The tobacco mosaic virus has a rod-shaped, rigid capsid made of over a thousand molecules of a single type of protein arranged in a helix. Viruses with a rod-shaped structure due to their helical arrangement are called helical viruses. The capsid of adenoviruses, many of which infect the respiratory tract of animals, is polyhedral and consists of 252 identical protein molecules arranged to form 20 triangular faces, making an icosahedron. Therefore, these viruses and all others with the same shape are called icosahedral viruses.

Viruses that infect bacteria has been found to have many of the most complex capsids, known as bacteriophages or simply phages. The three phages T2, T4, and T6 have many structural similarities. Their DNA is enclosed in an elongated icosahedral head, to which a protein tail with tail fibers is attached, enabling them to attach to the bacteria they infect.

Some viruses possess additional structures that aid in infecting their hosts. Such structure is the viral envelope, which is a membranous layer surrounding the capsids of influenza and various other animal viruses. This envelope is derived from the host cell's membranes and incorporates phospholipids and membrane proteins from the host cell.

Glycoprotein molecules of viral origin protrude from the envelope's outer surface, attaching on the host cell surface to specific receptor molecules. These glycoproteins on the viral envelope bind to specific receptor molecules of the host cell, facilitating viral entry. The envelopes of some viruses do not originate from the host cell’s cytoplasmic membrane.

These few viral parts work together with the host cell's components to produce many viral progenies [6].

2.2.3 Function

To understand how a virus can pose such a significant global threat and why timely prevention and management are necessary, it is essential to analyze how viruses operate.

Viruses can only reproduce within their host cells. They do not have the metabolic enzymes and equipment, required for protein synthesis. Thus, they are obligate intracellular parasites, meaning they can only replicate within a host cell. Viruses are sets of genes transferred from one host cell to another.

Moreover, they show specificity in the cells they infect. Only a certain amount of host cells can be infected by each virus type, known as the virus's host range. This specificity is due to the evolution of virus-host recognition systems. Viruses recognize host cells through proteins on their surfaces that fit like a lock and key with specific receptor molecules on the host cell's surface. Some viruses have a wider host range, while others are so limited that they infect only a single species. In multicellular eukaryotes, there is an additional level of specialization, with viruses usually restricted to specific tissues.

Viral infection starts when the virus attaches to the host cell and the viral genome enters into the cell. Depending on the host cell and virus type, the mechanism of genome entry differs. For example, T2, T4, and T6 phages inject their DNA into bacteria using their tail apparatus. Other viruses enter host cells through endocytosis or, for enveloped viruses, by the fusion of their envelope with the cell membrane. After the viral genome and capsid enter the cell and the capsid is broken down by cellular enzymes, the viral genome is released into the cytoplasm.

Then encoded by the viral genome the viral proteins command the host's materials and machinery, reprogramming the cell to replicate the viral nucleic acid and synthesize viral proteins. The host provides the nucleotides for making viral nucleic acids and the amino acids, enzymes, transfer RNAs, ribosomes, adenosine triphosphate (ATP), and other components needed for synthesizing viral proteins.

Most DNA viruses use the host cell's DNA polymerases to synthesize copies of their genome using the viral DNA as a template. RNA viruses, on the other hand, replicate their genome using polymerases encoded by their genome, which can use RNA as a template. The envelope glycoproteins are transported to the cytoplasmic membrane via vesicles. A capsid assembles around each viral genome molecule. New viruses exit the cell, each bearing numerous glycoproteins on its membrane, which is derived from the host cell. After exiting the host cell, the viruses can infect other cells.

A simple type of viral reproductive cycle concludes with the release of thousands of viruses from the infected cell, frequently leading to the host cell's destruction. Death and cellular damage of the host cell, along with the body's response to viral infection, are responsible for many symptoms associated with viral infections [6].

2.2.4 Classification

Phages

Phages are viruses that infect bacteria. They have been extensively studied and are known for their complexity. The study of phages contributed to the discovery that a number of viruses with double-stranded DNA as their genetic material reproduce through two alternative mechanisms: the lytic cycle and the lysogenic cycle. In the lytic cycle, the phage injects its DNA into the bacterial cell, hijacks the cell's machinery to produce new phages, and eventually causes the cell to burst, releasing new phages. In the lysogenic cycle the phage DNA integrates into the bacterial genome and replicates along with it, without killing the host. Under certain conditions, it can switch to the lytic cycle.

Plant viruses

There are over 2.000 known types of plant diseases caused by viruses. It is estimated that these diseases can cause a global loss of 15 billion dollars annually due to the damage they cause to agricultural and horticultural crops. Usual symptoms of viral infections in plants are discolored spots on fruits and leaves, stunted growth, and damage to roots or flowers. All of these symptoms reduce the yield and quality of the crop. Plant viruses exhibit a similar structure as the one animal viruses have and follow a similar way of reproduction. Most of the discovered plant viruses until now, have RNA genomes. A lot of them have helical capsids, while some have icosahedral capsids.

Animal viruses

Animal viruses exhibit many variations in their infection and reproduction cycles. An important variable is the nature of the virus's genome is whether it consists of DNA or RNA and whether it is single-stranded or double-stranded.

Families of viruses with double-stranded DNA:

- Adenoviruses (respiratory diseases, tumors, no envelope)
- Papovaviruses (papilloma virus, like warts and cervical cancer, polyoma virus, like tumors, no envelope)
- Herpesviruses (envelope)
- Poxviruses (envelope)

Family of viruses with single-stranded DNA:

- Parvoviruses (parvovirus B19, mild rash, no envelope)

Family of viruses with double-stranded RNA:

- Reoviruses (rotavirus, diarrhea, no envelope)

The family of single-stranded RNA viruses is categorized based on the utilization of RNA as mRNA, as a template for mRNA synthesis, or as a template for DNA synthesis. The last type is retroviruses, which have the most complex reproductive cycle among RNA viruses infecting animals. These viruses possess an enzyme, which is the reverse transcriptase that uses RNA as a template for DNA synthesis, causing the genetic information flow to reverse from the usual DNA-to-RNA direction to RNA-to-DNA.

RNA used as messenger ribonucleic acid (mRNA):

- Picornaviruses (rhinoviruses (common cold), poliovirus, hepatitis A virus, other enteroviruses, no envelope)
- Coronaviruses (severe acute respiratory syndrome, envelope)
- Flaviviruses (zika, yellow fever, hepatitis C, west nile virus, envelope)

- Togaviruses (rubella virus, envelope)

RNA as a template for mRNA synthesis:

- Filoviruses (ebola virus, envelope)
- Orthomyxoviruses (influenza virus, envelope)
- Paramyxoviruses (measles virus, envelope)
- Rhabdoviruses (rabies virus, envelope)

RNA as a template for DNA Synthesis:

- Retroviruses (HIV (human immunodeficiency virus) and viruses causing leukemia, envelope) [6]

2.3 Immune system

2.3.1 Mechanisms

Immunology refers to the study of the physiological defense mechanisms through which the host organism recognizes, destroys, or neutralizes foreign bodies that invade the body, whether living or non-living matter. The defense mechanisms protect against infections from pathogens such as microorganisms and viruses, including bacteria and fungi, remove or isolate foreign bodies, and destroy cancer cells that are formed in the body. The immunological defense mechanisms can be categorized into two types the innate and the adaptive, which interact with one another.

The innate immunological defense mechanisms provide an immediate response to a wide range of foreign substances or cells invading the organism without identifying their specific identity. Consequently, these responses are not unique to any particular invader and are therefore referred to as non-specific immune responses.

The adaptive immunological defense mechanisms provide a specific response adjusted to particular pathogens and has the unique ability to remember past infections. This memory allows for a more efficient and rapid response during subsequent encounters with the same pathogen. While the innate response takes less time to activate than the adaptive response but the adaptive provides longer and highly specific protection.

The adaptive and innate immune systems are interconnected and cooperate together to form a defense strategy. The innate immune system controls the beginning of infection and provides crucial signals that shape the adaptive response. For instance, dendritic cells, part of the innate system, act as antigen-presenting cells that capture and present pathogen antigens to T lymphocytes (T cells), thus initiating and guiding the adaptive immune response. This interaction ensures that the adaptive immune system is activated and customized specifically to the pathogen encountered.

In conclusion, together they provide a robust and dynamic defense mechanism, with the adaptive system providing a targeted and longer response and the innate system offering immediate protection [7].

2.3.2 Antigen and antibodies

In general, an antigen is anything that can trigger an immune response because it is identified as foreign by the body's defense system. In the context of viruses, an antigen is a component found on the surface of the virus that the immune system recognizes as an invader. To combat this, the immune system produces antibodies, proteins designed to specifically bind to these viral antigens. By doing so, antibodies help neutralize the virus and assist in its removal from the body, effectively protecting against infections [7].

2.3.3 Cells

The immune system consists of many different cells that combat diseases and are found both in the blood and in tissues and organs throughout the body. These cells are different types of white blood cells known as leukocytes and they can be classified into two groups: lymphoid cells and myeloid cells. The myeloid cells consist of neutrophils, eosinophils, basophils, monocytes, macrophages, dendritic and mast cells and the lymphoid of natural killer (NK), B lymphocytes (B Cells), T Cells and plasma cells.

Neutrophils, produced in the bone marrow, perform phagocytosis (the process by which cells engulf and digest foreign particles or microorganisms) and release enzymes to kill microorganisms. Eosinophils, also from the bone marrow, combat parasitic infections and participate in allergic responses, while basophils release histamine during allergic reactions. Monocytes, which originate in the bone marrow and differentiate into dendritic cells and macrophages in tissues, are involved in phagocytosis and the presentation of the antigens. Macrophages further specialize in these functions and release cytokines to signal other immune cells.

Dendritic cells, located in various tissues, are key in antigen presentation and initiating adaptive immunity. NK cells, from the bone marrow, target virus-infected and tumor cells. B cells, also bone marrow derived, produce antibodies and present antigens to T cells.

T lymphocytes include helper T cells that aid other immune cells and cytotoxic T cells that destroy cells that are infected. Plasma cells, differentiated from B cells, release antibodies and mast cells, found in tissues, release histamine during allergic reactions. Additionally, cytokines, proteins released by immune cells, act as messengers to regulate immune responses, cell growth, and differentiation. These cells collaborate to protect the body from infections, remove damaged cells and coordinate immune responses [7].

2.3.4 Immune response to viral entry

When viral entry happens, the immune system initiates a coordinated response to identify, neutralize, and eliminate the pathogen. The entry can happen through various routes such as the respiratory and gastrointestinal tract or breaches in the skin.

Antigen-presenting cells (APCs), such as macrophages, dendritic cells and B cells, encounter the virus. These cells ingest the virus and break it down into smaller protein fragments (antigens). These fragments are then displayed on the surface of the APCs. Helper T-cells recognize these antigens on the APCs, they become activated and they release cytokines to enhance the response of B-cells and cytotoxic T-cells.

Cytotoxic T-cells directly kill cells that are infected, preventing further viral replication. B-cells, with assistance from Th cells, differentiate into plasma cells that generate specific for the virus antibodies. These antibodies bind to the virus, blocking its entry into host cells and tagging it for destruction by macrophages and neutrophils.

Some of the cytotoxic T cells, helper T cells, and B cells that are activated transform to memory cells. These cells remain in the body way after the infection has been resolved and provide a quicker and more robust response if the same virus invades the body again in the future [7].

2.3.5 Symptoms

When a virus enters the body, it can manifest in a range of symptoms that reflect the body’s efforts to combat the infection. Typically, individuals may experience a fever, which acts as a natural defense mechanism to inhibit viral replication. Respiratory viruses often trigger a persistent cough as the body attempts to clear the virus from the airways, accompanied by a sore throat. Fatigue and muscle aches frequently accompany viral infections and headaches may arise as a secondary effect. Additionally, respiratory infections can lead to a runny or stuffy nose, while gastrointestinal viruses might cause diarrhea or vomiting. These symptoms collectively illustrate the intricate interaction between viral invasion and the body’s defense mechanisms, highlighting the complex nature of the immune response [7].

2.4 Flaviviruses

2.4.1 Arboviruses

Vector-borne diseases are caused by pathogens and parasites that are transmitted to humans and other animals through vectors which are organisms that carry and transmit these infectious agents. Arboviruses unlike other viral groups defined by genetic relationships, are characterized by their shared mode of transmission rather than their evolutionary lineage. These viruses are all transferred by an arthropod vector and are considered responsible for several major diseases such as Yellow fever Dengue, Zika, West Nile and Japanese Encephalitis virus.

Arthropods are an ancient group of animals that have been present on Earth million years ago. They are ubiquitous and have developed diverse adaptations, including bloodsucking behaviors in several insects, some mites and all ticks. The transmission cycle of arboviruses involves an arthropod vector that becomes infected after feeding on the blood of a vertebrate host that is infected. The virus then replicates within the vector and is transferred to a new host.

Environmental factors can critically influence the survival and replication of both the arthropod vectors and the viruses they carry, impacting the prevalence and spread of arbovirus-related diseases. Understanding and controlling the transmission of arboviruses is vital for public health, given their potential to cause widespread outbreaks with severe health consequences [8].

2.4.2 Proteases

Proteins are molecules composed of amino acid chains that are connected by peptide bonds. Each amino acid has the same main structure. It includes an alpha carbon bonded to a hydrogen atom, central carbon atom, an amino group (NH_2), and a carboxyl group (COOH).

The side chain (R group) attached to this central carbon is what varies and it differs between each of the 20 standard amino acids. The side chain decides the properties and characteristics of the amino acid, such as its size, acidity, polarity, and whether it is hydrophobic (water-repelling) or hydrophilic (water-attracting).

Peptide bonds are covalent bonds created by a condensation reaction between the amino group of one carboxyl acid and the amino group of another, resulting in the release of a water molecule. They are generally strong bonds, meaning that enough energy must be supplied to the molecule to overcome the attractive forces that holds the atoms together.

Structurally, proteins can be categorized into four levels: primary, secondary, tertiary, and quaternary. The primary structure is the sequence of amino acids that are linked together by peptide bonds. It determines the protein’s overall shape and function, as the particular order of amino acids influences how the protein will fold and interact with other molecules.

The secondary structure addresses the local folding of the polypeptide chain into specific shapes stabilized by hydrogen bonds that link different parts of it. Hydrogen bonds are weak, non-covalent interactions that occur between a hydrogen atom that is covalently bonded to an electronegative atom and another electronegative atom.

The tertiary structure refers to the three-dimensional shape of a single polypeptide chain, created by the folding of secondary structures into a compact, globular form. This shape is stabilized by interactions among the side chains of the amino acids within the polypeptide chain, such as ionic and hydrogen bonds, van der Waals forces, and disulfide bridges.

Ionic bonds, also known as salt bridges, are stronger bonds than hydrogen bonds that form between negatively and positively charged side chains. Van der Waals forces are weak, non-covalent interactions that occur between all atoms in close proximity and disulfide bridges are covalent bonds that form through an oxidation reaction where two cysteine residues form a disulfide bond.

The tertiary structure is essential for the protein's functionality, as the specific folding determines the active site and interaction sites with other molecules**.** Some proteins have a quaternary structure, where multiple polypeptide chains come together to form a functional complex. Similar types of interactions as in tertiary structure hold the chains together.

Proteins, also, have two distinct ends, the N-terminus and the C-terminus. The N-terminus, characterized by a free amino group, represents the beginning of the amino acid sequence and is often referred to as the 5' end. Conversely, the C-terminus, with a free carboxyl group, signifies the end of the amino acid sequence and is known as the 3' end.

Each protein can vary greatly in size. Peptides are short chains of fewer than 50 amino acids, polypeptides are longer chains ranging from 50 to 1000 amino acids but not yet fully functional proteins, proteins are fully folded polypeptides that many different lengths and multi-subunit proteins consist of multiple polypeptide chains.

Furthermore, each one’s unique structure, determined by the sequence of amino acids, dictates its specific function within the cell. They can serve as structural components, as enzymes, which catalyze biochemical reactions, as transporters, and as signaling molecules. Proteins are also involved in immune responses, cell movement, and regulation of gene expression. Many proteins act as enzymes, which are catalysts that speed up chemical reactions in the body. Amino acids in the enzyme’s active site interact with substrates, facilitating the conversion of reactants to products [9].

Most catalysts in biological systems are enzymes, and almost all enzymes are proteins. Enzymes are highly specialized and they can increase the speed in which chemical reactions are happening in living organisms. To understand their function, it is important to understand their structure.

Substrates are molecules that enzymes selectively recognize and bind to. This high specificity ensures that the enzyme catalyzes only the intended reaction. When a substrate binds to an enzyme, it interacts with the enzyme's active site binding to it through various interactions, forming an enzyme-substrate complex.

The active site is a specially shaped pocket on the enzyme’s surface that is complementary to the substrate, allowing it to fit like a key in a lock. This site, though typically a small part of the enzyme’s entire structure, is crucial for its catalytic function. The specificity of an enzyme is largely determined by the unique structure and chemistry of its active site. This site is composed of amino acid residues that create a specific chemical environment. For instance, in many enzymes, a catalytic triad of amino acids works together to catalyze reactions.

At the entry of the substrate to the active site, the enzyme experiences a minor conformational change, optimizing the interaction and reducing the activation energy needed for the reaction. Lowering the activation energy means that the enzyme makes it easier for the chemical reaction to happen by reducing the energy barrier that the substrate must overcome to be converted into the product. This leads to the efficient conversion of the substrate into the product.

Once the reaction is complete, the products, which have lower affinity for the active site than the substrate, are released. This reduced affinity means that the products do not bind as tightly, allowing them to be released from the active site. The enzyme returns to its original conformation, prepared to bind with a new substrate molecule and do the same process all over again.

Additionally, cofactors and coenzymes play a crucial role in assisting the enzyme's catalytic activity. Cofactors, which can be metal ions like or organic molecules, are necessary for many enzymes to be fully active. Coenzymes, a type of organic cofactor usually derived from vitamins, assist in enzyme activity by carrying chemical groups between molecules during reactions.

Enzymes can also be regulated allosterically, where molecules bind to sites other than the active site, known as allosteric sites. When a molecule binds that way, it makes a conformational change in the enzyme's structure. This change can either enhance or inhibit the enzyme’s activity. When an allosteric molecule binds, it can increase the receptivity of the active site by making it more accessible or stabilizing a conformation that enhances catalytic efficiency. This regulatory mechanism enables the cell to adjust enzyme activity as needed, ensuring that reactions proceed at optimal rates

The enzymes have six major classes. The oxidoreductases catalyze oxidation-reduction reactions. They facilitate the transfer of electrons between molecules, usually by transferring hydrogen or oxygen atoms. The transferases transfer functional groups between molecules. Lyases destroy various chemical bonds without using oxidation and hydrolysis. Isomerases arrange atoms inside a molecule, converting it into its isomer and ligases join two molecules. Finally, hydrolases catalyze the cleavage of bonds through the addition of water (hydrolysis). This class of enzymes includes proteases, lipases, and nucleases.

Proteases are a specific type of enzyme that play a critical role in protein metabolism by decomposing proteins into smaller peptides or individual amino acids. This process, known as proteolysis, cleaves the peptide bonds between amino acids in a protein chain. In the context of viral infections, such as those caused by flaviviruses, proteases are essential for the virus's life cycle.

The flavivirus protease is classified as a serine protease because it utilizes a catalytic triad, composed of histidine, serine and aspartate, which operates in a mechanism similar to that of trypsin. Like trypsin, it uses its serine residue to cleave peptide bonds, targeting specific sites just after certain amino acids. Given their essential role in viral replication, proteases make excellent targets for antiviral drugs, as inhibiting them prevents viral spread [10].

2.4.3 Treatment

Vaccines is the primary way to prevent viral infections. They are harmless forms of a pathogen or its derivatives capable of stimulating the body's immune system to develop defenses against the harmful pathogen. They can completely eradicate certain diseases and prevent some viral diseases, but current medical technology cannot cure most of the viral diseases once they occur.

For treating a virus, often antiviral drugs are needed. They can be a critical tool in managing and curing viral infections, especially when vaccines are unavailable or ineffective. They target specific viral enzymes critical to the virus's ability to replicate and spread. The most common practice is inhibiting viral proteases, enzymes that cleave viral polyproteins into functional parts necessary for the virus to mature. These antiviral drugs, known as protease inhibitors, are designed to interact with either the active site or the allosteric site of the enzyme.

When a drug targets the active site of the protease, it acts as a competitive inhibitor. This means the drug molecule is structurally the same to the enzyme’s natural substrate and competes directly to bind to the active site. By occupying this site, the drug prevents the enzyme from processing its natural substrate, which in the case of viral proteases, means the virus cannot cleave its polyproteins into functional proteins. This blockage effectively stops the virus from producing the components needed to assemble new viral particles, halting replication.

In contrast, when a drug binds to the allosteric site, a different region on the enzyme, it functions as a non-competitive inhibitor. Binding to the allosteric site changes the conformation of the enzyme, which then alters the structure of the active site. This alteration can reduce the enzyme's ability to bind its natural substrate or render the active site completely inactive. As a result, the enzyme can no longer catalyze the necessary cleavage reactions, further inhibiting the virus's ability to replicate.

Both mechanisms, competitive and non-competitive inhibition, specifically target viral proteases disrupting the virus's lifecycle without affecting human proteases, offering a targeted therapeutic approach. However, it's important to note that viruses can mutate over time, potentially developing resistance to these drugs, which underscores the need for ongoing research and development of new antiviral therapies [10].

Currently, there are no safe therapeutic options for treating flavivirus infections and supportive care is the primary approach, although it is often not very effective. The symptoms of viral infections vary depending on how the virus interacts with host cells, and many symptoms, such as fever and pain, are primarily due to the immune system's response as it attempts to eliminate the virus. For West Nile virus and dengue, no licensed vaccines exist, and treatment remains focused on symptom management rather than targeting the virus.

Similarly, while vaccines are available for Japanese encephalitis and yellow fever, they have limitations, and their use is mainly recommended for those in endemic areas. The challenge is further compounded by the lack of specific antiviral drugs and the difficulty in scaling vaccine production to meet global demand, particularly for dengue. Substantial gaps remain in our understanding of flaviviruses and their associated diseases, highlighting the need for continued research and development [9].

2.4.4 Structure

The flavivirus is a small, icosahedral, enveloped particle, 40–60 nm in diameter, with a 30 nm nucleocapsid core. It contains three structural proteins: the membrane protein M, the capsid protein C and envelope protein E. The C protein encases the RNA genome, forming the nucleocapsid, which is encircled by a lipid bilayer originating from the host cell, anchoring the M and E proteins. The E protein is vital for recognizing host-cell receptors and serves as the main target for neutralizing antibodies.

The flavivirus genome is a positive, single-stranded RNA, approximately 11.000 nucleotides long, functioning as a single RNA messenger, which is used to produce the proteins needed by the virus. This RNA has a part called an "open reading frame" (ORF), which is a sequence of genetic code that can be translated into a protein.

On either side of the ORF, there are regions of the RNA called noncoding regions (NCRs). Although they don't make proteins, these NCRs play important roles. They can fold into specific shapes, known as RNA secondary structures, which may help the virus replicate its genome, produce proteins, and package the RNA into new virus particles.

Non-structural proteins such as NS1, NS2A, NS2B, NS3, NS4A, NS4B, and NS5 are crucial for the replication of the virus and assembly. NS1, a homodimer, is vital for viral RNA replication, as do NS4A and NS4B, which are small, hydrophobic, membrane-associated proteins. NS2A, a small viral protein, aids in replication by preparing RNA templates and supporting the machinery necessary for this process. Additionally, NS2A can interfere with the host's immune response, helping the virus evade detection.

NS2B, another small membrane-associated protein, functions as a cofactor for the NS2B–NS3 complex, which activates a protease responsible for cleaving the viral polyprotein at specific sites. NS3, a large cytoplasmic protein, is involved in several enzymatic activities, including polyprotein processing and viral RNA replication. Notably, NS3 also seems to play a role in the assembly of the virus. NS5, the biggest and most conserved protein, is crucial for the virus's life cycle, underlining its importance across different flavivirus species [9].

2.4.5 Expansion and symptoms

Yellow Fever

Yellow fever (YF) is an African disease caused by the yellow fever virus (YFV), which has a single serotype and several genotypes [9]. The first documented outbreak occurred in Barbados in 1647, spreading to the Yucatan peninsula in 1648 and throughout the Caribbean, with regular outbreaks in Cuba. Major outbreaks followed, including a significant one in Philadelphia in 1793 that killed almost 10% of the population, and the biggest outbreak in America in 1878 along the Mississippi River, that claimed over 20.000 lives.

