MOLECULAR FINGERPRINT SCREENING OF MARINE NATURAL COMPOUNDS USING ML APPROACH

Submitted in partial fulfillment of the requirements for the award of
Bachelor of Science degree in
Bioinformatics and Data Science

By

ASWINI T

(40738003)

DEPARTMENT OF BIOINFORMATICS AND DATA SCIENCE

SCHOOL OF BIO AND CHEMICAL ENGINEERING

SATHYABAMA INSTITUTE OF SCIENCE AND TECHNOLOGY

(DEEMED TO BE UNIVERSITY)

Accredited with Grade "A" by NAAC I 12B Status

by UGC I Approved by AICTE

JEPPIAAR NAGAR, RAJIV GANDHI SALAI,

CHENNAI - 600 119

MAY - 2023

SATHYABAMA

INSTITUTE OF SCIENCE AND TECHNOLOGY

(DEEMED TO BE UNIVERSITY)

Accredited "A" Grade by NAAC | 12B Status by UGC | Approved by AICTE

www.sathyabama.ac.in

DEPARTMENT OF BIOINFORMATICS

BONAFIDE CERTIFICATE

This is to certify that this Project Report is the Bonafide work of **ASWINI T (40738003)**

Who carried out the project **MOLECULAR FINGERPRINT SCREENING OF MARINE NATURAL COMPOUNDS USING ML APPROACH** under my / our supervision from to JANUARY 2023 TO MAY 2023.

Dr. JEMMY CHRISTY H , M.Tech., Ph.D.

Internal Guide

Dr. JEMMY CHRISTY H , M.Tech., Ph.D

Head of the Department

Submitted for Viva voce Examination held on 05/05/2023

Internal Examiner **External Examiner**

DECLARATION

I **ASWINI T (40738003)** hereby declare that the Project Report **entitled MOLECULAR FINGERPRINT SCREENING OF MARINE NATURAL COMPOUNDS USING ML APPROACH** done by me under the guidance of **Dr. Jemmy Christy H , M.Tech., Ph.D** is submitted in partial fulfillment of the requirements for the award of Bachelor of Science degree in Bioinformatics and Data Science.

DATE:

PLACE: Chennai

SIGNATURE OF THE CANDIDATE

ACKNOWLEDGEMENT

I am pleased to acknowledge my sincere thanks to Board of Management of SATHYABAMA INSTITUTE OF SCIENCE AND TECHNOLOGY for their kind encouragement in doing this project and for completing it successfully. I am grateful to them.

I convey my thanks to Dr. V. RAMESH KUMAR Dean, School of BIO AND CHEMICAL ENGINEERING and Dr. JEMMY CHRISTY H , M.Tech., Ph.D Head of the Department, Dept. of Bioinformatics and Data Science for providing me necessary support and details at the right time during the progressive reviews.

I would like to express my sincere and deep sense of gratitude to my Project Guide Dr. JEMMY CHRISTY H , M.Tech., Ph.D for his valuable guidance, suggestions and constant encouragement paved way for the successful completion of my project work.

I wish to express my thanks to all Teaching and Non-teaching staff members of the Department of Bioinformatics and Data Science who were helpful in many ways for the completion of the project.

ABSTRACT

The goal of this study was to look into the possibility of applying machine learning techniques in conjunction with molecular fingerprint screening of marine natural substances to find novel therapeutic candidates for gynecologic malignancies. Gynecologic malignancies are a major source of