The last outbreak in America was in New Orleans in 1905. Initially believed to be caused by "miasma," YF was later proven to be mosquito-borne by the Yellow Fever Commission in 1900, confirming earlier hypotheses by J.C. Nott and Carlos Finlay [11]. Dramatic upsurges in YFV activity occurred in Africa in the 1960s and late 1980s, involving over 100.000 cases, with more recent outbreaks in Brazil, Paraguay, Argentina, Uganda, Sudan, and Ethiopia. Despite vaccination efforts, surveillance remains predominantly passive, leading to an underestimation of the true incidence, especially in endemic areas [9].

YFV infection ranges from sub-clinical to severe hemorrhagic disease and death. The disease progresses through an "infection" phase, with flu-like symptoms like fever and malaise, to a severe "intoxication" phase, characterized by jaundice, hemorrhage, multi-organ dysfunction and death seven to ten days after onset. YFV directly infects liver cells, causing acute injury and potentially leading to vascular leakage and coagulation issues. While there are reports of disseminated intravascular coagulation and loss of coagulation factors in patients that were infected, much about YF pathogenesis remains unclear, necessitating further research [11].

Dengue

"Dandy Fever" and "break-bone fever" were terms used in the late 1700s to describe what were later identified as outbreaks of dengue and Chikungunya virus (CHIKV) infections. Dengue, caused by DENV infection, was distinguished from CHIKV by the absence of persistent arthritis and the presence of a rash and headaches. The first official description of dengue was by David Bylon during 1779.

In the 1800s, it was hypothesized that both yellow fever and dengue were transmitted similarly and it was also noted that they occurred in the same locations. Both diseases were identified as filterable agents linked to the Aedes aegypti mosquito, and later, Aedes albopictus was also identified as a vector for dengue.

Using various immunological tests, Sabin and Schlesinger identified two immunological types of dengue virus. Immunological types refer to distinct variations of the virus that elicit different immune responses from the host. Further serological assessments revealed that in dengue virus (DENV), there are four immunological types, also known as serotypes, DENV-1, DENV-2, DENV-3, and DENV-4.

Each serotype has unique surface proteins that the immune system identifies and responds to. When an individual is infected with one serotype, they have immunity to that particular serotype. However, this immunity does not fully protect against the other three serotypes. In fact, subsequent infections with a different serotype can result in more severe disease.

A significant outbreak in America in 1922 affected one to two million people. Today, about 40% of the global population is vulnerable to DENV infection, primarily in tropical and subtropical regions, with an estimated 390 million cases annually [11].

Specifically, the virus affects the continents of Africa, Asia, Oceania, and the Americas. However, information on cases in African endemic countries is limited, and the true prevalence is obscured by numerous asymptomatic infections in big urban regions. These factors make dengue the most prevalent arbovirus in the world. The complexity of dengue pathogenesis continues to be a focus of ongoing research [9]. with a variety of clinical presentations, so the severity of the disease was determined based on clinical observations.

Dengue is a disease with a variety of clinical presentations, so the severity of the disease was based on clinical observations. The typical symptoms include a fever, nausea, headache, rash, vomiting, and muscle and joint pains. This initial phase generally lasts for five to seven days, with recovery being straightforward for most individuals. In some cases, however, patients may experience a sudden worsening after the febrile phase, marked by warning signs such as abdominal pain, lethargy, fluid accumulation, persistent vomiting, liver enlargement, and mucosal bleeding.

Severe dengue, or dengue hemorrhagic fever (DHF), is characterized by severe plasma leakage, leading to fluid accumulation in body cavities and organs, which can cause respiratory distress and hypotension. Without prompt treatment, severe dengue can progress to dengue shock syndrome (DSS) and result in significant bleeding and multi-organ impairment, affecting the liver, heart, central nervous system, and pancreas [9].

DHF/DSS is more common in secondary infections, especially in children or in newborns. In practical terms, the grading system for severe DENV infection is not well-defined prompting efforts to refine and enhance the classification [11].

Japanese Encephalitis

Epidemics of encephalitis in Japan date back to 1871, with a notable outbreak in 1924 affecting 6.000 people and causing a 60% fatality rate. The causative agent, Japanese encephalitis virus (JEV), was first isolated from non-human primates in 1933 and further characterized in mice during a 1935 outbreak.

JEV has five distinct genotypes but there are no differences among them, they all form a single serotype. Approximately three billion people across 24 countries, mainly in Asia, live in JEV-endemic areas. The annual incidence is around 70.000 cases, influenced by geographic and climatic factors as well as vaccination rates, with an estimated 14.000–20.500 fatalities.

The virus is transmitted by Culex mosquitoes that are the primary vectors and with vertebrate hosts like pigs and domesticated birds. Pigs serve as significant amplifying hosts, while birds primarily facilitate the spread to new areas.

The majority of human infections with JEV are asymptomatic. Acute encephalitis is the most frequently observed clinical manifestation. with symptoms like headache, myalgia, diarrhea, and vomiting. It can escalate to neurological complications such as acute flaccid paralysis, convulsions, mental confusion, severe encephalitis, coma and death. Approximately 30% of survivors of severe disease experience neurological sequelae, including seizures, physical disabilities, and cognitive deficits [11].

West Nile

West Nile virus (WNV) was initially identified as a human pathogen in 1937 in Uganda. Later on, its presence was reported in Asia, Africa, Europe and Australia. It was not considered as a significant human health issue until the late 1990s due to its rare outbreaks, which were sporadic, linked to low pathogenicity and typically resulted in mild neurological diseases.

The first outbreak in humans was documented in Israel in 1950. Notable outbreaks then occurred in France, South Africa, Algeria, Romania, Tunisia, and Russia between the 1950s and 1990s. In August 1999, the WNV was first identified in New York City and it successfully established and dispersed throughout America. Today, WNV is globally distributed and it is transmitted by *Culex* spp. mosquitoes with Passeriformes birds being the hosts.

Most of WNV infections in humans are asymptomatic. Typical symptoms are an abrupt fever, myalgia, headache, fatigue, nausea, weakness, diarrhea and vomiting, often developing two to fourteen days after virus infection. The illness usually remains for two to five days.

The neuroinvasive infection by WNV is characterized by altered mental status (disorientation, coma and stupor) meningitis, encephalitis, and/or poliomyelitis with long-lasting neurological complications. Infrequent complications are myocarditis, Guillain–Barrι syndrome, respiratory failure and death. The fatality rate in patients with neurological symptoms is approximately 10 %, with the elderly and immunocompromised patients being more at risk [9].

2.5 Zika virus

2.5.1 Expansion

The Zika virus was first isolated in 1947 from a rhesus monkey in Uganda's Zika Forest [12]. It remained largely unnoticed for decades due to its limited geographical spread in East and West Africa and its typically mild or asymptomatic infections [13]. However, this changed dramatically in 2007 with an outbreak on the Pacific Island of Yap, marking the first documented human epidemic. The virus continued to spread, causing significant outbreaks in French Polynesia and other South Pacific islands between 2013 and 2014 [12].

In 2015, ZIKV was detected in Brazil [14], where it rapidly proliferated, leading to an epidemic that affected hundreds of thousands of people across the Americas and the Caribbean [12]. The Brazilian Health Ministry estimated over one million infections by the end of 2016 [14]. Since then, ZIKV has caused a global health emergency, with approximately 84.000 reported cases in 2016 [15], and it has impacted millions of people in over 40 countries, including those in North and South America, Europe, and Asia [12]. The virus's ability to cause large-scale outbreaks is largely due to its vectors and the regions in which they circulate [14].

2.5.2 Transmission

Zika virus (ZIKV) is mainly transmitted through *Aedes* mosquitoes, particularly *Aedes aegypti* and *Aedes albopictus* [15]. There is also evidence suggesting that mosquitoes can acquire ZIKV if the water is contaminated with human urine, highlighting the risk of transmission in areas with poor sanitation. However, the virus can also spread through non-vector means, including sexual transmission [14], as it has been detected in human spermatozoa [12].

Additionally, ZIKV can be transmitted to the fetus from the mother while she is pregnant, as the virus is capable of crossing the placental barrier, leading to severe birth defects. Another potential route of transmission is through breast milk, as viable ZIKV has been identified there, posing a risk of mother-to-child transmission [9].

2.5.3 Symptoms

ZIKV infection often causes mild symptoms, like rash, fever, conjunctivitis, muscle and joint pain, headache, and fatigue. These symptoms typically last for two to seven days and are often so mild that they go unnoticed [9]. However, ZIKV poses a significant health threat due to its association with severe neurological and immunological complications such as microcephaly in newborns and Guillain-Barré syndrome in adults.

Microcephaly is a medical condition characterized by abnormal fetal growth in the uterus, impacting the development of the central nervous system, which can result in permanent cognitive impairment or death. Guillain-Barré syndrome, is more common in young adults where the immune system attacks the peripheral and spinal nerves causing the demyelination of them and leading to muscle weakness or paralysis [14].

2.5.4 Structure

Zika virus, is an arbovirus member of the Flavivirus genus within the Flaviviridae family. It has a single-stranded RNA genome comprising of almost 10.794 nucleotides of positive polarity and an envelope [15]. This genome encodes a single polyprotein that is cleaved by viral and cellular proteases into structural proteins that form the viral particle (capsid, precursor membrane/membrane, and envelope) and nonstructural proteins (NS1, NS2A, NS2B, NS3, NS4A, NS4B, and NS5) that are involved in the replication of the virus within host cells.

The viral NS2B-NS3 complex is a chymotrypsin-like serine protease with a key role in processing the ZIKV polyprotein, facilitating the release and maturation of individual proteins [14]. The catalytic triad of the protease is Ser135, His51, and Asp75, with NS2B acting as a cofactor to enhance NS3's activity. The protease can adopt an "open" conformation when a substrate or inhibitor is absent, where NS2B shows limited interaction with NS3, and a "closed" conformation upon binding a substrate or inhibitor, resulting in a more compact structure [16]. Understanding these structural details is crucial for developing targeted antiviral treatments [15].

2.5.5 Treatment

The creation of efficient and safe vaccines and antiviral drugs is crucial for controlling and managing outbreaks and complications associated with ZIKV infections. Although multiple vaccines have demonstrated considerable promise in human clinical trials, none have yet received approval from the World Health Organization (WHO) or other global health agencies.

In the absence of approved vaccines, the exploration of clinically available antiviral drugs offers a potential alternative for containing Zika infections [12]. However, no specific antiviral drugs have been approved either [13]. As a result, current clinical management relies on supportive care, including hydration, rest and the use of nonsteroidal anti-inflammatory and antipyretics drugs. While these measures help alleviate symptoms, they offer only limited effectiveness, underscoring the need for continued research into more targeted therapies [14].

Significant efforts have been dedicated to the development and identification of compounds with inhibitory potential against ZIKV, particularly focusing on small molecule protease inhibitors. In recent years, several ZIKV NS2B-NS3 protease inhibitors have been identified through virtual screening approaches. The inhibition of the NS2B-NS3 protease complex is especially noteworthy because it plays a pivotal role in the replication of the virus, making it a promising molecular target for the development of antiviral drugs against this emerging flavivirus.

The application of virtual screening and computational modeling has accelerated the traditional drug discovery process, enabling the rapid identification of potential inhibitors. This approach is recognized as one of the most effective strategies for discovering and developing new, safer drugs. Furthermore, leveraging existing knowledge from studies on compounds effective against other flaviviruses has proven invaluable in expediting the discovery of therapeutics against ZIKV [12].

Notably, most reported effective small molecules are designed to target allosteric sites instead of the enzyme's active site. [13]. The preference for allosteric inhibition arises due to limitations associated with directly targeting the active site. The active site of the NS3 enzyme is inherently small, structurally rigid and relatively inaccessible. These characteristics make it challenging for competitive inhibitors to bind effectively and retain specificity. Moreover, even before any structural changes occur, this specificity can complicate drug development and increase the risk of resistance as the virus mutates.

Things become even more difficult after the enzyme undergoes a shape change either due to an NS3 inhibitor or a substrate binding to the active site of the enzyme. Specifically, the NS2B protein adopts a closed conformation that obscures the active site, further obscuring access to it. This can partially or completely block the inhibitor's access to essential binding regions, resulting in incomplete binding and, consequently, ineffective inhibition of the enzyme's catalytic activity.

Even if the inhibitor manages to bind, the altered enzyme structure may weaken the interaction, leading to a fragile bond, thereby reducing the drug's overall effectiveness. This insufficient inhibition allows the enzyme to maintain some level of activity, diminishing the therapeutic impact of the drug. Over time, incomplete inhibition can facilitate the continued replication of the pathogen, potentially leading to the development of drug resistance as the pathogen adapts to evade the inhibitory effects.

Ultimately, these issues can culminate in therapeutic failure, where the treatment does not achieve the desired clinical outcome. Addressing these structural challenges is therefore critical for successful drug design and effective disease control.

In contrast to competitive inhibitors that compete with the substrate for the protein's active site, allosteric inhibitors function by binding to other sites on the surface of the protein. That way they can modulate protease activity by inducing conformational changes rather than completely blocking the enzyme, thereby allowing for a more controlled inhibition and a way less invasive to influence the catalytic function of the enzyme.

This subtle modulation reduces the risk of side effects and minimizes toxicity to host cells. Toxicity in host cells refers to the harmful effects that a drug or compound can have on the cells of the organism being treated, rather than just on the pathogen it is targeting. Additionally, allosteric sites are generally less susceptible to mutations compared to the active site, which can further reduce the likelihood of resistance.

Furthermore, allosteric inhibition offers the potential for synergistic interactions with active site inhibitors, providing an avenue for developing more robust and durable antiviral therapies against ZIKV.

However, in allosteric inhibition, sites for binding are often more challenging to discover and define. They are often less conserved across strains, meaning that viral mutations could easily evolve to render allosteric inhibitors ineffective, leading to quicker resistance development. Since allosteric inhibition works by inducing conformational changes in the enzyme, the resulting effects on enzyme activity can be less predictable and harder to measure than with competitive inhibitors.

On the other hand, competitive inhibitors directly block the substrate from binding, providing an immediate and measurable impact on enzyme activity, which is critical in the initial phases of drug development. The structure of the active site is often easier to map, allowing for structure-based drug design and more straightforward optimization of drug candidates. Each approach has positives and negatives. The choice between them depends on the specific virus and enzyme [14].

Several Food and Drug Administration (FDA) approved drugs, originally developed for other uses, have shown promise in combating Zika virus. Sofosbuvir is approved for Hepatitis C, hydroxychloroquine for malaria, lupus, and rheumatoid arthritis, while azithromycin and novobiocin are antibiotics. These repurposed drugs highlight the strategy of leveraging existing therapeutics for new disease applications, including combating Zika virus [17].

METHODOLOGY

For the prediction of potential antiviral compounds that may be effective against the Zika virus, three different methodologies were used: statistical analysis, machine learning, and in silico molecular docking experiments. The primary advantage of these approaches, and the reason they are increasingly applied in scientific research, particularly in the health sector, is their ability to reduce cost and time.

Specifically, statistical analysis aids in a better understanding and prediction of results, while the creation of machine learning models allows for the processing of large datasets and the recognition of patterns that are not easily discernible by humans. Concurrently, in silico analysis can expedite the search for new compounds by enabling the assessment of the antiviral activity of thousands of potential molecular structures. This approach also helps avoid the production and experimental testing of these numerous compounds in the laboratory, a process that is both time consuming and costly.

3.1 Data preprocessing

3.1.1 Statistical analysis & machine learning

For data preparation, the ChEMBL [18] database was utilized to identify potential antiviral compounds against the Zika virus. Since this database does not have the

NS2B-NS3 protein specific to the Zika virus, a search was conducted using the target organism.

Navigating to the activity charts Half Maximal Inhibitory Concentration (IC50) and Half Maximal Effective Concentration (EC50) values, which are critical measures of a compound's potency and efficacy, were analyzed. Lower values indicate higher potency, making them essential criteria for selecting promising antiviral compounds.

The data were extracted and processed, mainly by labeling the compounds based on their IC50 values, classifying those with an IC50 less than 10.000 nM as "Active" and those with higher values as "Not Active." This allowed to efficiently identify and classify the antiviral compounds for further investigation.

To find the molecular descriptors of this compounds the ChemDes [19] online tool was used. In the platform the descriptors were computed with RDKit [20] which is an open-source cheminformatics library that offers a comprehensive suite of tools for processing and analyzing chemical data.

The input was the Simplified Molecular Input Line Entry System (SMILES) of the compounds that is a text-based notation for representing chemical structures using ASCII characters. Once the molecular descriptors were calculated the results were downloaded as a CSV file.

The code read the data from that CSV file, handled missing values by filling them with the most frequent value of each column, and classified data based on a threshold value of 10.000 that it splits the data into 'Active' and 'Not Active' categories. The code further dropped irrelevant columns and applied a correlation cleaner function to remove highly correlated features. The correlation cleaner iteratively removed the most correlated feature until no pair exceeded a specified correlation cutoff of 0.9. This resulted in a cleaner dataset that is well-prepared for subsequent analysis.

To ensure a robust evaluation, the dataset was split into training and testing sets with 80% of the data for training and 20% for testing. The random state was reproducible so its time the code was executed the same split and the exact same data points remained. This was particularly useful for debugging and for ensuring that results can be reliably compared across different runs of the experiment.

The training data were then used to select molecular descriptors according to their Fishers Score. This method aims to identify descriptors that best separate different classes in a dataset. The primary goal is to select those that maximize the distance between classes while minimizing the variance within each class, making it easier for classifiers to distinguish between different categories.

The Fisher Score for a particular descriptor is calculated:

$$\text{Fisher Score}(f_i) = \frac{\sum_{c=1}^{C} N_c(\mu_{c,i} - \mu_i)^2}{\sum_{c=1}^{C} N_c\sigma_{c,i}^2}$$

- C is the number of classes.
- N_c is the number of samples in class c.
- $\mu_{c,i}$ is the mean of the feature f_i for class c.
- μ_i is the overall mean of the feature f_i across all classes.
- $\sigma_{c,i}^2$ is the variance of the feature f_i within class c.

The numerator of the equation calculates the between-class variance. It measures how much the feature varies between different classes. Specifically, it sums up the weighted squared differences between the mean of the feature in each class and the overall mean. A larger value indicates that the feature can differentiate well between classes.

The denominator represents the within-class variance. It sums the variances of the feature within each class, weighted by the amount of samples in the class. A smaller value here suggests that the feature has less variability within each class, which is desirable because it implies that the feature is consistent within each class.

As a result, a high Fishers Score indicates that the feature has a large between-class variance and a small within-class variance, making it a good feature for distinguishing between classes [21].

The top 50 features with the higher score were retained for training and testing a Random Forest classifier. Training means learning patterns between the target labels and the features and then this model is utilized to predict the labels of the test dataset. This step assesses how effectively the model generalizes to new data.

Random Forest is an advanced machine learning algorithm that builds on the concept of decision trees. A decision tree resembles a flowchart, where every internal node symbolizes a decision according to a particular feature of the data, branches represent the results from these decisions, and leaf nodes indicate the final prediction or classification. The tree starts with a root node that contains the entire dataset and splits it based on the feature that best separates the data, according to methods like Gini impurity. This process of splitting continues recursively, creating a structure where every path from the root to a leaf represents a sequence of decisions leading to a final outcome [22].

Gini impurity is an automated method to assess the quality of a split at a node. It calculates the probability that a randomly chosen feature from the node will be incorrectly classified if it were labeled at random based on the current class distribution in the node. By calculating the Gini impurity, the classifier can identify which features are better suited for placement higher up in the tree, starting from the root and progressing through the internal nodes.

During the construction of a decision tree, Gini impurity helps decide which feature to split on by evaluating how each potential split affects the impurity of the resulting child nodes. The feature that leads to the highest reduction in Gini impurity from the parent node to the child nodes is chosen for the split.

The mathematical formula for the Gini impurity is given by:

$$\text{Gini} = 1 - \sum_{i=1}^{n} p_i^2$$

- pi represents the probability of a sample being classified into a specific class i
- n is the number of classes [23]

A Random Forest classifier improve the overall predictive performance by producing an ensemble of multiple decision trees. Each tree in this forest is trained on a various subset of the data, generated through a process called bootstrapping, where the original dataset is randomly sampled with replacement. This implies that some data points may be chosen multiple times, while others might not be selected at all, ensuring diversity in the training data for each tree.

Additionally, the random subspace method is used. In a standard decision tree, when making a split at any node, the algorithm considers all available features in the dataset to determine the best possible split. This approach, while effective, can lead to overfitting, especially if some features are particularly dominant. If these dominant features are always chosen first, the resulting trees can become highly correlated, meaning they make similar mistakes and fail to generalize well to new data.

To mitigate this issue, Random Forest introduces a layer of randomness during the tree-building process. At each node in a decision tree, instead of evaluating all features, Random Forest randomly selects a subset of features. Only these randomly chosen features are then considered for splitting the node.

Once the individual trees are built, the Random Forest classifier aggregates their predictions to make a final decision. For classification tasks, this is usually accomplished through majority voting, where the class that receives the highest votes from the trees is selected as the final prediction. For regression tasks, the predictions from all the trees are averaged.

The Random Forest classifier was selected for both feature selection and later on for classification due to its unique combination of advantages. It can handle high-dimensional data well, meaning it can manage datasets with a large number of features without getting overwhelmed.

For classification, its ensemble approach, where multiple decision trees are trained on different subsets of the data, makes it highly robust and less prone to overfitting that happens when a model is too closely tailored to the training data. By averaging the predictions of these diverse trees, Random Forest produces a model that generalizes well to new data.

It also excels in capturing complex, non-linear relationships between features, which many other classifiers might miss. Moreover, Random Forest is relatively immune to issues like multicollinearity, that occurs when two or more features in a dataset are highly correlated. This boosts its versatility and reliability for both feature selection and classification tasks [24].

To understand which descriptors contributes most to the model's predictions, the permutation importance of each feature was calculated. This involved randomly shuffling the values of each feature and assessing the resulting decrease in model accuracy. Features that resulted in a substantial drop in accuracy were considered more valuable. The features were then ranked according to their importance scores, and the top 10 possibly most important features were identified and extracted from the original dataset. These features were used for further analysis.

3.1.2 Computational chemistry

In our study, computational chemistry methods were utilized to evaluate whether a series of natural compounds could interact with the NS2B-NS3 protease of the Zika virus in a similar way to its bound ligand, which acts as a potential inhibitor, thereby assessing their potential as antiviral drugs.

The analysis began by identifying from the Protein Data Bank (PDB) [25] a suitable co-crystallized protein and ligand complex. PDB is a database that stores three-dimensional (3D) structural data of biological molecules, such as nucleic acids, proteins and complex assemblies. These structures are produced using experimental methods like X-ray crystallography.

The selected structure was 5LC0 that represents the NS2B-NS3 protease of the Zika virus with a boronate inhibitor bound to its active site. A boronate inhibitor is a type of chemical compound that contains a boronic acid or boronate group and is designed to inhibit the activity of specific enzymes, particularly targeting serine or cysteine proteases. The International Union of Pure and Applied Chemistry (IUPAC) name of the compound used in this structure is N-((S)-3-(4-(aminomethyl) phenyl)-1-(((R)-4-guanidino1-(5-hydroxy-1,3,2-dioxaborinan-2yl) butyl) amino)-1-oxopropan-2-yl) benzamide (6T8).

In the crystal structure of the ZIKV NS2B/NS3 protease, the two chains, labeled A and B, are identical. Each chain is made up of the NS3 protease and the NS2B cofactor, forming a special kind of pair known as a homodimer. This homodimer exhibits quasi-twofold symmetry, where the two identical parts are arranged almost like mirror images of each other. This specific arrangement in the dimer, which is unusual for flavivirus proteases, could be essential for the virus's ability to replicate within the host [16].

5LC0 is considered a good structure for studying the Zika virus for several reasons. It is highly relevant as it represents the NS2B-NS3 protease complex of the Zika virus with a bound inhibitor, which provides an example of how the potential inhibitors interact into the active site of the examined complex, making it a valuable template for designing new drugs. The resolution of the structure is good for crystallographic data, providing accurate details of the atomic positions, which is crucial for reliable computational chemistry procedures.

Finally, it has a well-defined view of the active site and it has been widely used in research and published studies so it is validated as a reliable structure. Once the structure was selected, it was inspected and cleaned by removing non-essential elements like water and other small molecules or additional chains that are not part of the target complex. Following this, proper protonation was achieved by adding hydrogen atoms, enhancing the accuracy of simulations and improving the understanding of the protein's behavior and function.

Understanding the interactions between a protein and its complexed ligand is essential for comprehending how enzyme inhibition occurs. To identify these interactions the Protein-Ligand Interaction Profiler (PLIP) [26] software was used. It examines 3D structures, from PDB, of protein and ligand complexes and finds non-covalent interactions such as hydrophobic interactions, hydrogen bonds, π-stacking, salt bridges and π-cation.

Hydrophobic interactions occur when non-polar molecules or parts of molecules group together to avoid contact with water. When two flat, ring-shaped molecules, such as aromatic rings, stack on top of each other because of the special type of electrons they have, called π-electrons, π-stacking happens. These π-electrons, which are found in π-bonds, a type of bond where electrons are spread above and below the atoms, rather than directly between them, create a weak attraction between the rings.

This attraction occurs because the electrons in one ring are drawn to the electron cloud of another ring. As a result, the stacking stabilizes the structure of the molecules. On the other hand, π-cation interactions occur when a positively charged ion, called a cation, is attracted to the π-electrons in an aromatic ring. Since the π-electrons form a negatively charged cloud above and below the ring, the positively charged cation is drawn to it, creating a strong attraction [27].

Based on the identified interactions that were mostly hydrogen bonds, pharmacophore models were created, using Pharmit [28], which is a web-based tool, that screens large libraries of compounds to predict which ones may have the necessary features to interact effectively with a biological target. The 5LC0 PDB was used as input so the 3D coordinates of the protein and the ligand were extracted. After identifying the binding site, a pharmacophore model was generated automatically, the features of whom, were altered to adjust to the search being conducted.

It was decided that two models would be constructed, the first includes six features, five hydrogen bond donors and one hydrogen bond acceptor, while the second with three hydrogen bond donors. The choice to include two models derived because the one with more features would have a higher specificity, meaning that it would likely identify ligands that closely match the binding interactions of the known ligand. It also reduced the chance of non-relevant compounds being identified as potential hits. The second one with fewer features, would be more flexible, potentially allowing for a broader range of structurally diverse compounds to be identified as hits.

To select this type of features the focus was on two key criteria, the hydrogen bonds and the Root Mean Square Deviation (RMSD) score that was produced following the screening with the ZINC compound database [29].

Prioritizing hydrogen bonds allowed to target compounds that form strong interactions with the enzyme's active site, as these bonds are vital for stabilizing the inhibitor and they were also the most common interaction found. To further refine the selection process, RMSD was used as an additional filter. This is a measure used to quantify the differences between two molecular structures.

Specifically, in drug discovery, how much the predicted binding pose of a compound (inhibitor) differs from a reference pose (pose of the inhibitor in a crystal structure). Low RMSD indicates that the inhibitor binds in a way that is very similar to the reference pose. This suggests that the pharmacophore features and docking predictions are accurate. The inhibitor also has a better chance of fitting into the enzyme's active site and be more stable in it.

If the predicted binding pose has a significant deviation from the reference, suggesting that the RMSD is higher, it might mean that the inhibitor will not bind as well or could bind in a way that is less effective. Compounds with the lower RMSD were the ones considered as potential candidates for experimental validation.

Furthermore, a selection of 2.000 natural compounds sourced from the ZINC library was obtained. Natural compounds were chosen because they are often biocompatible, as they are derived from natural sources that organisms have been exposed to over evolutionary time. They can interact with specific biological targets with good precision minimizing unintended interactions with other biological systems.

These reduce the likelihood of triggering immune responses or causing other adverse effects making them less toxic compared to synthetic compounds. It is worth mentioning that they also have a long history of use in traditional medicine further proving their safety. These benefits make natural compounds particularly desirable in drug development [30].

The ZINC library was selected in both cases due to its comprehensive database of commercially available compounds, which includes a large and diverse collection of natural products. The fact that these are commercially available means that they are well-documented and ready for experimental validation.

These compounds are also carefully validated to meet specific criteria and come in formats directly compatible with virtual screening tools and computational chemistry software, eliminating the need for extensive preprocessing. Additionally, it is widely recognized and used by the scientific community having been referenced in numerous studies. All the above make it a reliable choice.

In conclusion the final dataset comprised of 2.200 compounds. A total of one hundred compounds were selected from the pharmacophore model with the highest feature count, each with an RMSD score ranging from 0.319 to 0.604. Similarly, one hundred compounds were extracted from the second pharmacophore model, with RMSD values between 0.028 and 0.140. Along with the 2.000 compounds directly derived from the library, they all were subsequently utilized for molecular docking experiments.

3.2 Characteristics

Table 3.1: Description of features that have originated from the results of statistical analysis and machine learning methodologies [31]

3.3 Objectives & implementation

3.3.1 Objectives & implementation of statistical analysis methodology

	NAME	DESCRIPTION	CATEGORY
1	MinEStateIndex	Returns a tuple of EState indices for the molecule	Topological
2	MaxAbsEStateIndex	Returns a tuple of EState indices for the molecule	Topological
3	MinAbsEStateIndex	Returns a tuple of EState indices for the molecule	Topological
4	Qed	It stands for quantitative estimation of drug-likeness and it reflects the underlying distribution of molecular properties including molecular weight, topological polar surface area, number of hydrogen bond donors and acceptors, the number of aromatic rings and rotatable bonds, and the presence of unwanted chemical functionalities	Physicochemical
5	MaxAbsPartialCharge	Returns molecular charge descriptors	Topological
6	MinAbsPartialCharge	Returns molecular charge descriptors	Topological
7	FpDensityMorgan3	Morgan fingerprint density	Connectivity
8	BCUT2D_MWLOW	Lowest eigenvalue weighted by atomic masses It contains information on molecular size and topology	Topological
9	BCUT2D_CHGHI	Highest eigenvalue weighted by gasteiger charges	Topological

In the process of statistical analysis, it was necessary to find the features that had statistically significant differences between the two categories, active or non-active compounds. To achieve this, the appropriate statistical test had to be implemented. The options were the t-test (Student's t-test) and the rank sum test (Wilcoxon rank sum test or Mann-Whitney U test).

The first test requires the samples to have a normal distribution, meaning there should be homogeneity of variance between the two samples, and it is generally more sensitive to extreme values. The second type of test does not require a known distribution and is more robust against extreme values, which is why it was preferred.

The rank sum test is a non-parametric test that uses the mean rank for two independent samples. More specifically, the observations of the two categories are placed one after the other in a column in ascending order. The next step is to assign a rank to each observation, with the lowest taking the first rank and the next taking the second, and so on. In the case of ties between observations, the average rank is calculated. Then, the U parameter needs to be found for each category.

$$U_1 = n_1 \times n_2 + \frac{n_1 \times (n_1+1)}{2} - T_1$$

$$U_2 = n_1 \times n_2 + \frac{n_2 \times (n_2+1)}{2} - T_2$$

The n1 represents the number of observations belonging to the first category, and n2 similarly represents the number of observations belonging to the second category. T1 and T2 are the sum of the ranks for each category. From U1 and U2, the one with the smaller value is chosen. Using U, along with n1 and n2, the p-value is calculated using various methods. If the p-value is below a predefined level of significance, then the null hypothesis is rejected, meaning there is a statistically significant difference between the two categories [32].

As previously mentioned, the goal is to find the features with statistically significant differences, as these are more likely to provide the best discriminative ability between the two categories. In this study, this was achieved with an algorithm using Python, where the test was performed on each of the 10 potentially best features. Each feature was divided into two categories depending on whether the compounds were active or not, and for each, the p-value was calculated. This was compared to a significance level of 0.001, and if the p-value was smaller, then the feature showed a statistically significant difference. The distributions of the values of the two categories for each feature that showed a statistically significant difference were depicted in box plots.

Box plots are a graphical tool that summarize data distribution, making them useful for comparing a feature across two different categories (Figure 4.1). The box represents the interquartile range (IQR), which includes the central 50% of the data between the first quartile (Q1) and the third quartile (Q3). The lower edge of the box corresponds to Q1, while the upper edge corresponds to Q3, with the line inside marking the median (Q2), or central value.

This box effectively illustrates where most of the values are, indicating how concentrated or dispersed they are around the median. The whiskers extend from the box to the minimum and maximum values within 1.5 times the IQR, while any data points beyond the whiskers are regarded as outliers. representing extreme values. When the medians are clearly separated and the boxes (IQRs) show minimal overlap, there is strong visual evidence of a statistical difference between the two categories.

Moreover, having fewer outliers suggests that the data is more consistent and follows a predictable pattern, with fewer extreme deviations. This enhances the reliability of the analysis, making it easier to detect true patterns, draw accurate conclusions, and avoid misleading interpretations [23].

3.3.2 Objectives & implementation of machine learning methodology

A classifier is a machine learning model that is trained through a dataset to learn patterns in the data associated with each category. Once trained, it can predict the category of new data. In this study, to create the model that would predict as accurate as possible if a compound, based on its features, is active or non-active many classifiers were tested. These were the Nearest Centroid, k-Nearest Neighbors, Linear Discriminant Analysis, Gaussian Naive Bayes, Logistic Regression, Perceptron, Multilayer Perceptron, Random Forest, linear Support Vector Machine and Decision Tree Classifier.

The Nearest Centroid Classifier calculates the distance of a data point to the center of each class and assigns it to the closest class and it works best for data that are well-separated. The k-Nearest Neighbors (KNN) assigns a class to a data point based on the majority class among its k nearest in distance points (neighbors) and its most effective for small datasets.

The Naive Bayes classifier works by calculating the probability of each class according to the features of the existing data. When a new compound is introduced, it uses the same method, which is based on Bayes' Theorem, to calculate the probability of that compound belonging to each class, considering the compound's features. It then classifies the new compound into the class with the biggest probability. It is called "naive" because it insinuates that the features of the data are not dependent of each other making it fast for problems where this applies.

The Linear Discriminant Analysis (LDA) finds a straight line that best separates classes in the data. It assumes each class follows a normal distribution and calculates both the class means and the spread (variance) of data within each class. LDA aims to maximize the distance between the class means while minimizing the variance within each class.

By combining these factors, LDA determines an optimal line for classification, where new data points are classified according to which side of the line they are. This approach works well for linearly separable problems.

The Logistic Regression applies a mathematical function called the logistic or sigmoid function to produce a value between zero and one. This value indicates the likelihood that a given input belongs to a specific class. If the probability is greater than 0.5, the data point is more likely to belong to one class, whereas a probability below 0.5 suggests it belongs to the other class. This is especially useful when it is important to know not only the predicted class but also the confidence level of that prediction.

The Perceptron is the simplest type of neural network, which is a computational model inspired by the function of the human brain. It takes multiple input features and assigns each a weight. These weights represent how much influence each feature has in determining the outcome. It then computes a weighted sum of all the inputs, which means it multiplies each input by its weight and adds them together. After that, it applies a threshold function to decide whether the data point belongs to one class or the other. During training, the Perceptron continuously adjusts its weights by learning from the errors it makes. The ultimate goal is for the Perceptron to identify a linear decision boundary that divides the two classes with as much accuracy as possible.

The Multilayer Perceptron (MLP) is an extension of the basic perceptron and it can handle non-linear relationships between data. It consists of multiple layers which are the input layer, one or more hidden layers, and the output layer. Each node in the hidden layers is connected to nodes in the next layer through weights. These hidden layers allow the MLP to learn more complex patterns and solve problems where a simple straight line isn't enough to separate the classes.

The linear Support Vector Machine (SVM) aims to separate data into different classes using a straight line or a plane in higher dimensions. It tries to identify the line that maximizes the distance, called the margin, between itself and the closest points from each class. These closest points are called support vectors. A larger margin means there’s more space between the classes, which generally leads to better classification results. SVM is mostly used for high-dimensional data [24].

The classifiers that produced the best results were the KNN, Decision Tree, which builds a single tree based on the training data to make classifications, and the Random Forest classifier. While all three classifiers performed well, the Random Forest was ultimately selected due to its higher classification accuracy, as well as all the advantages that were mentioned before.

In the dataset the number of active and not active compounds of the 10 possibly most important descriptors were 296. The Random Forest classifier that was used for classification demands a significant amount of data to effectively train multiple decision trees and capture the underlying patterns in the data. A larger dataset as a result enables the classifier to better generalize by reducing variance, avoiding overfitting, and improving the model's predictive accuracy.

To achieve that resampling was utilized. It is a statistical technique used to create additional samples from a limited dataset, particularly in situations where the original data is insufficient. Using this method the 296 compounds increased to 1.000 entries.

To the resampled data noise was added, which was random values that were generated to follow a normal (Gaussian) distribution, which is a bell-shaped curve where the majority of the values are concentrated near the mean, and fewer values appear as you move towards the tails of the distribution, representing more extreme deviations.

In the tests the mean was zero to ensure that the noise is centered around the original data points, while the standard deviation was 0.5 to control the spread of the noise, introducing slight variations without drastically altering the data. This can prevent the model from simply memorizing the specific details of the training samples and encourages the model to learn more generalized patterns, making it more robust and better equipped to handle new, unseen data.

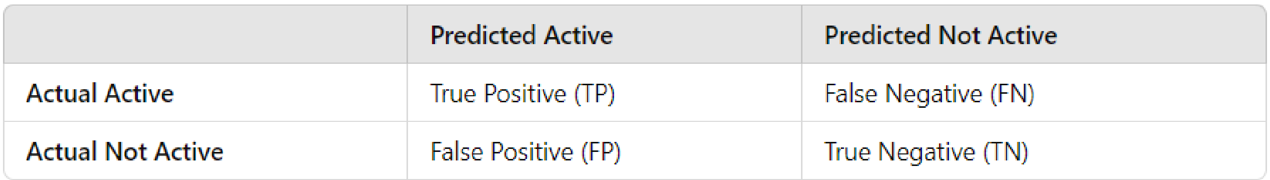
Synthetic Minority Over-sampling Technique (SMOTE) was applied to address class imbalance by generating synthetic examples for the minority class. For each data point in the minority class, it identifies a certain number of nearest in distance neighbors. It selects one of these nearest neighbors at random. It then generates a new data point (a synthetic example) by taking a point somewhere along the line segment that connects the original data point and the selected neighbor.

This means the new data point will have values that are a weighted average of the original and the neighbor. The specific location on the line is chosen randomly, so each synthetic example is slightly different. This process is repeated for many data points in the minority class, generating as many synthetic examples as needed to balance the class distribution. This leads to better decision-making when the model is trained, as it no longer heavily favors the majority class.

After applying SMOTE, normalization was performed to scale the data, ensuring that all features contribute equally to the model's learning process. This happened by adjusting the values of the features, so they lie within a specific range, typically between 0 and 1, preventing features with larger numerical differences from dominating the model’s predictions.

Then, an exhaustive search method was employed to identify the optimal combination of features so the classification model achieved the highest possible accuracy. This technique explores every potential subset of features, from individual features to combinations that include the entire set, leaving no possible feature interaction unchecked. For each subset, the model was trained, and its performance was evaluated using the bootstrap method. The bootstrap approach involves creating multiple datasets by randomly drawing samples with replacement from the original data. For each bootstrap sample, confusion matrices were generated.

Table 3.2: Confusion matrix



	Predicted Active	Predicted Not Active
Actual Active	True Positive (TP)	False Negative (FN)
Actual Not Active	False Positive (FP)	True Negative (TN)

- True Positives (TP): The number of compounds correctly predicted as active
- False Positives (FP): The number of compounds incorrectly predicted as "active" when they are actually not active
- True Negatives (TN): The number of compounds correctly predicted as not active
- False Negatives (FN): The number of compounds incorrectly predicted as not active when they are actually active

Additionally, critical evaluation metrics such as sensitivity (the model’s ability to correctly identify active compounds), specificity (the ability to correctly identify not active compounds), and overall accuracy were computed.

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$

$$\text{Specificity} = \frac{TN}{TN + FP}$$

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN}$$

This combination of exhaustive search and bootstrap evaluation provided a thorough and reliable assessment of the model's performance, ensuring that the most effective subset of features was identified while maintaining confidence in the model's ability to generalize across different samples.

Receiver Operating Characteristic (ROC) curves visualize how well a classification model performs by showing the relationship between the True Positive Rate (TPR) and the False Positive Rate (FPR). The TPR, also known as sensitivity, indicates how many actual positive cases are correctly identified compared to the FPR that represents how many negative cases incorrectly classified as positive.

$$\text{FPR} = \frac{\text{False Positives}}{\text{False Positives} + \text{True Negatives}}$$

The Area Under the Curve (AUC) quantifies the overall ability of the model to distinguish between the active and inactive classes. Value of 1 indicates perfect classification, 0.5 signifies that the model is no better than random guessing, and values less than 0.5 suggest that the model performs worse than random guessing. A curve that is steep and closer to the top-left corner with a big AUC score and a accuracy greater than 90% signifies a model with strong discriminatory ability. This indicates a high true positive rate with a low false positive rate [33].

3.3.3 Objectives & implementation of computational chemistry methodology

In the molecular docking experiments, the aim is to identify compounds that could potentially inhibit the Zika virus protease complex NS2B-NS3. To accomplish that the softwares Webina [34] and Maestro [35] were utilized.

Webina is a web-based platform that integrates AutoDock Vina, to predict how small compounds fit into a protein's binding site. This aids researchers in assessing the strength of the interaction between the ligand and the receptor. Webina requires three files as input. One that will contain only the receptor, another with the known complexed ligand, and a third one with the compound being tested in conjunction with the receptor.

To determine the correct pose for each test compound, PyRX [36], an open-source virtual screening software was used. Two main criteria were applied. Each compound had to possess a low-energy (stable) conformation and ideally exhibit an orientation similar to that of the known ligand, especially since it has already been established that it is an effective binder to the target site. Aligning the orientation ensures correct interactions but minor variations are allowed if the main interactions are maintained.

For every compound combined with the receptor to be able to accurately predict the binding affinity between them it is essential and can be accomplished by adding hydrogen atoms at a specific pH. If a ligand is not in its proper protonation state, it may not fit optimally in the binding site or may fail to interact effectively with crucial residues, leading to inaccurate docking results. The optimal pH was determined through the BRENDA database [37], a comprehensive resource for enzyme information. The Zika virus has a pH range of 7.4–8.5 in which the enzyme is most stable, and a pH of 7.5 was selected for the experiments.

Moreover, to create a custom grid box in Webina is critical as it defines the specific area of the protein where the docking analysis will be conducted. This creates a three-dimensional box around the protein's active site, specifying the exact region where the ligand will be tested for binding. By setting the grid box, researchers can narrow the docking search to the most relevant part of the protein, ensuring that calculations focus on the target area rather than the entire protein structure. Properly defining the grid box size and position is essential for obtaining reliable docking results.

Through the docking process in Webina the binding affinity is produced, which is an indication of the strength of the interaction between the ligand and the receptor. A more negative binding affinity value suggests a stronger interaction, implying that the compound is more likely to bind effectively to the target protein. It generates multiple possible orientations (poses) of the ligand within the binding site, ranked by their predicted binding affinities. The top-ranked pose is considered the most likely interaction of the compound with the protein.

Maestro, developed by Schrödinger, is a commercial software that provides advanced molecular docking, visualization, and analysis tools. It enables the simultaneous screening of a library of compounds against a target protein, evaluating their binding affinity and potential poses in a single run. This makes it suitable for studies involving a large number of compounds, enabling quick identification of potential inhibitors. It offers two docking modes the Glide XP (Extra Precision) and the Glide SP (Standard Precision). Glide SP provides a balance between speed and accuracy in contrast to XP which is more stringent, delivering precise and detailed results by applying additional scoring functions.

The increased strictness of Maestro compared to Webina arises from its advanced scoring functions, algorithms, and docking protocols, particularly in Glide XP mode. This results in more accurate and reliable docking outcomes than those provided by Webina, which utilizes the simpler AutoDock Vina scoring function. This led to the initial use of Webina to filter out compounds, retaining only the most suitable candidates. These selected compounds were then subjected to further analysis in Maestro's two docking modes to identify the top compounds with greater precision.

Initially, internal validation was performed, by re-docking the original ligand back into the receptor, in both software programs, to estimate the binding affinity of the complex. This information was crucial, as it provided a reference range for evaluating how well the other compounds would bind in comparison.

Then to eliminate or not compounds, the criteria included achieving the highest absolute docking score close to that of the original complex, in addition to evaluating the number of interactions, especially hydrogen bonds, since they contribute most to stabilizing the connection of the test compounds with the receptor. Also important was for the interactions to match the interactions the original complexed ligand had with the protease complex. To examine the interactions PLIP was employed.

Ultimately, 10 compounds that met these criteria were selected from Webina and processed through Glide Sp and Glide Xp mode in Maestro to identify the top five potential inhibitors.

3.3.4 Objectives & implementation of combined methodologies

One method of verifying the results from the machine learning methodology, was the combination of these with the potential compounds identified from the molecular docking experiments. The machine learning results refer to the combination of features that provided the highest classification accuracy. The goal of this process is to categorize the compounds from the molecular docking experiments, using the machine learning features, as active.

To implement this endeavor, the MetaboAnalyst [38] software was used. The data that were uploaded are the original database containing the compounds from ChEMBL but with only the features that emerged from the machine learning process. The appropriate type of analysis was selected, which in this case was the classification analysis, and the algorithm used was the Random Forest, as it was also the one utilized in the classification process during the machine learning stage.

Additionally, data segmentation was implemented, where the entire dataset was split into two subsets. One of the sets was used for training the model and the other was a test set, which was used for evaluating the model's final performance. The data was split with 70% belonging to the training set and 30% to the test set.

Once the analysis was complete and the Random Forest classifier was trained, it was necessary, for the features of the compounds that were nominated from the molecular docking experiments to be found. This was achieved using the RDKit library. More specifically, an algorithm was created, that took as input the SMILES of the compounds.

These were converted into Mol objects, which are various molecular properties, using the Chem library. Then the features of the compounds were produced from the Mol objects. Only the ones that appeared in the machine learning results were extracted and subsequently used as input into MetaboAnalyst.

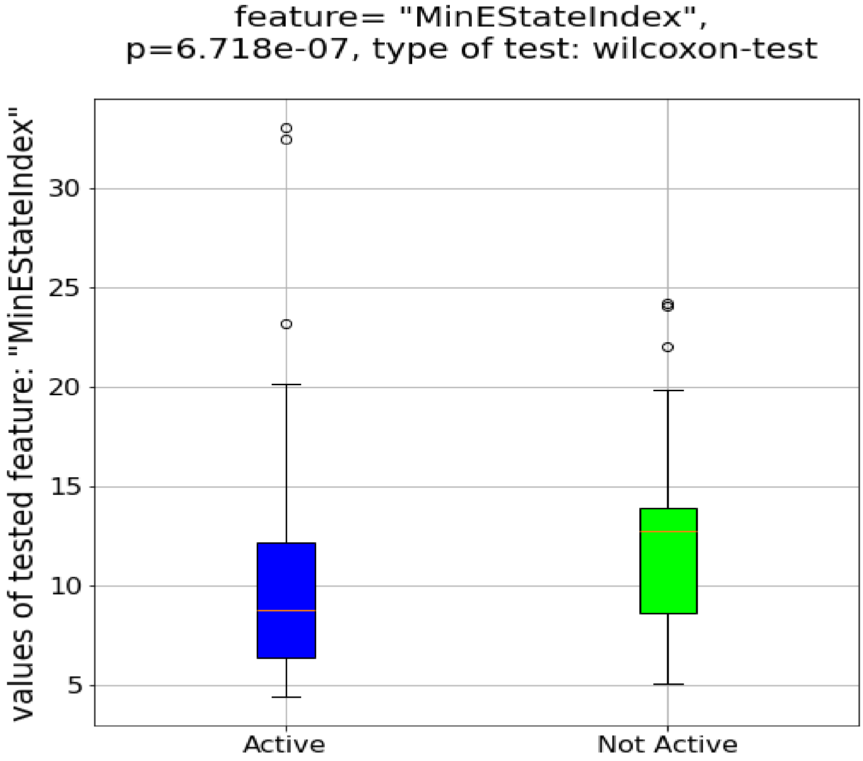
Finally, the trained Random Forest classifier was used to evaluate and categorize the compounds as active or inactive. The classification accuracy was depicted through a ROC curve, a scatter plot, and a box plot diagram.

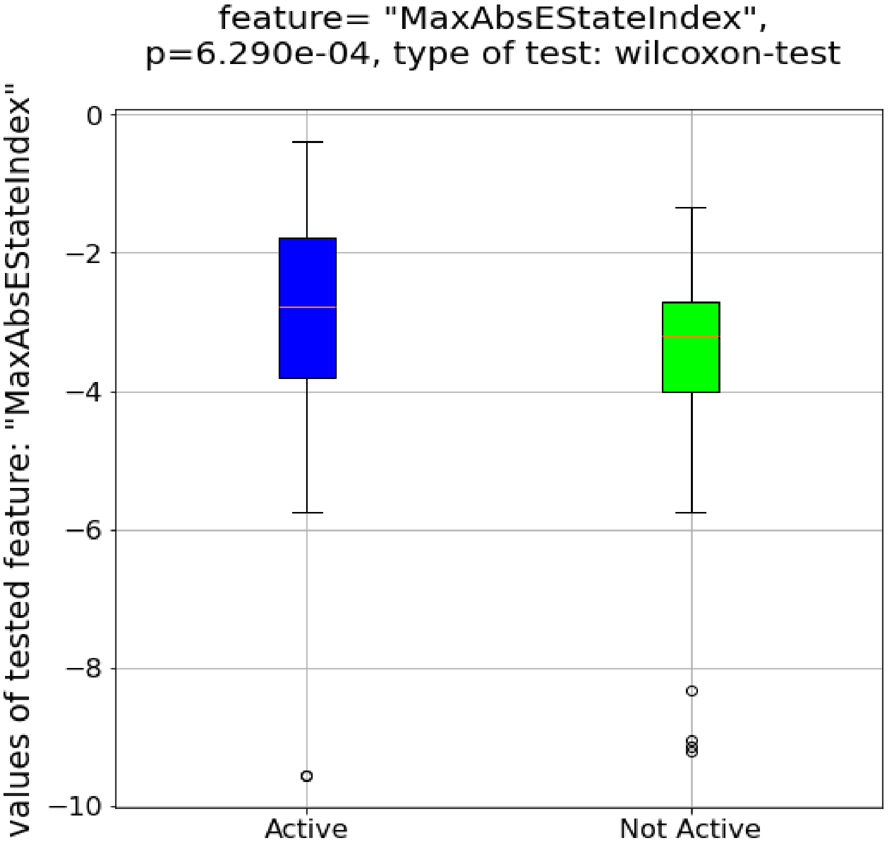
RESULTS

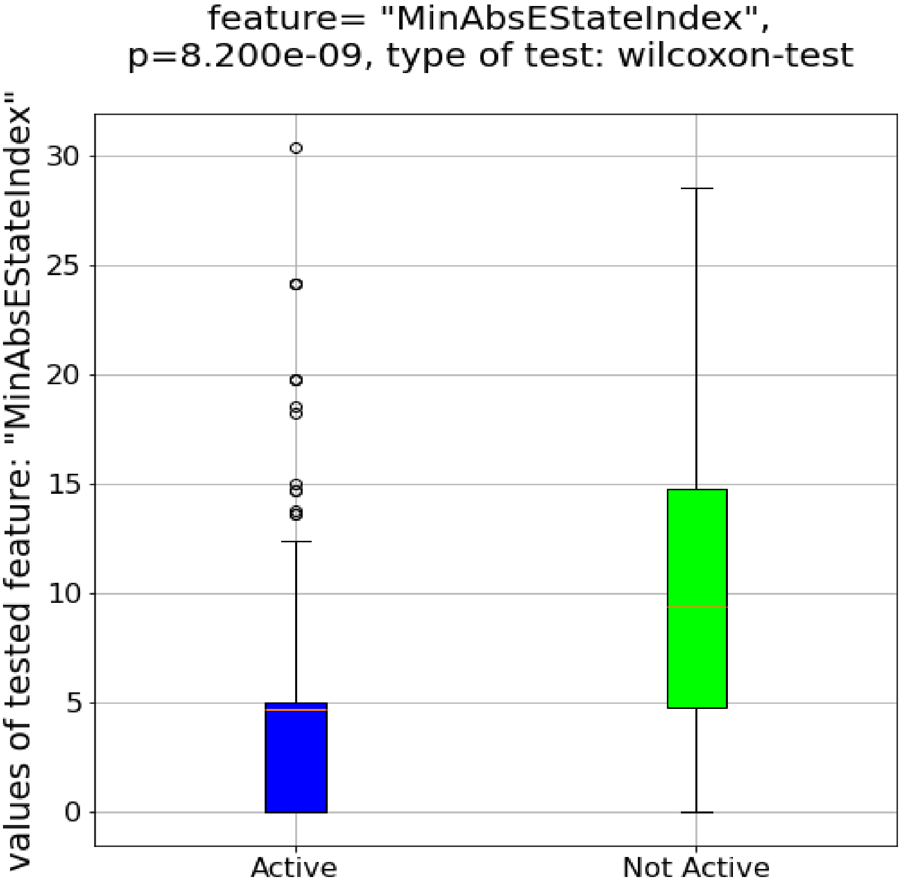
4.1 Results of statistical analysis

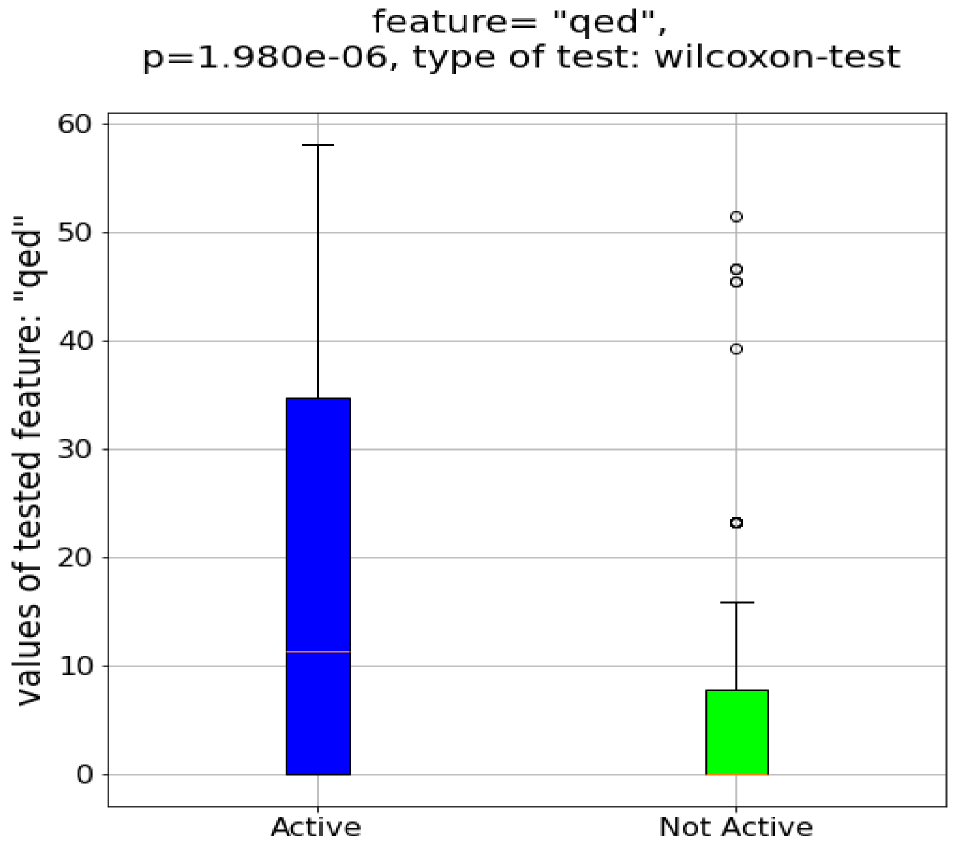
Table 4.1: Characteristics that show a statistically significant difference at a significance level of 0.001

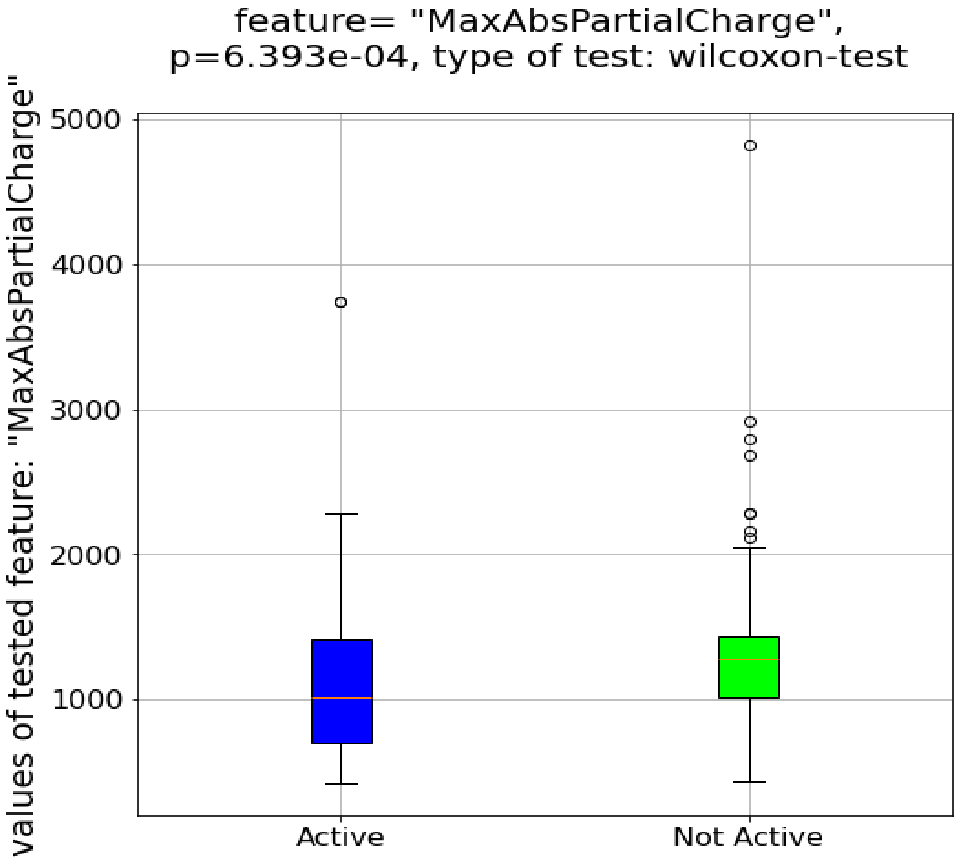
	Name of the Characteristic	p-value
1	MinEStateIndex	6.718e-07
2	MaxAbsEStateIndex	6.290e-04
3	MinAbsEStateIndex	8.200e-09
4	Qed	1.980e-06
5	MaxAbsPartialCharge	6.393e-04
6	MinAbsPartialCharge	4.701e-04
7	FpDensityMorgan3	2.133e-06
8	BCUT2D_MWLOW	9.514e-10

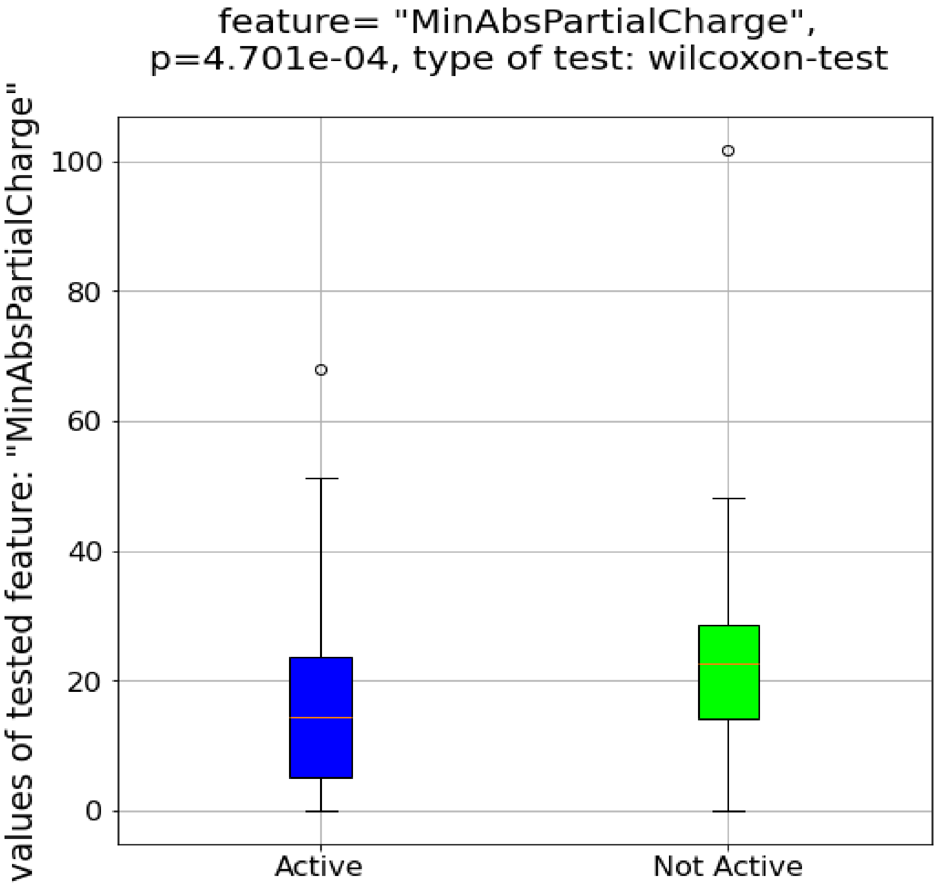


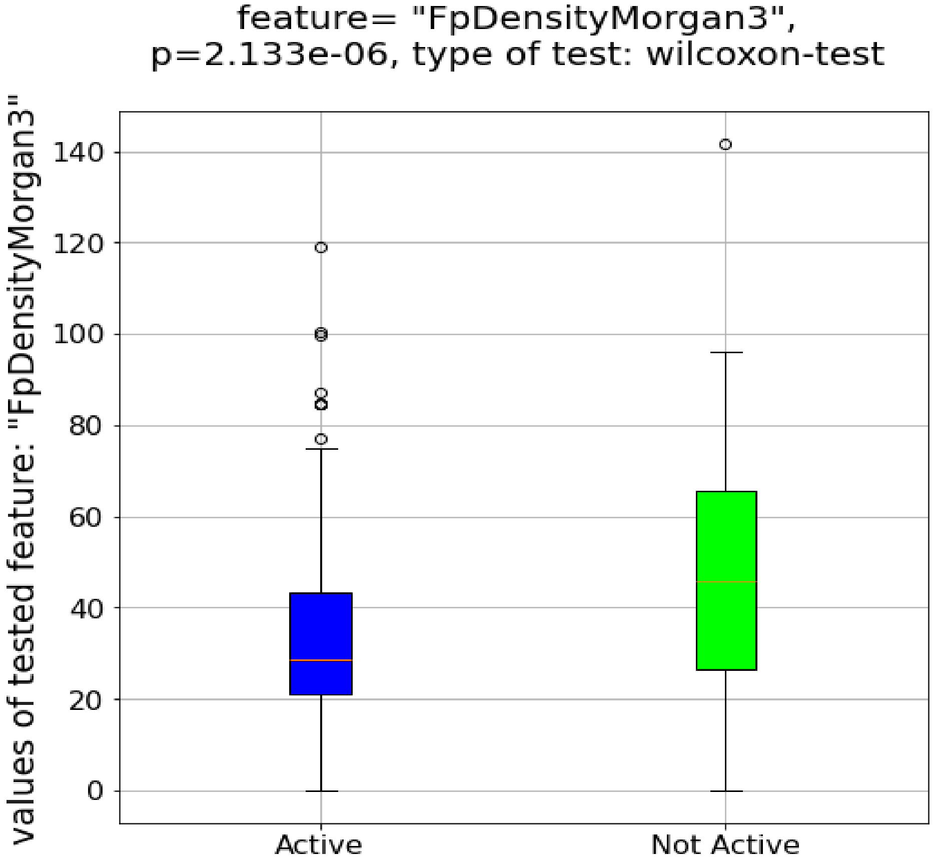

feature= "MaxAbsEStateIndex",
p=6.290e-04, type of test: wilcoxon-test
values of tested feature: "MaxAbsEStateIndex"
0
−2
−4
−6
−8
−10
Active
Not Active


feature= "MinAbsEStateIndex",
p=8.200e-09, type of test: wilcoxon-test
values of tested feature: "MinAbsEStateIndex"
30
25
20
15
10
5
0
Active
Not Active


feature= "qed",
p=1.980e-06, type of test: wilcoxon-test
values of tested feature: "qed"
0
10
20
30
40
50
60
Active
Not Active


feature= "MaxAbsPartialCharge",
p=6.393e-04, type of test: wilcoxon-test
values of tested feature: "MaxAbsPartialCharge"
1000
2000
3000
4000
5000
Active
Not Active


feature= "MinAbsPartialCharge",
p=4.701e-04, type of test: wilcoxon-test
values of tested feature: "MinAbsPartialCharge"
0
20
40
60
80
100
Active
Not Active


feature= "FpDensityMorgan3",
p=2.133e-06, type of test: wilcoxon-test
values of tested feature: "FpDensityMorgan3"
0
20
40
60
80
100
120
140
Active
Not Active

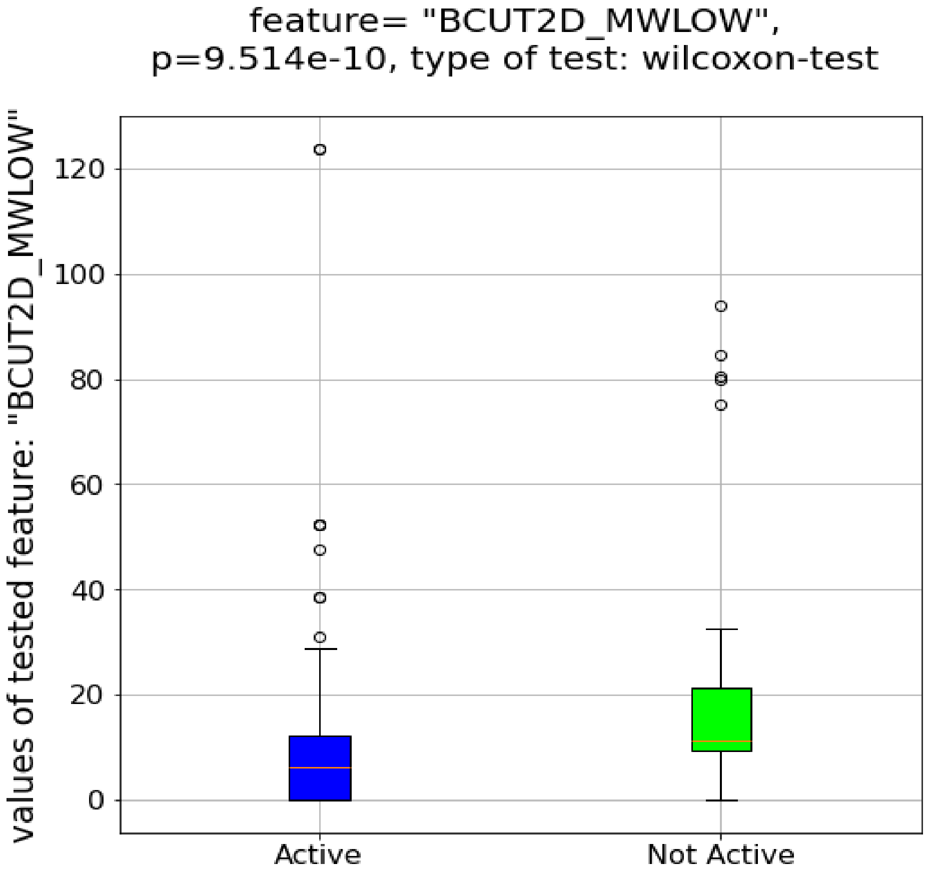


Figure 4.1: Box plots for the characteristics from table 4.1 for the active compounds and the not active compounds

4.2 Results of machine learning

Table 4.2: The classification with the best performance, according to the confusion matrix, was produced by the combination of the 7 characteristics

Optimal Feature Combination	Confusion matrix	Sensitivity	Specificity	Accuracy
BCUT2D_CHGHI MinAbsEStateIndex MinAbsPartialCharge BCUT2D_MWLOW FpDensityMorgan3 MaxAbsPartialCharge MinEStateIndex	[534 23] [27 536]	95.84%	95.10%	95.46%

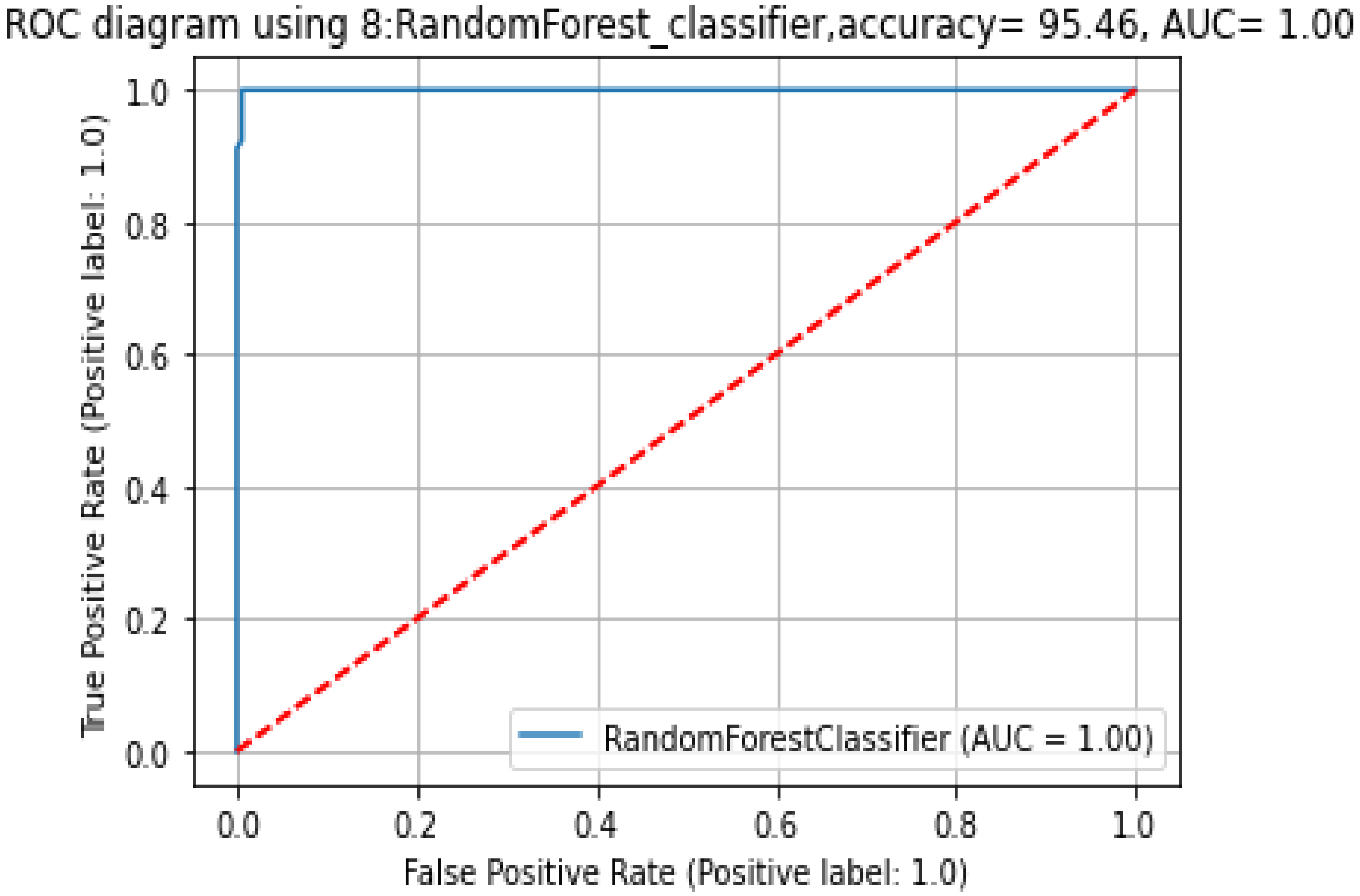


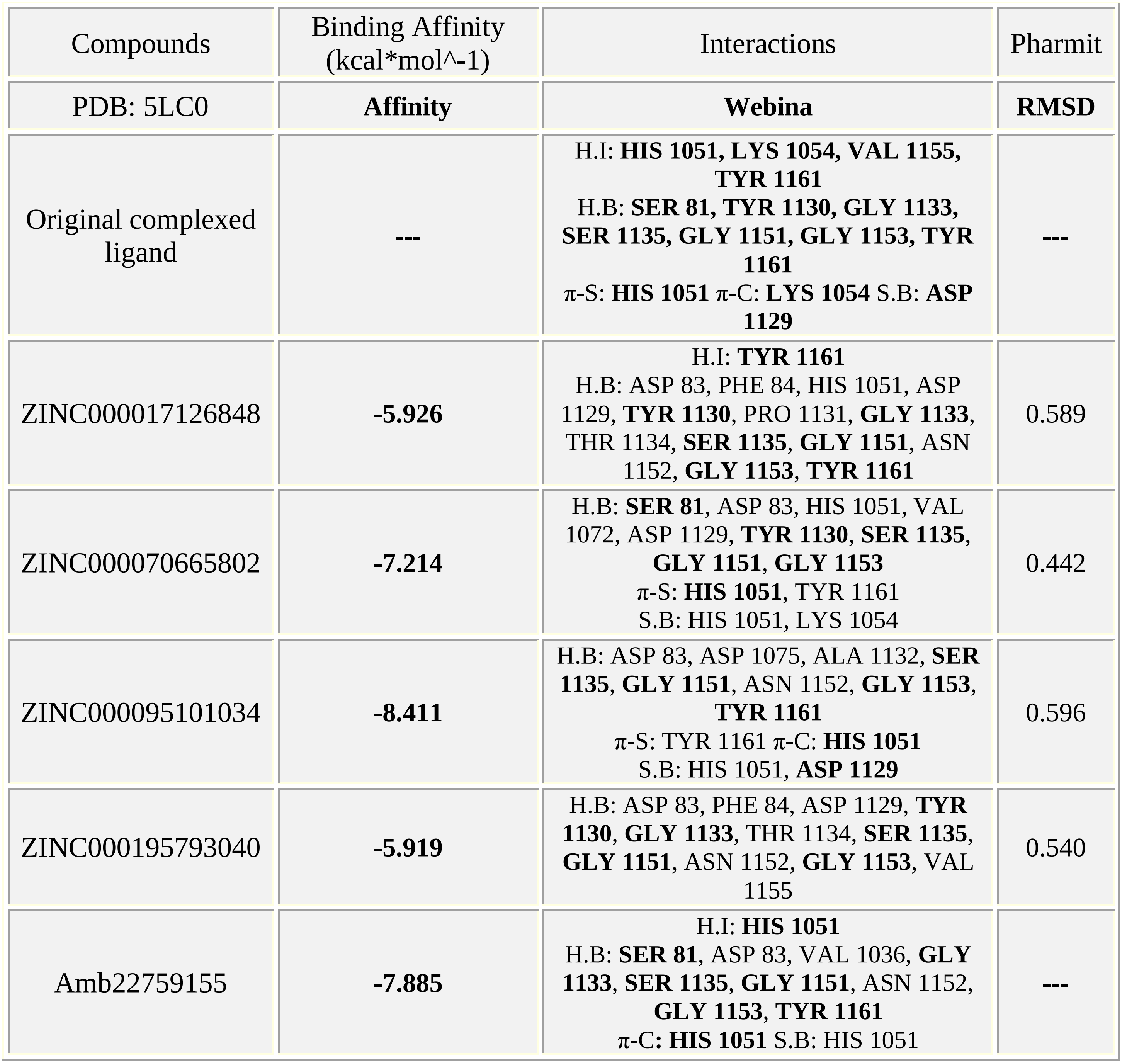
Figure 4.2: Receiver Operating Characteristic curve for the optimal combination of features with an accuracy of 95.46% and an AUC equal to 1

4.3 Results of computational chemistry

Table 4.3: Binding affinities and interactions of the original complexed ligand and the five potential inhibitors that were not selected with the virus protease complex, as identified through the Maestro program for both modes, Glide-SP and Glide-XP. With bold are the common interactions

Compounds	Binding affinity (kcal*mol^-1)		Interactions	
PDB: 5LC0	**Glide-SP**	**Glide-XP**	**Glide-SP**	**Glide-XP**
Original complexed ligand	**-5.265**	**-6.481**	H.I: **HIS 1051, LYS 1054, VAL 1155, TYR 1161** H.B: **SER 81, TYR 1130, GLY 1133, SER 1135, GLY 1151, GLY 1153, TYR 1161** π-S: **HIS 1051** π-C: **LYS 1054** S.B: **ASP 1129**	
ZINC000017126848	**-4.912**	**-9.615**	H.B: ASP 83, ASP 1129	H.B: **SER 81**, TRP 1050, LYS 1054, PRO 1131, **GLY 1151**
ZINC000070665802	**-7.581**	**-10.367**	H.B: **SER 81**, ASP 83, VAL1036, HIE 1051, PRO 1131, ASN 1152	H.B: ASP 83, VAL 1036, VAL 1072, **SER 1135**, **TYR 1161**
ZINC000095101034	**-5.108**	**-8.431**	H.B: VAL 1036, ASP 1129, **TYR 1130**	H.B: ASP 83, HIE 1051, VAL 1072, ASP 1075, ASN 1152
ZINC000195793040	**-5.393**	**-10.356**	H.B: ASP 83, VAL 1036, ASP 1075, ASP 1129, **TYR 1130**	H.B: **SER 81**, ASP 83, HIE 1051, LYS 1054, **TYR 1130**, PRO 1131 π-C: TYR 1161
Amb22759155	**-6.394**	**-10.373**	H.B: ASP 83, **TYR 1130**, PRO 1131, **GLY 1153**, **TYR 1161**	H.B: **TYR 1130**, PRO 1131, ASN 1152, **GLY 1153**, **TYR 1161**

Table 4.4: Binding affinities and interactions of the original complexed ligand and the five potential inhibitors that were not selected with the virus protease complex, as identified through the free access software Webina. With bold are the common interactions.



Compounds	Binding Affinity (kcal*mol^-1)	Interactions	Pharmit
PDB: 5LC0	**Affinity**	**Webina**	**RMSD**
Original complexed ligand	**---**	H.I: **HIS 1051, LYS 1054, VAL 1155, TYR 1161** H.B: **SER 81, TYR 1130, GLY 1133, SER 1135, GLY 1151, GLY 1153, TYR 1161** π-S: **HIS 1051** π-C: **LYS 1054** S.B: **ASP 1129**	**---**
ZINC000017126848	**-5.926**	H.I: **TYR 1161** H.B: ASP 83, PHE 84, HIS 1051, ASP 1129, **TYR 1130**, PRO 1131, **GLY 1133**, THR 1134, **SER 1135**, **GLY 1151**, ASN 1152, **GLY 1153**, **TYR 1161**	0.589
ZINC000070665802	**-7.214**	H.B: **SER 81**, ASP 83, HIS 1051, VAL 1072, ASP 1129, **TYR 1130**, **SER 1135**, **GLY 1151**, **GLY 1153** π-S: **HIS 1051**, TYR 1161 S.B: HIS 1051, LYS 1054	0.442
ZINC000095101034	**-8.411**	H.B: ASP 83, ASP 1075, ALA 1132, **SER 1135**, **GLY 1151**, ASN 1152, **GLY 1153**, **TYR 1161** π-S: TYR 1161 π-C: **HIS 1051** S.B: HIS 1051, **ASP 1129**	0.596
ZINC000195793040	**-5.919**	H.B: ASP 83, PHE 84, ASP 1129, **TYR 1130**, **GLY 1133**, THR 1134, **SER 1135**, **GLY 1151**, ASN 1152, **GLY 1153**, VAL 1155	0.540
Amb22759155	**-7.885**	H.I: **HIS 1051** H.B: **SER 81**, ASP 83, VAL 1036, **GLY 1133**, **SER 1135**, **GLY 1151**, ASN 1152, **GLY 1153**, **TYR 1161** π-C**: HIS 1051** S.B: HIS 1051	**---**

Table 4.5: Binding affinities and interactions of the original complexed ligand and the five potential inhibitors that were selected with the virus protease complex, as identified through the Maestro program for both modes, Glide-SP and Glide-XP. With bold are the common interactions

Compounds	Binding affinity (kcal*mol^-1)		Interactions	
PDB: 5LC0	**Glide-SP**	**Glide-XP**	**Glide-SP**	**Glide-XP**
Original complexed ligand	**-5.265**	**-6.481**	H.I: **HIS 1051, LYS 1054, VAL 1155, TYR 1161** H.B: **SER 81, TYR 1130, GLY 1133, SER 1135, GLY 1151, GLY 1153, TYR 1161** π-S: **HIS 1051** π-C: **LYS 1054** S.B: **ASP 1129**	
ZINC000013424720	**-4.635**	**-6.743**	H.B: ASP 1075, ASP 1129, **TYR 1130**, **GLY 1153**, **TYR 1161** π-C: **LYS 1054**	H.B: **SER 81**, ASP 83, LYS 1054, **TYR 1130**
ZINC000253389742	**-5.154**	**-7.083**	H.B: **SER 81**, **TYR 1130**, **GLY 1151**, **TYR 1161**	H.B: **SER 81**, **TYR 1130**, **GLY 1151**, **TYR 1161**
ZINC000253529689	**-5.201**	**-9.424**	H.B: ASP 83, LYS 1054, ASP 1129, **GLY 1153**	H.B: **SER 81**, LYS 1054, ASP 1129, **TYR 1130**, **GLY 1151**, **GLY 1153** π-S: TYR 1161
ZINC000271778003	**-5.772**	**-6.987**	H.B: ASP 83, HIE 1051, LYS 1054, **GLY 1151**, **GLY 1153**, VAL 1155, **TYR 1161**	H.B: ASP 83, HIE 1051, LYS 1054, **GLY 1151**, **TYR 1161**
Neoeriocitrin	**-5.456**	**-7.792**	H.B: ASP 83, LYS 1054, ASP 1129, **GLY 1153**	H.B: ASP 83, LYS 1054, ASP 1129, PRO 1131, **GLY 1153**

Table 4.6: Binding affinities and interactions of the original complexed ligand and the five potential inhibitors that were not selected with the virus protease complex, as identified through the free access software Webina. With bold are the common interactions. The internal validation from Webina could not produce an affinity due to the boronic acid the original complexed ligand had

Compounds	Binding affinity (kcal*mol^-1)	Interactions	Pharmit
PDB: 5LC0	**Affinity**	**Webina**	**RMSD**
Original complexed ligand	**---**	H.I: **HIS 1051, LYS 1054, VAL 1155, TYR 1161** H.B: **SER 81, TYR 1130, GLY 1133, SER 1135, GLY 1151, GLY 1153, TYR 1161** π-S: **HIS 1051** π-C: **LYS 1054** S.B: **ASP 1129**	**---**
ZINC000013424720	**-8.746**	H.I: **HIS 1051** H.B: **SER 81**, VAL 1072, **TYR 1130**, **GLY 1133**, THR 1134, **SER 1135**, **TYR 1161** π-S: **HIS 1051**	0.108
ZINC000253389742	**-7.05**	H.I: **TYR 1161** H.B: **SER 81**, ASP 83, VAL 1036, **TYR 1130**, **GLY 1133**, **SER 1135**, **GLY 1153**, **TYR 1161** π-S: **HIS 1051** S.B: HIS 1051	0.461
ZINC000253529689	**-8.005**	H.B: VAL 1036, **TYR 1130**, **GLY 1133**, **SER 1135**, **GLY 1151**, ASN 1152, **GLY 1153**, **TYR 1161** π-S: TYR 1161 S.B: HIS 1051	0.090
ZINC000271778003	**-7.083**	H.I: ASP 83 H.B: **SER 81**, ASP 83, ASP 1129, **TYR 1130**, ALA 1132, TYR 1150, **GLY 1151**, ASN 1152, **GLY 1153**, **TYR 1161** π-S: **HIS 1051**, TYR 1161 S.B: HIS 1051	0.530
Neoeriocitrin	**-8.765**	H.I: **HIS 1051**, ALA 1132 H.B: **SER 81**, ASP 83, HIS 1051, **TYR 1130**, **GLY 1151**, ASN 1152, **GLY 1153**, **TYR 1161** π-S: **HIS 1051** S.B: HIS 1051, LYS 1054	**---**

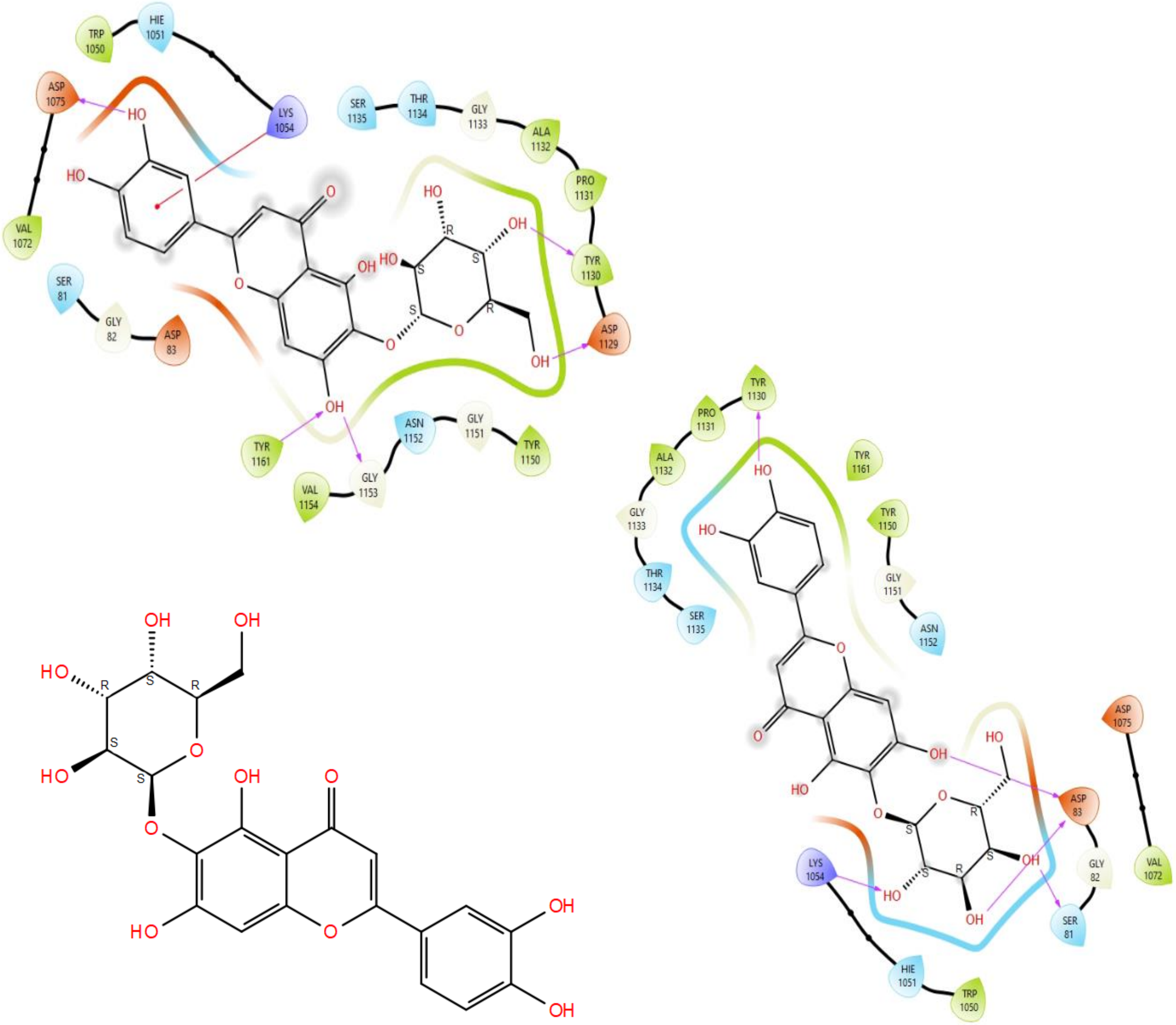


Figure 4.3: Two-dimensional representation of the compound ZINC000013424720 through Glide-SP (top left) and Glide-XP (top right) with the interactions into the receptor, and though Maestro (bottom) without the interactions

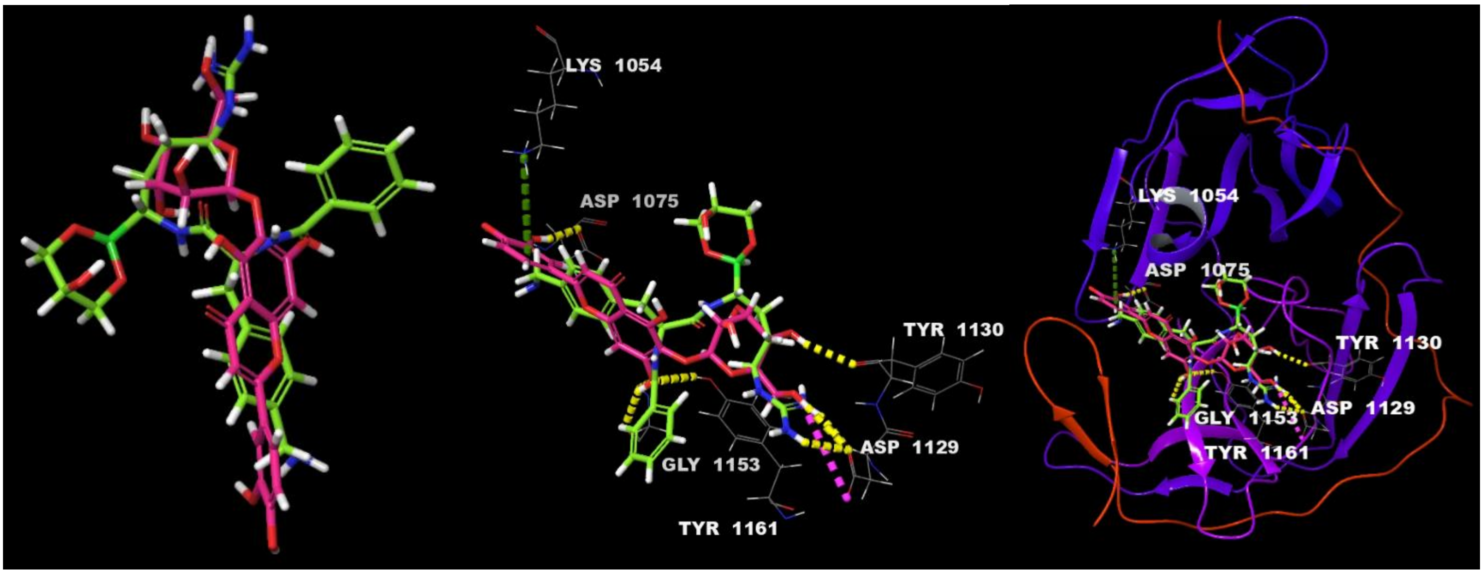


Figure 4.4: Three-dimensional representation of the compound ZINC000013424720 through Glide-SP (pink) superimposed with the original complexed ligand (green)

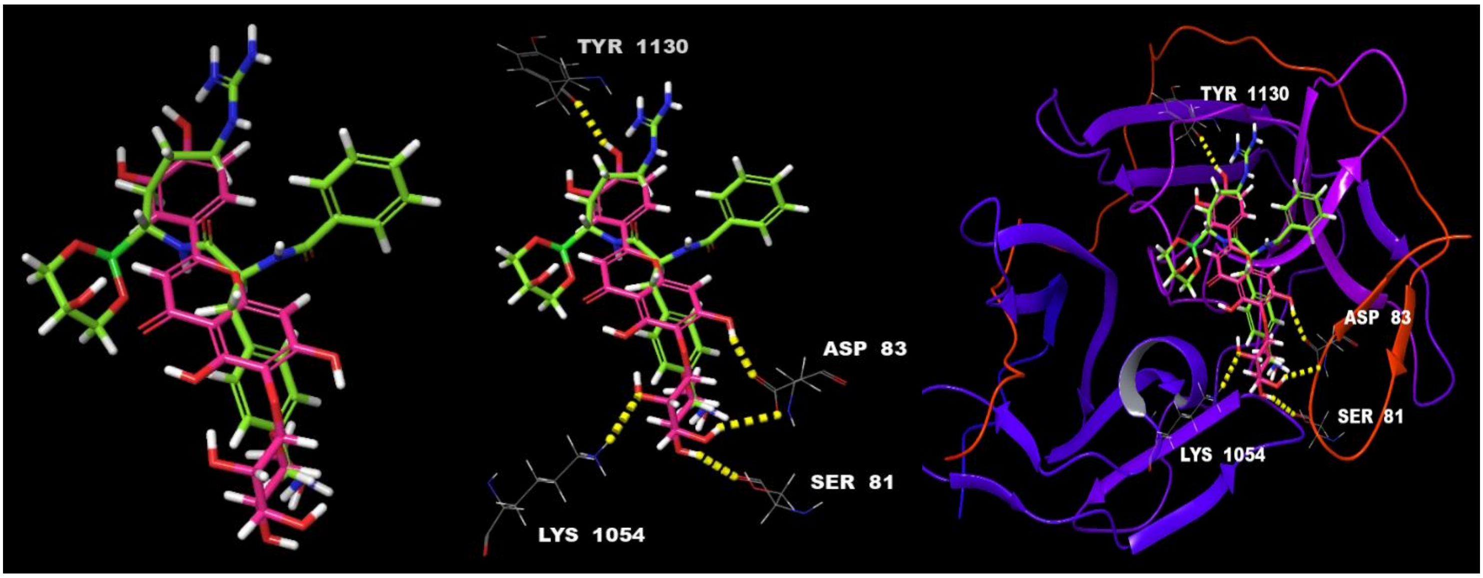


Figure 4.5: Three-dimensional representation of the compound ZINC000013424720 through Glide-XP (pink) superimposed with the original complexed ligand (green)

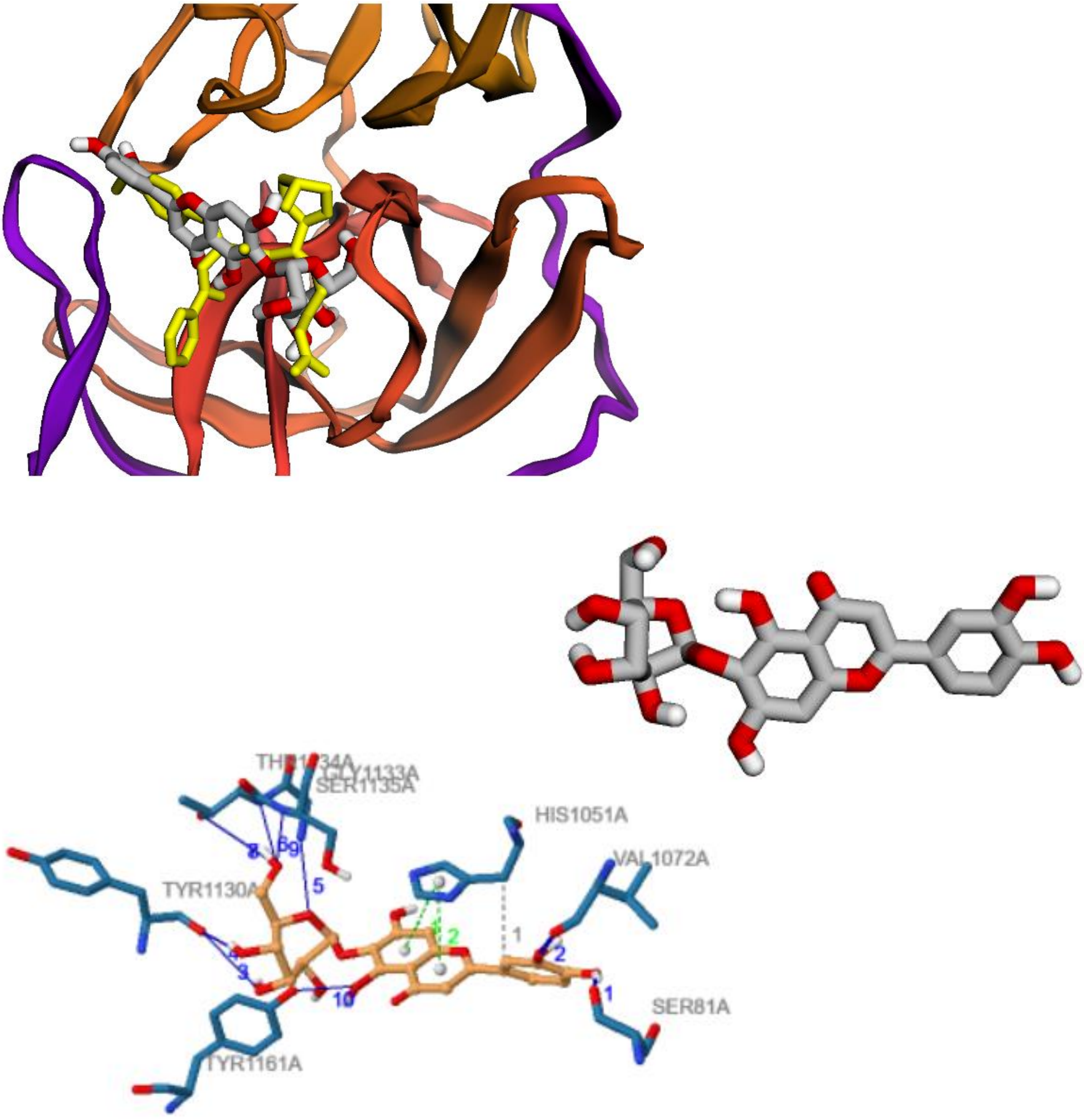


Figure 4.6: Three-dimensional representation of the compound ZINC000013424720 through Webina (top left and right) (grey) superimposed with the original complexed ligand (yellow) and two-dimensional representation of the compound though and PLIP (bottom left) with the interactions with the receptor

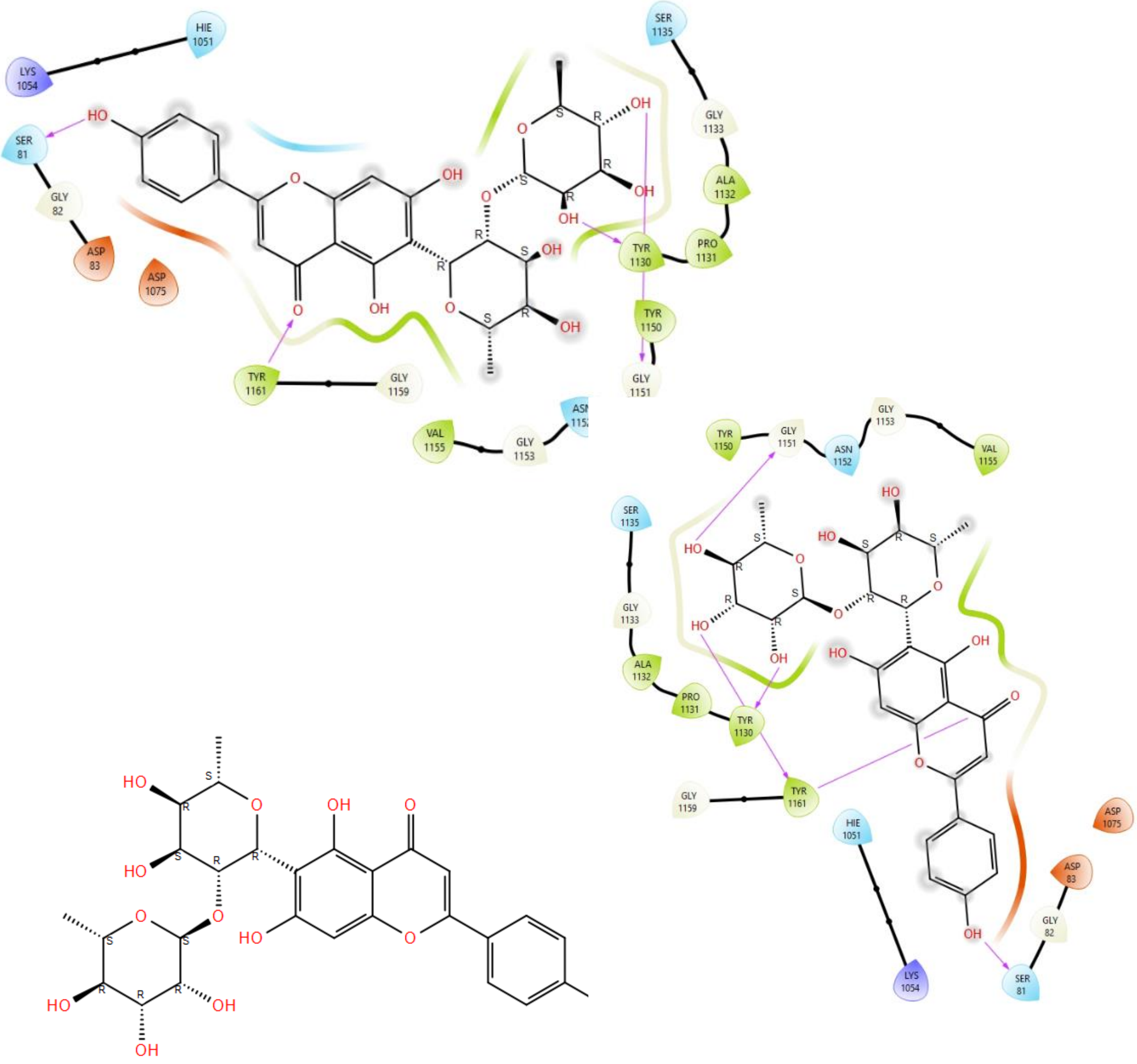


Figure 4.7: Two-dimensional representation of the compound ZINC000253389742 through Glide-SP (top left) and Glide-XP (top right) with the interactions with the receptor, and though Maestro (bottom) without the interactions

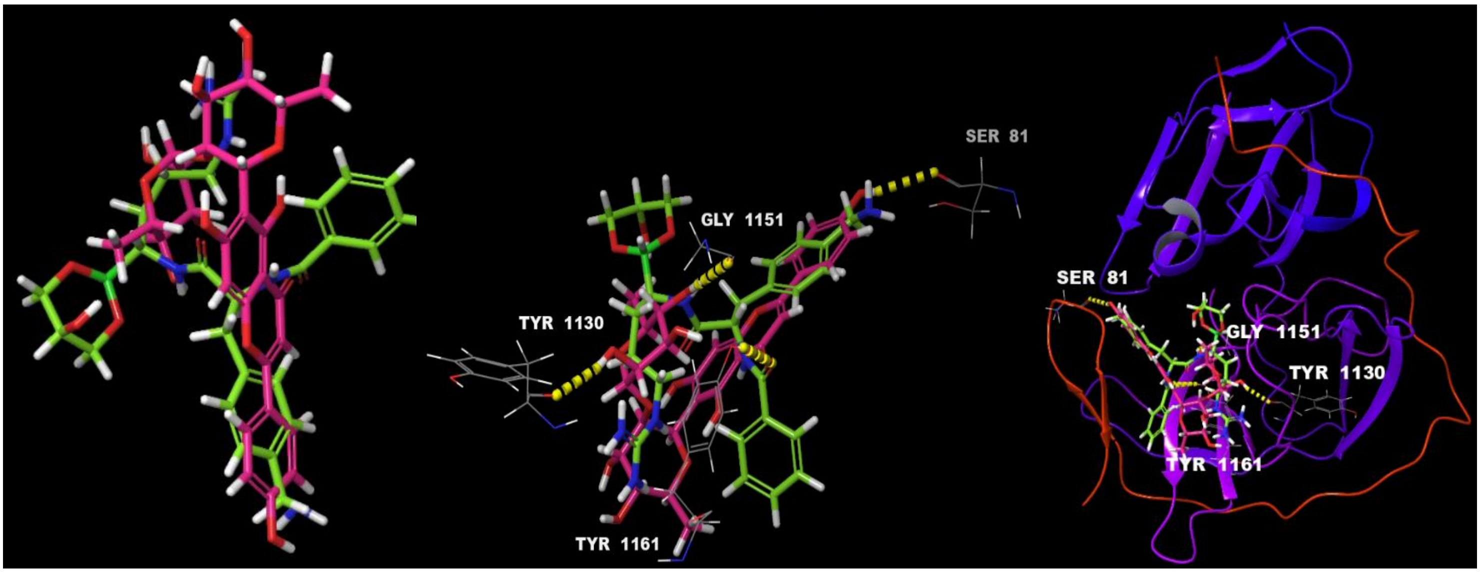


Figure 4.8: Three-dimensional representation of the compound ZINC000253389742 through Glide-SP (pink) superimposed with the original complexed ligand (green)

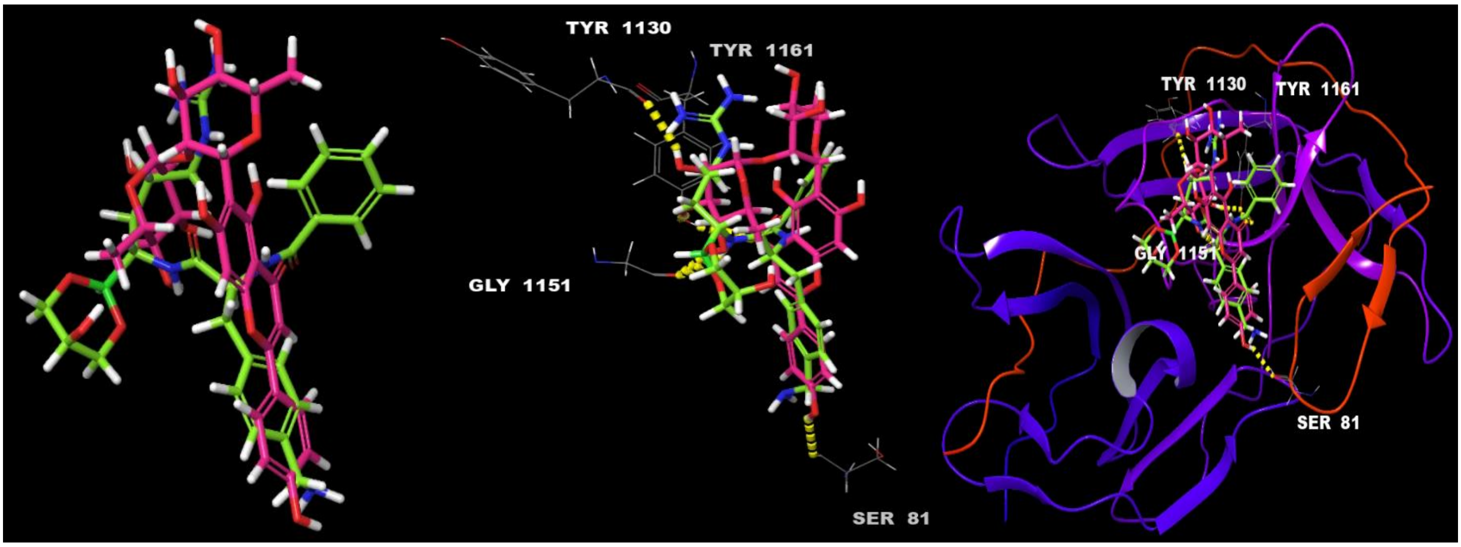


Figure 4.9: Three-dimensional representation of the compound ZINC000253389742 through Glide-XP (pink) superimposed with the original complexed ligand (green)

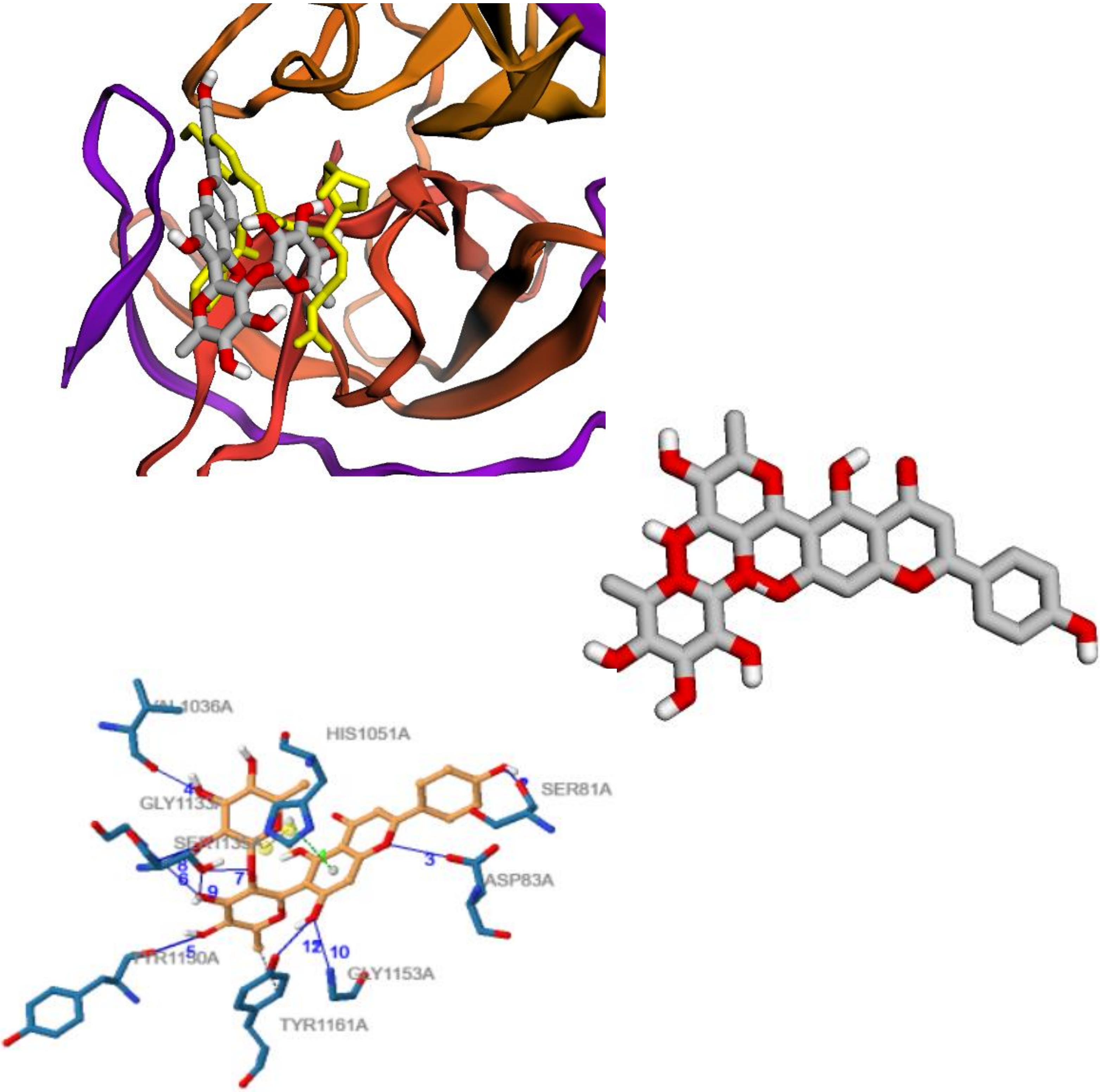


Figure 4.10: Three-dimensional representation of the compound ZINC000253389742 through Webina (top left and right) (grey) superimposed with the original complexed ligand (yellow) and two-dimensional representation of the compound though and PLIP (bottom left) with the interactions with the receptor

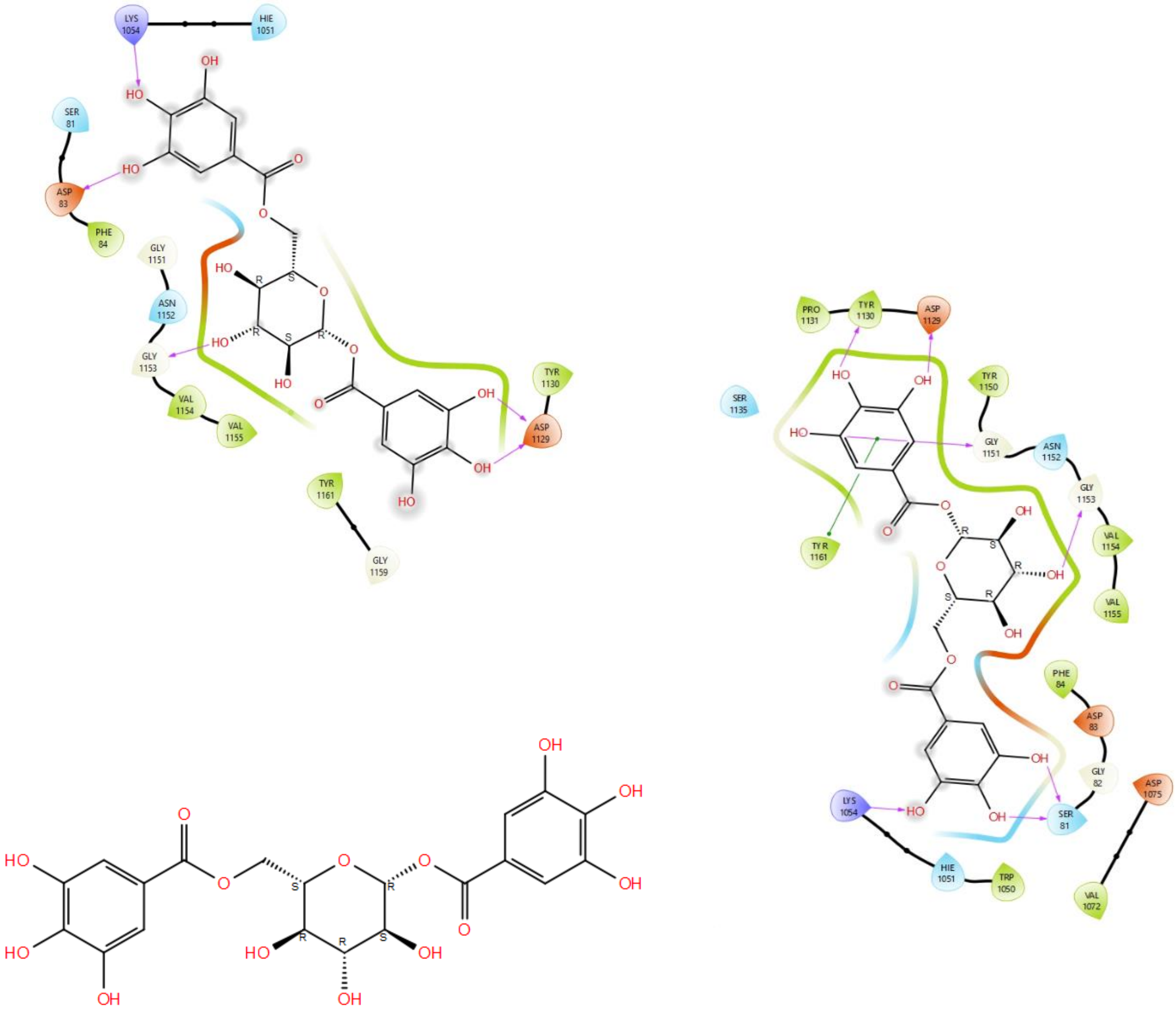


Figure 4.11: Two-dimensional representation of the compound ZINC000253529689 through Glide-SP (top left) and Glide-XP (top right) with the interactions with the receptor, and though Maestro (bottom) without the interactions

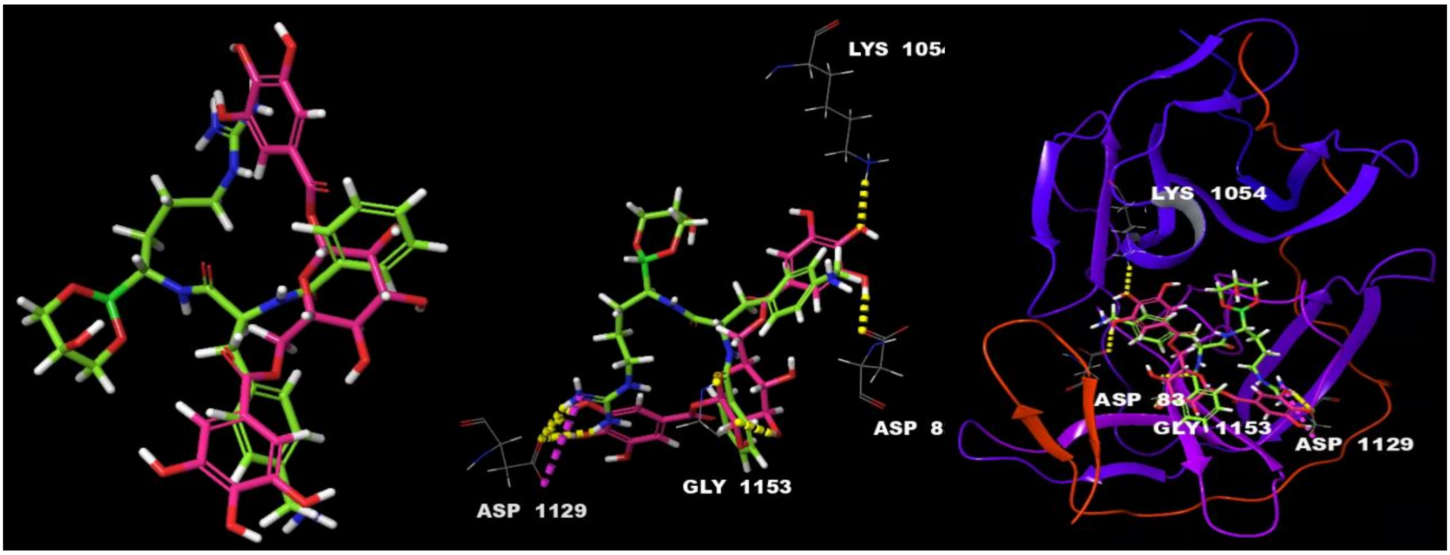


Figure 4.12: Three-dimensional representation of the compound ZINC000253529689 through Glide-SP (pink) superimposed with the original complexed ligand (green)

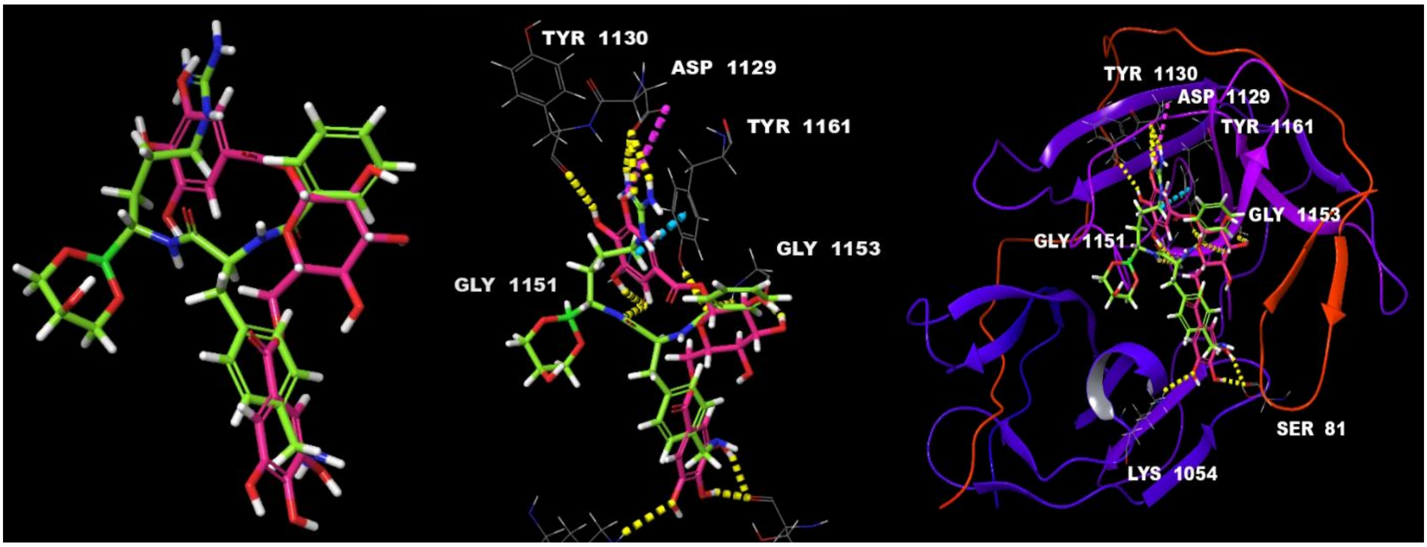


Figure 4.13: Three-dimensional representation of the compound ZINC000253529689 through Glide-XP (pink) superimposed with the original complexed ligand (green)

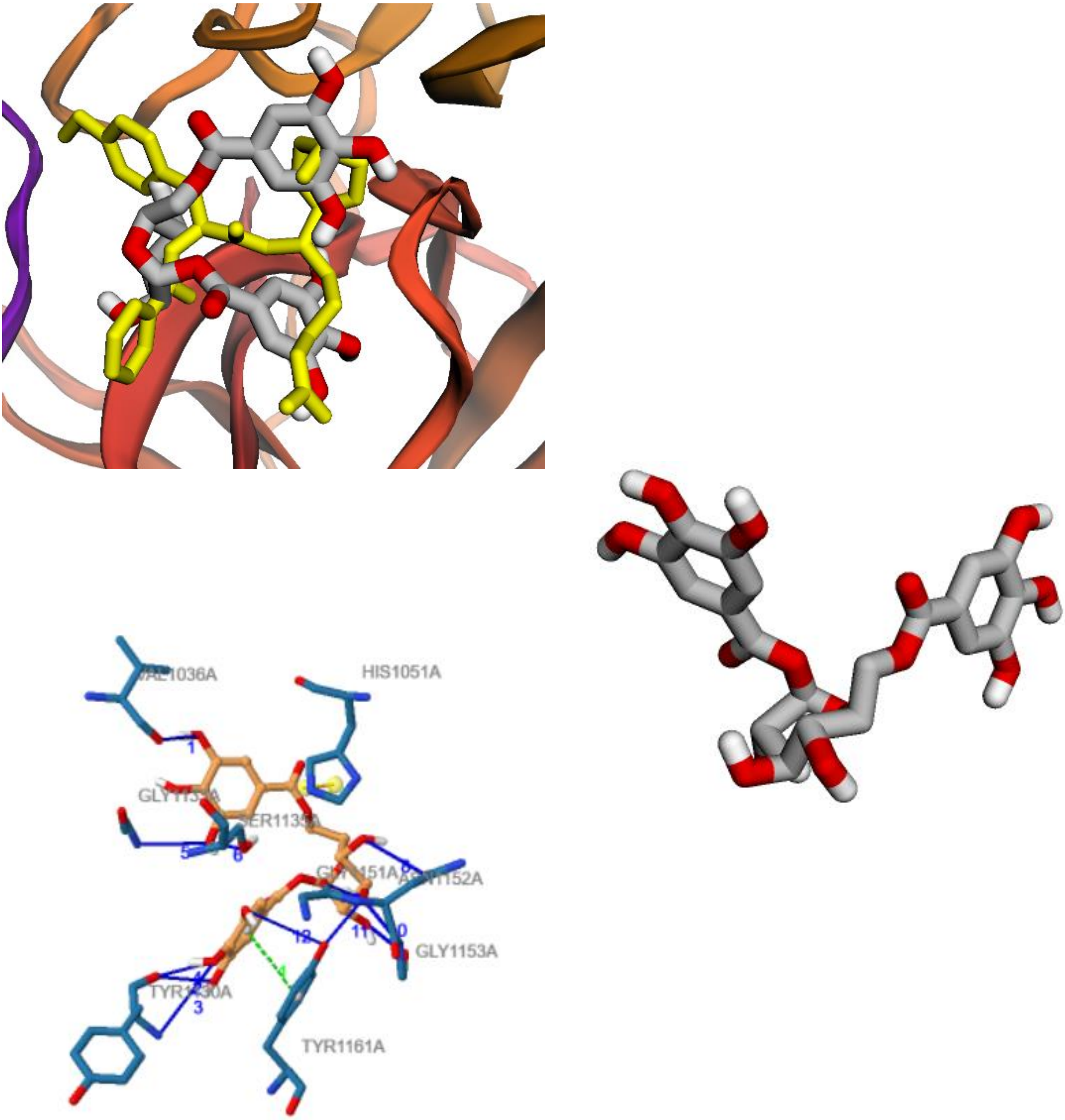


Figure 4.14: Three-dimensional representation of the compound ZINC000253529689 through Webina (top left and right) (grey) superimposed with the original complexed ligand (yellow) and two-dimensional representation of the compound though and PLIP (bottom left) with the interactions with the receptor

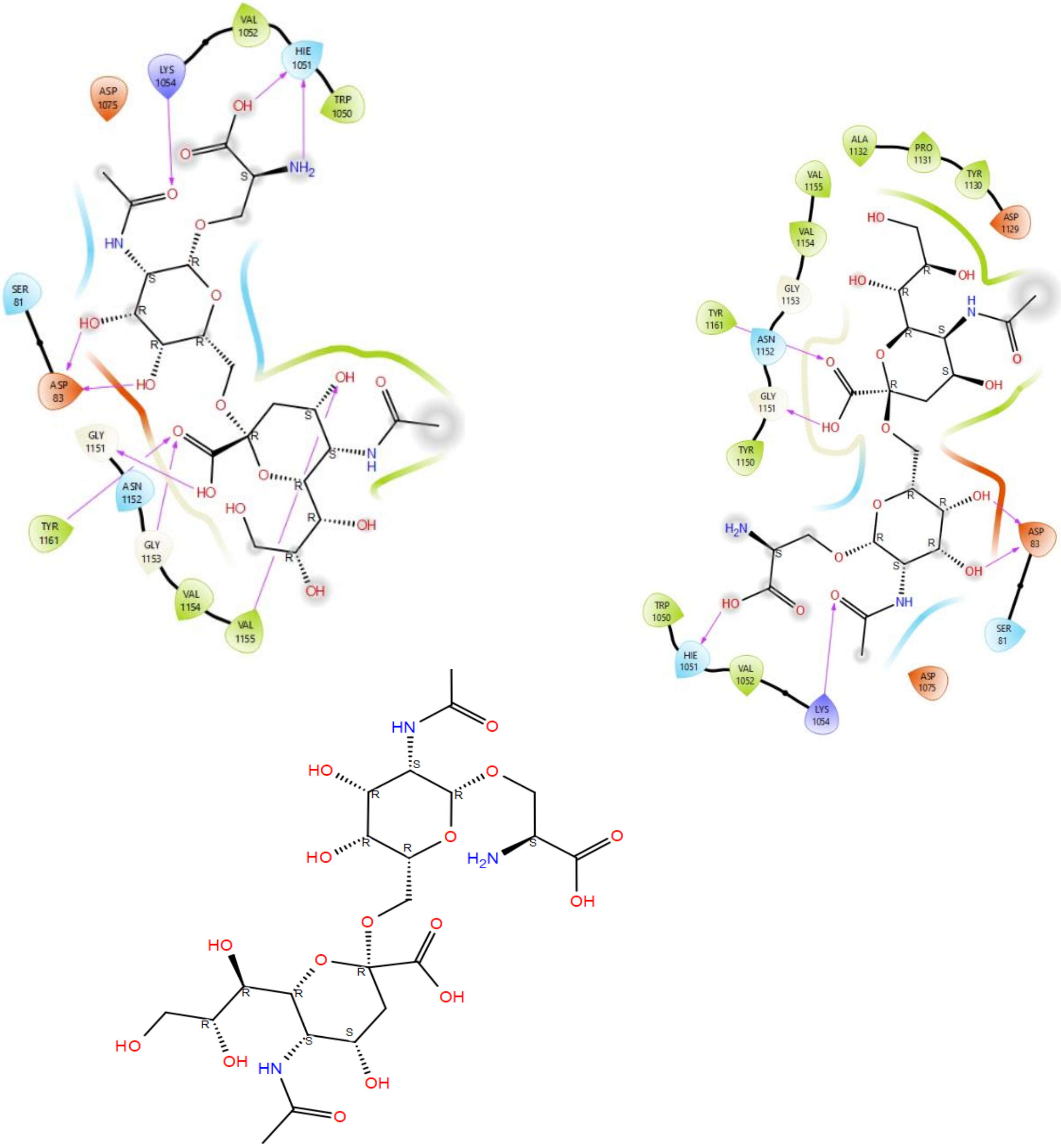


Figure 4.15: Two-dimensional representation of the compound ZINC000271778003 through Glide-SP (top left) and Glide-XP (top right) with the interactions with the receptor, and though Maestro (bottom) without the interactions

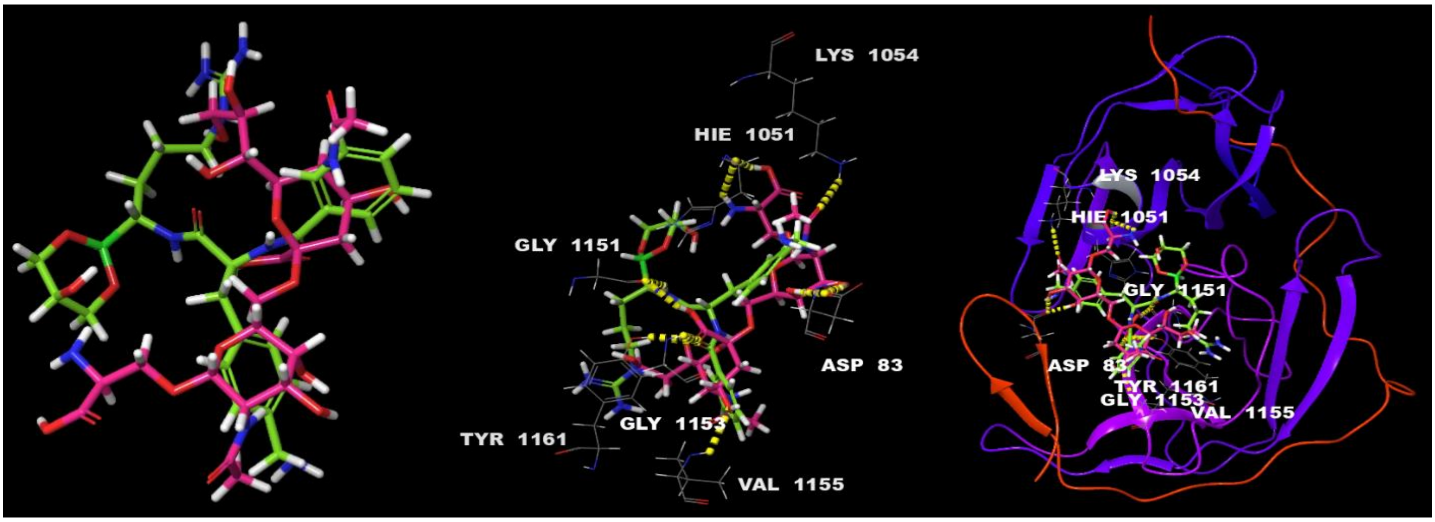


Figure 4.16: Three-dimensional representation of the compound ZINC000271778003 through Glide-SP (pink) superimposed with the original complexed ligand (green)

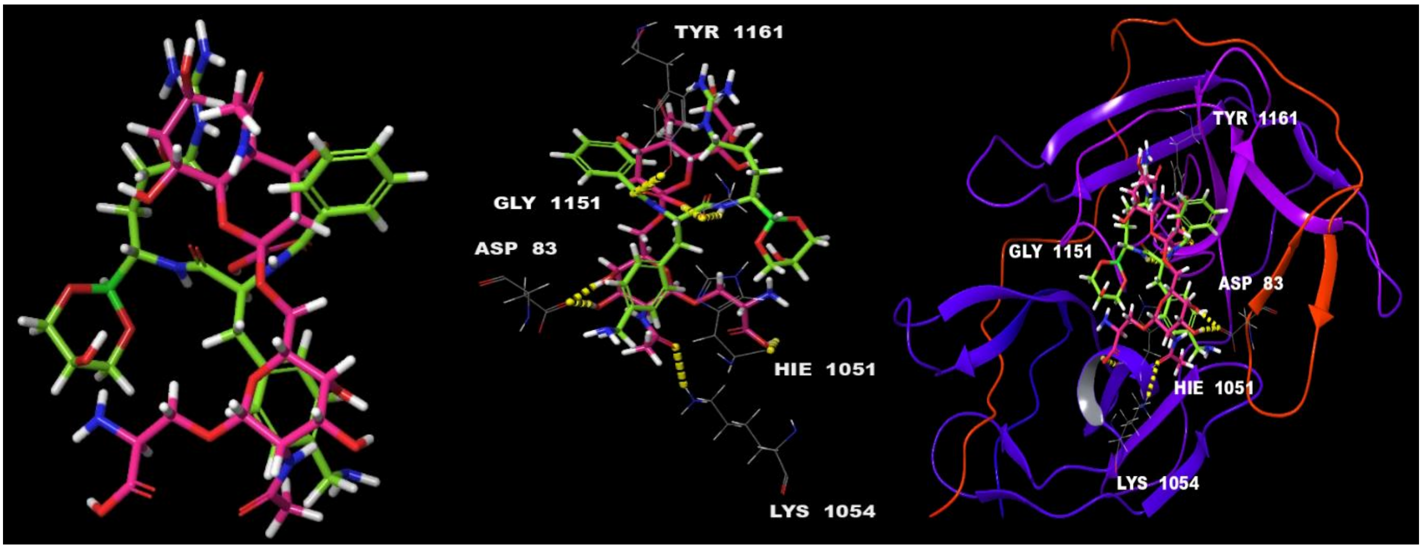


Figure 4.17: Three-dimensional representation of the compound ZINC000271778003 through Glide-XP (pink) superimposed with the original complexed ligand (green)

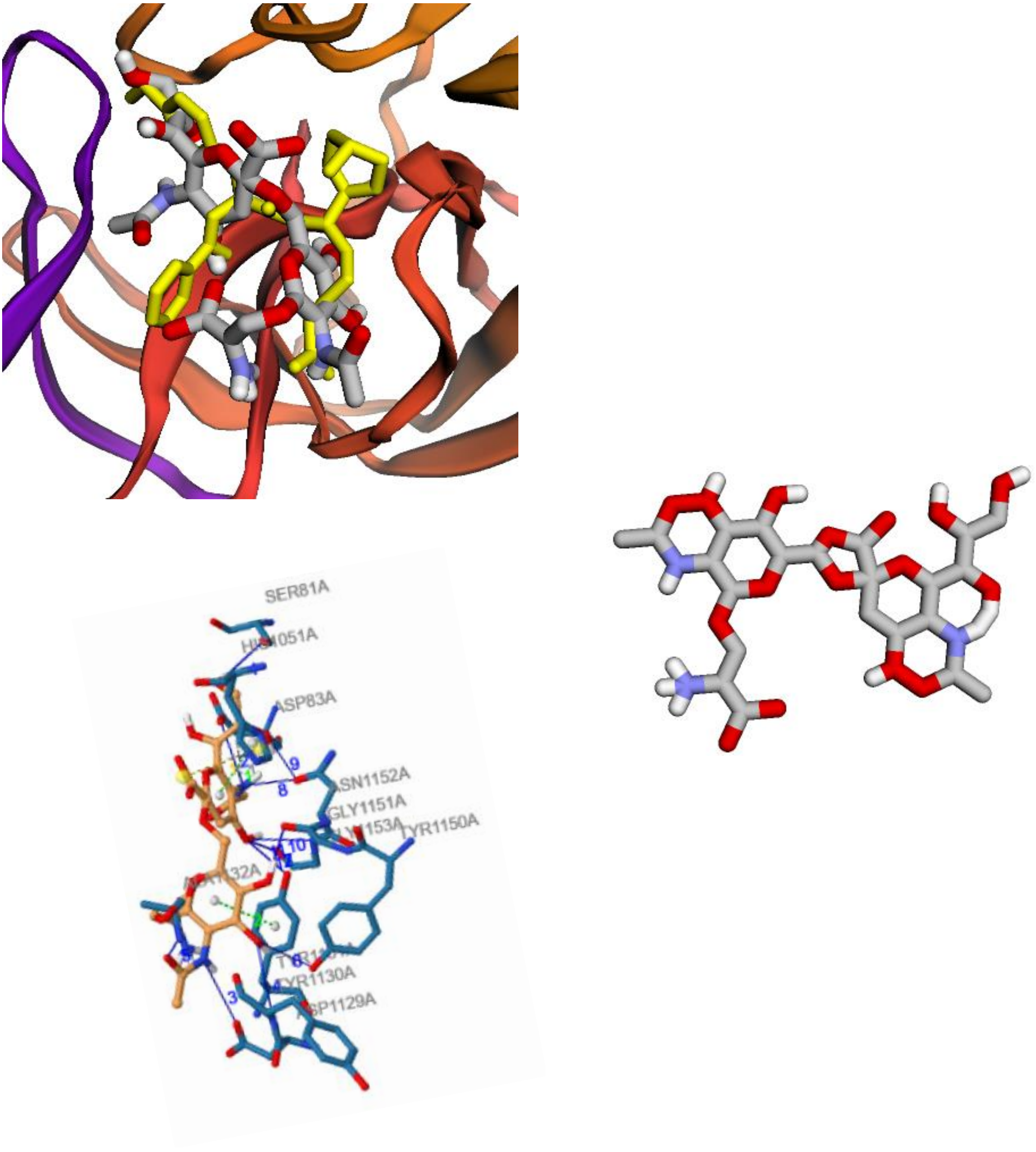


Figure 4.18: Three-dimensional representation of the compound ZINC000271778003 through Webina (top left and right) (grey) superimposed with the original complexed ligand (yellow) and two-dimensional representation of the compound though and PLIP (bottom left) with the interactions with the receptor

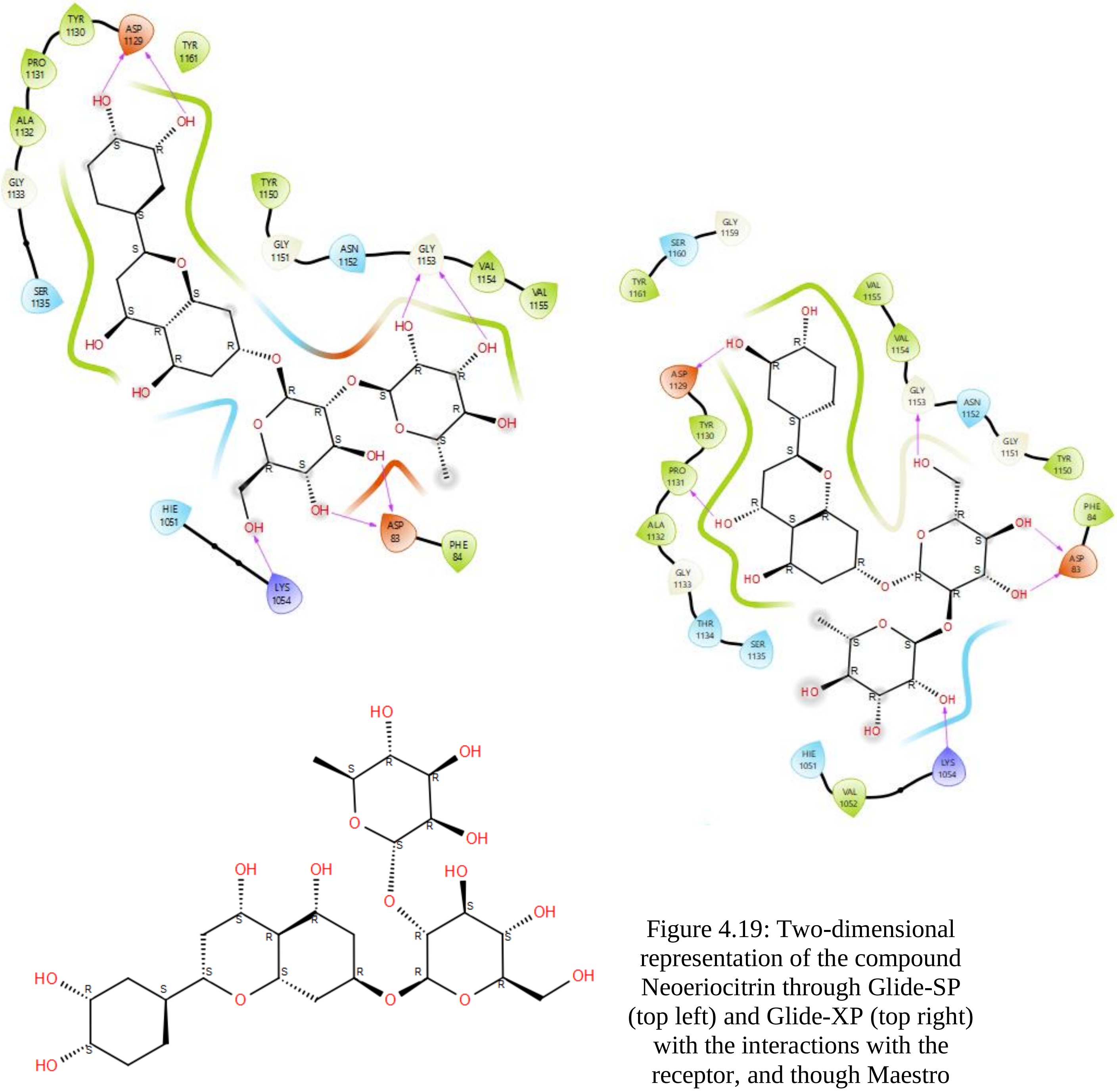


Figure 4.19: Two-dimensional representation of the compound Neoeriocitrin through Glide-SP (top left) and Glide-XP (top right) with the interactions with the receptor, and though Maestro (bottom) without the interactions

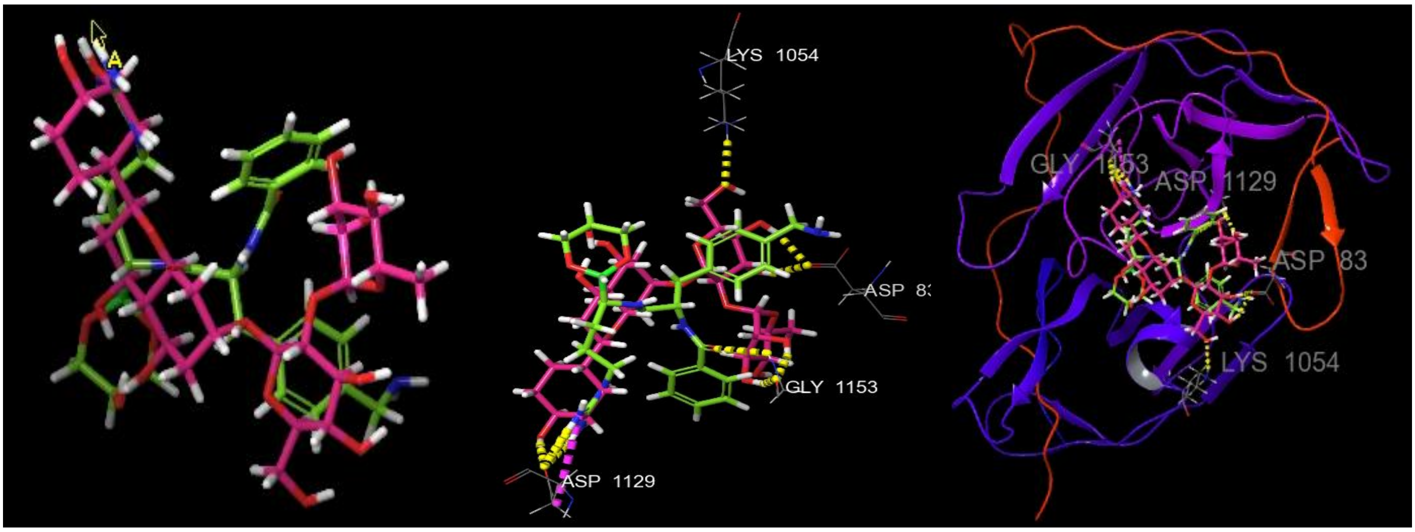


Figure 4.20: Three-dimensional representation of the compound Neoeriocitrin through Glide-SP (pink) superimposed with the original complexed ligand (green)

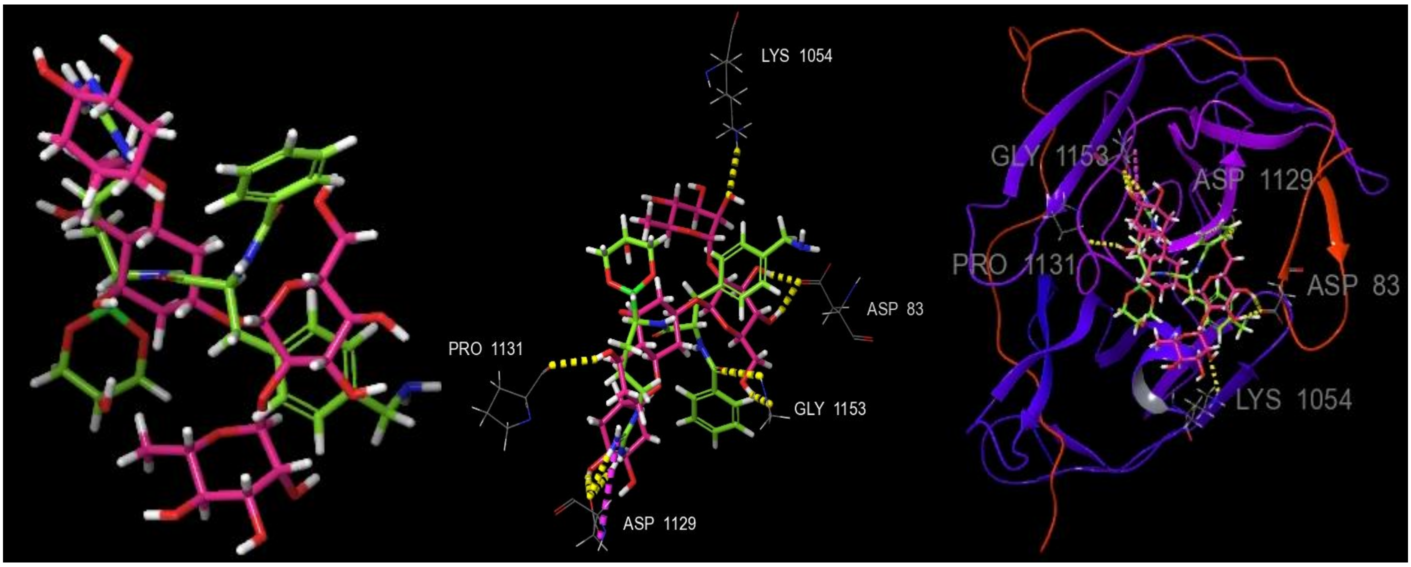


Figure 4.21: Three-dimensional representation of the compound Neoeriocitrin through Glide-XP (pink) superimposed with the original complexed ligand (green)

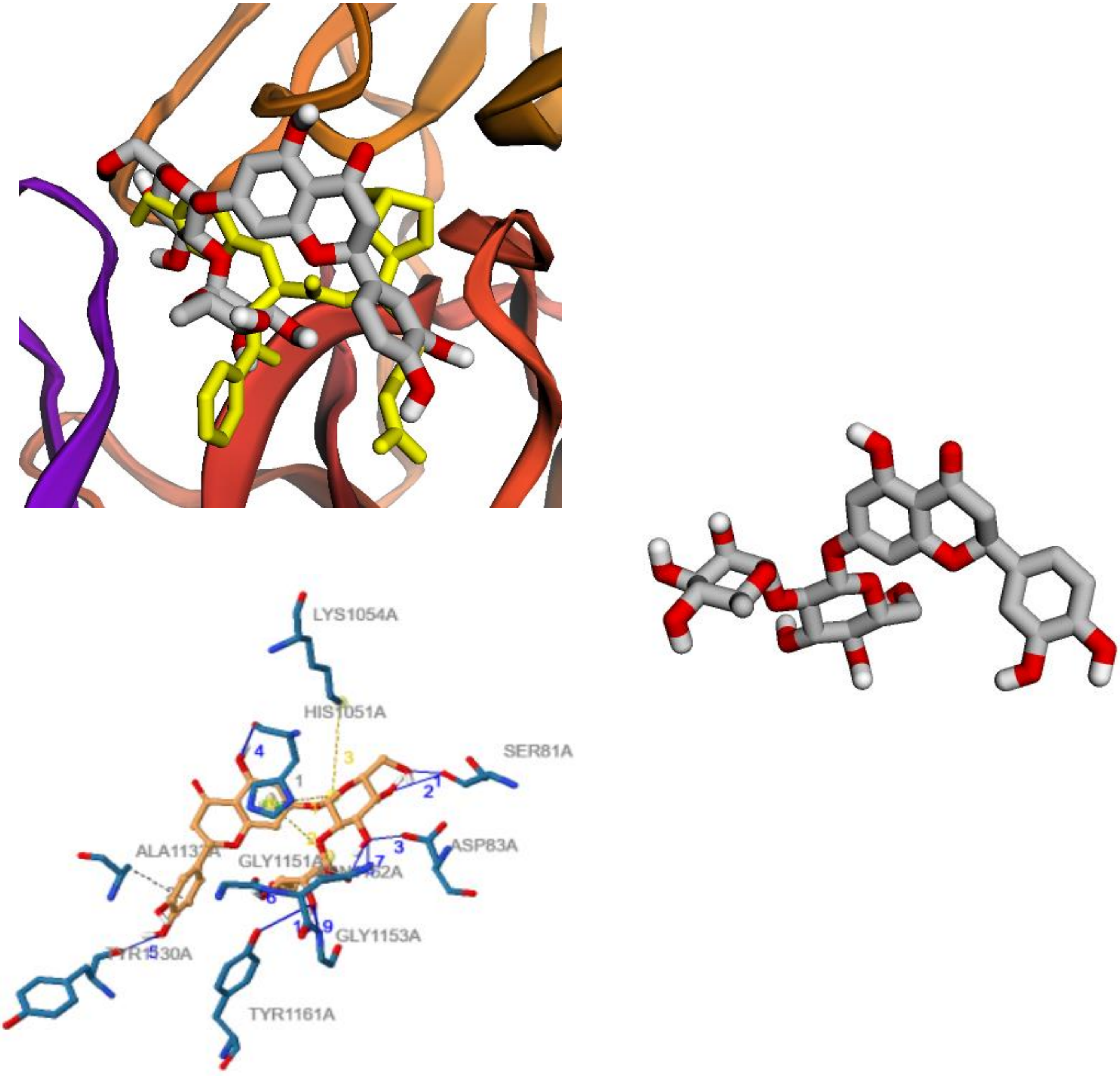
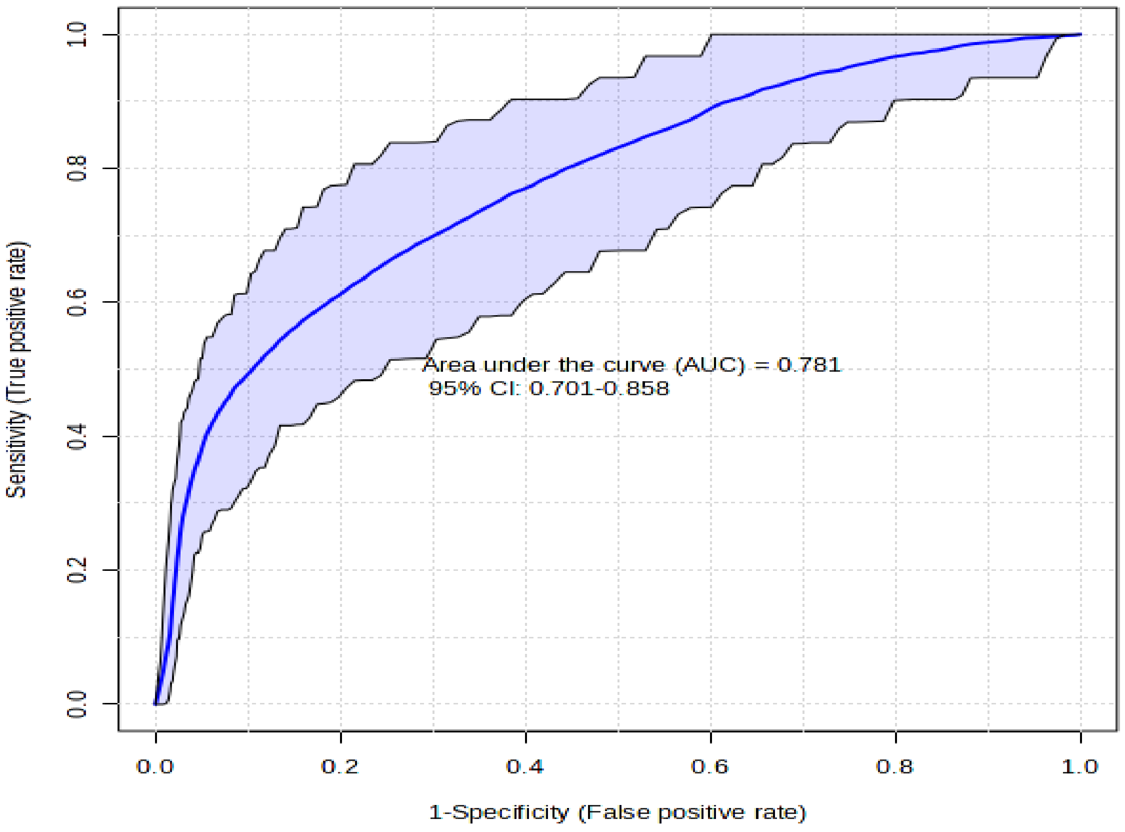


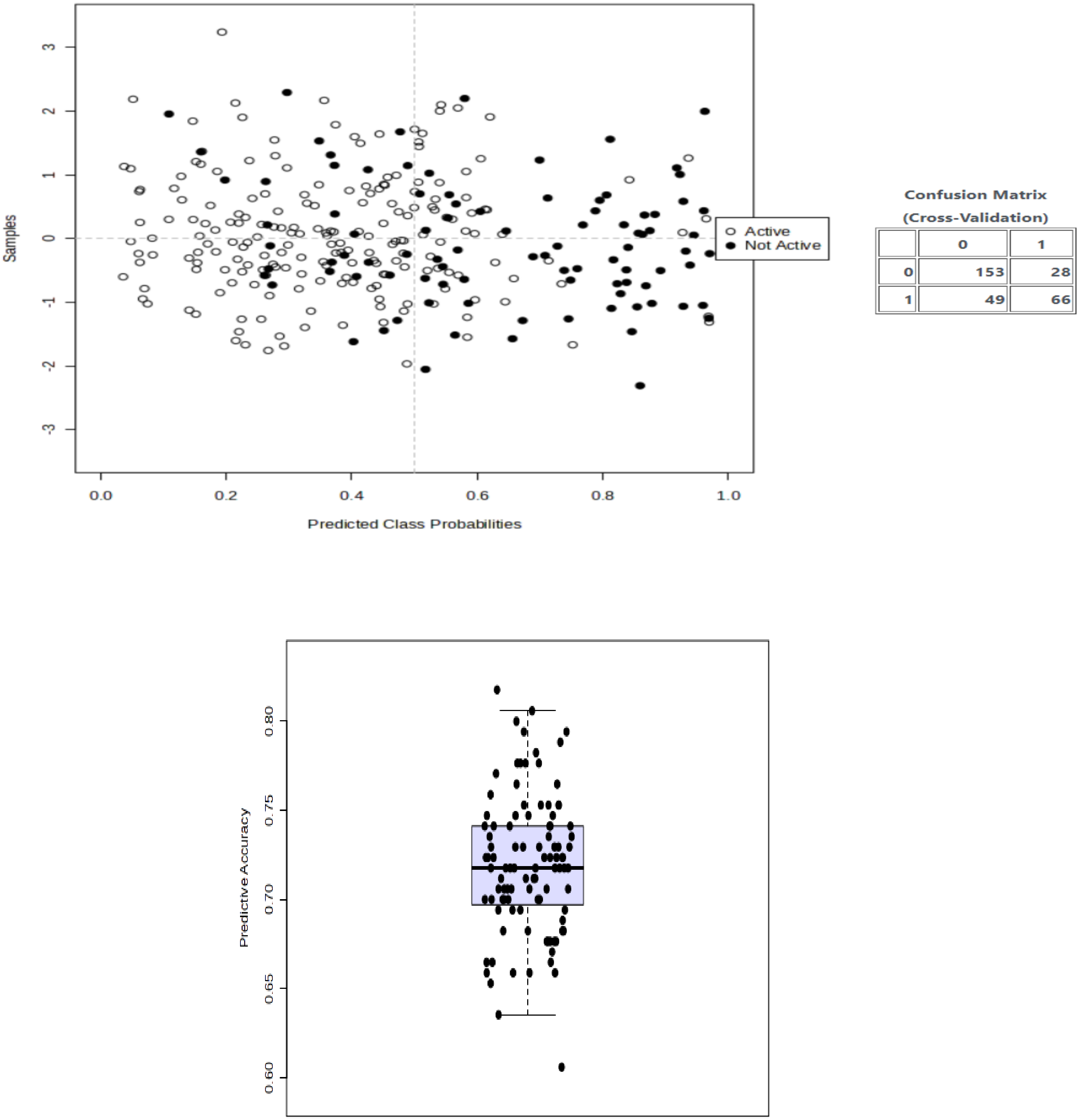
Figure 4.22: Three-dimensional representation of the compound Neoeriocitrin through Webina (top left and right) (grey) superimposed with the original complexed ligand (yellow) and two-dimensional representation of the compound though and PLIP (bottom left) with the interactions with the receptor

4.4 Results of combined methodologies

Table 4.7: Classification of the five selected compounds, as active or inactive, based on a model created with MetaboAnalyst

Compound Name	Possibility	Category
ZINC000013424720	0.69	Active
ZINC000253389742	0.64	Active
ZINC000253529689	0.61	Active
ZINC000271778003	0.70	Active
Neoeriocitrin	0.63	Active





	0	1
0	153	28
1	49	66

Figure 4.23: Receiver Operating Characteristic curve, box plot, and scatter plot that provide a probability of correct classification close to 70%

DISCUSSION

The Wilcoxon rank sum test proved that out of the 10 potentially most important features, eight showed a statistically significant difference at a significance level of 0.001 (Table 4.1). This indicates that each of these features had a distinct value distribution between the active and inactive compound categories, making them effective for distinguishing between them. As a result, these features can contribute to more accurate predictions of compound activity.

Figure 4.2 displays a Receiver Operating Characteristic (ROC) curve generated using the Random Forest classifier, which achieved an accuracy of 95.46%. This accuracy shows that the model is highly effective in correctly classifying compounds as active or not active. The optimal combination of seven molecular descriptors, as shown in Table 4.2, was determined to be the most efficient for this classification. This feature set enabled the model to achieve a sensitivity of 95.84% and a specificity of 95.10%, demonstrating that it handles both true positive and true negative predictions with minimal errors, thus ensuring a well-balanced and robust performance.

Notably, the Area Under the Curve (AUC) reached a perfect score of 1, suggesting flawless classification. However, such a perfect AUC is often unrealistic. While this indicates the model performed exceptionally well on the data used, it’s crucial to validate it further with independent datasets to ensure its reliability in real-world applications.

From the in silico molecular docking experiments, five potential inhibitors were selected (table 4.5 – 4.6) based on several key criteria that assessed their ability to effectively dock to the receptor. The selection process prioritized compounds with low root mean square deviation values, indicating minimal structural deviation from the original ligand, which enhances the likelihood of favorable binding interactions.

Additionally, these compounds exhibited low binding affinities that were comparable to that of the original ligand, further supporting their potential as viable candidates for docking. A significant emphasis was placed on the nature and quantity of the interactions, particularly in the hydrogen bonds, which are crucial in stabilizing receptor and ligand complexes. The selected inhibitors had a good amount of hydrogen interactions, but also shared numerous common interactions with the ones the original ligand had, enhancing their potential for effective binding to the virus protease complex.

Flavonoids are a group of natural polyphenolic compounds commonly found in vegetables, fruits and certain beverages like tea and wine. They possess various biological activities, including anti-inflammatory, antioxidant and antiviral properties, making them suitable candidates for medicinal research [39]. All the selected compounds, except ZINC000271778003, belong to this family and primarily target enzymes.

ZINC000271778003, on the other hand, is a complex serine derivative known as 3-O-(2-acetamido-6-O-(N-acetylneuraminyl)-2-deoxygalactosyl) serine (STn Epitope). Although not classified as a typical flavonoid, it can still be considered a naturally occurring molecule due to its serine-based structure. Like flavonoids, it could exhibit significant biological activity, potentially enzyme inhibition or modulation.

Further validation of the selected compounds was achieved using a model created in MetaboAnalyst. This model classified all compounds as active, with an AUC value of 0.781, indicating a reliable performance in distinguishing active compounds from not active ones. This result reinforces the effectiveness of the selection process and the potential of these compounds as inhibitors against the Zika virus protease NS2B-NS3.

CONCLUSIONS AND FUTURE PERSPECTIVES

This study by employing a methodology of combining statistical analysis, machine learning, and molecular docking, identified promising candidates, primarily from the flavonoid family. These natural compounds have gained popularity for their antiviral properties due to their ability to interfere with various stages of viral infection, such as viral entry, replication, and assembly. Moreover, they exhibit relatively low toxicity, making them attractive candidates for developing antiviral therapies.

Other studies have supported the conclusion that flavonoids can serve as effective inhibitors. Notable examples include flavonoids from Pterogyne nitens, which have been shown to inhibit the Zika virus NS2B-NS3 protease [40] and flavonoids derived from the geopropolis of the Brazilian Jandaira bee, which effectively inhibit the replication of both Zika and Dengue viruses in vitro [41]. These findings and many more underscore the potential of flavonoids as promising antiviral drugs.

In the future, the aim is to conduct molecular dynamics simulations to evaluate the stability and interaction dynamics of the identified compounds with the NS2B-NS3 protease complex over time. This approach will improve our understanding of the binding affinities and conformational changes that occur, enhancing our predictions regarding their inhibitory effectiveness. Following these analyses, testing the most promising candidates in vitro will follow and depending on the results this methodology can be extended to explore other chemical libraries, increasing the chance of finding more potential inhibitors.

REFERENCES

1. Fathima AJ, Murugaboopathi G, Selvam P. Design and docking studies of small molecule derivatives as Zika virus NS2B-NS3 protease inhibitors: A computational approach. International Journal of Pure and Applied Mathematics. 2018;

2. Akaberi D, Chinthakindi PK, Båhlström A, Palanisamy N, Sandström A, Lundkvist Å, et al. Identification of a C2-symmetric diol-based human immunodeficiency virus protease inhibitor targeting Zika virus NS2B-NS3 protease. J Biomol Struct Dyn. 2020;38(18):5526–36.

3. Law WY, Asaruddin MR, Bhawani SA. Antiviral study of Schiff base vanillin derivatives against NS2B-NS3 protease of Zika virus based on pharmacophore modeling and molecular docking. 2023;27(6).

4. Altayb HN, Alatawi HA. Employing machine learning-based QSAR for targeting Zika virus NS3 protease: Molecular insights and inhibitor discovery. Pharmaceuticals. 2024;17(8):1067.

5. Bhukya PL, Mhaske ST, Sonkar SC, editors. Emerging human viral diseases, Volume I: Respiratory and hemorrhagic fever. Singapore: Springer Nature; 2023.

6. Campbell NA, Reece JB, Urry LA, Cain ML, Wasserman SA, Minorsky PV, et al. Biology. Vol. I: The chemistry of life – The cell – Genetics. Heraklion: University of Crete Press; 2019.

7. Widmaier EP, Raff H, Strang KT. Vander's Human Physiology: Mechanisms of body function. 13th ed. 2nd Greek ed. Athens: Broken Hill Publishers LTD; 2015.

8. Marcondes CB, editor. Arthropod-borne diseases. Cham: Springer International Publishing; 2017.

9. Loudon M, Parise J. Organic chemistry. 6th ed. Nicosia: Broken Hill Publishers Ltd; 2019.

10. Tymoczko JL, Berg JM, Stryer L. Biochemistry: A short course. 3rd ed. Broken Hill Publishers LTD; 2019.

11. Holbrook M. Historical perspectives on flavivirus research. Viruses. 2017;9(5):97.

12. Durgam L, Pagag J, Indra Neela Y, Guruprasad L. Mutational analyses, pharmacophore-based inhibitor design, and in silico validation for Zika virus NS3 helicase. J Biomol Struct Dyn. 2023;1–19.

13. Mirza MU, Alanko I, Vanmeert M, Muzzarelli KM, Salo-Ahen OMH, Abdullah I, et al. Discovery of Zika virus NS2B-NS3 inhibitors with antiviral activity via an integrated virtual screening approach. Eur J Pharm Sci. 2022;175:106220.

14. Nunes DADF, Santos FRDS, Da Fonseca STD, De Lima WG, Nizer WSDC, Ferreira JMS, et al. NS2B-NS3 protease inhibitors as promising compounds in the development of antivirals against Zika virus: A systematic review. J Med Virol. 2022;94(2):442–53.

15. Meewan I, Shiryaev SA, Kattoula J, Huang CT, Lin V, Chuang CH, et al. Allosteric inhibitors of Zika virus NS2B-NS3 protease targeting the protease in "super-open" conformation. Viruses. 2023;15(5):1106.

16. Lei J, Hansen G, Nitsche C, Klein CD, Zhang L, Hilgenfeld R. Crystal structure of Zika virus NS2B-NS3 protease in complex with a boronate inhibitor.

17. Song W, Zhang H, Zhang Y, Li R, Han Y, Lin Y, et al. Repurposing clinical drugs is a promising strategy to discover drugs against Zika virus infection. Front Med. 2021;15(3):404–15.

18. Gaulton A, Bellis LJ, Bento AP, et al. ChEMBL: A large-scale bioactivity database for drug discovery. Nucleic Acids Res. 2012;40(D1).

19. Dong J, Cao DS, Miao HY, et al. ChemDes: An integrated web-based platform for molecular descriptor and fingerprint computation. J Cheminform. 2015;7:60.

20. Landrum G. RDKit: Open-source cheminformatics software. 2006.

21. Toledo-Pérez DC, Aviles M, Toledo-Pérez RA, Rodríguez-Reséndiz J. Feature set to sEMG classification obtained with Fisher score. IEEE Access. 2024;12:13962–70.

22. Grus J. Data science: Basic principles and applications with Python. 2nd ed. Athens: Papasotiriou; 2021.

23. Sarkar D, Bali R, Sharma T. Practical machine learning with Python. Berkeley, CA: Apress; 2018.

24. Albon C. Machine learning with Python cookbook: Practical solutions from preprocessing to deep learning. 2018.

25. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The Protein Data Bank. Nucleic Acids Res. 2000;28(1):235-42.

26. Adasme MF, Linnemann KL, Bolz SN, Kaiser F, Salentin S, Haupt VJ, Schroeder M. PLIP 2021: Expanding the scope of the protein–ligand interaction profiler to DNA and RNA. Nucleic Acids Res. 2021;49(W1):W530-4.

27. Tro NJ. Principles of chemistry: A molecular approach. Nicosia: Broken Hill Publishers Ltd; 2012.

28. Sunseri J, Koes DR. Pharmit: Interactive exploration of chemical space. Nucleic Acids Res. 2016;44(W1): W442-8.

29. Irwin JJ, Sterling T, Mysinger MM, Bolstad ES, Coleman RG. ZINC: A free tool to discover chemistry for biology. J Chem Inf Model. 2012;52(7):1757-68.

30. Musarra-Pizzo M, Pennisi R, Ben-Amor I, Mandalari G, Sciortino MT. Antiviral activity exerted by natural products against human viruses. Viruses. 2021;13(5):828.

31. Todeschini R, Consonni V. Handbook of molecular descriptors. 1st ed. Wiley; 2000.

32. Papageorgiou EG. Probabilities - Biostatistics and applications with SPSS. Athens: New Technologies Publications; 2020.

33. Raschka S, Mirjalili V. Python machine learning: Machine learning and deep learning with Python, scikit-learn, and TensorFlow 2. 3rd ed. Birmingham Mumbai: Packt; 2019.

34. Trott O, Olson AJ. AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. J Comput Chem. 2010;31(2):455-61.

35. Schrödinger, LLC. Maestro, Version 12.5. New York, NY; 2020.

36. Dallakyan S, Olson AJ. Small-molecule library screening by docking with PyRx. In: Chemical biology. Humana Press, New York, NY; 2015. p. 243-50.

37. Schomburg I, Chang A, Ebeling C, Gremse M, Heldt C, Huhn G, Schomburg D. BRENDA, the enzyme database: Updates and major new developments. Nucleic Acids Res. 2004;32(suppl_1): D431-D433.

38. Pang Z, Chen J, Li S, et al. MetaboAnalyst 5.0: Narrowing the gap between metabolomics and systems biology. Nucleic Acids Res. 2021;49(W1): W388-W396.

39. Badshah SL, Faisal S, Muhammad A, Poulson BG, Emwas AH, Jaremko M. Antiviral activities of flavonoids. Biomed Pharmacother.2021;140:111596.

40. Lima CS, Mottin M, De Assis LR, Mesquita NCDMR, Sousa BKDP, Coimbra LD, et al. Flavonoids from Pterogyne nitens as Zika virus NS2B-NS3 protease inhibitors. Bioorg Chem. 2021;109: 104719.

41. Silva PGD, Chaves EJF, Silva TMS, Rocha GB, Dantas WM, Oliveira RND, et al. Antiviral activity of flavonoids from geopropolis of the Brazilian Jandaira bee against Zika and dengue viruses. Pharmaceutics. 2023 Oct 19;15(10):2494.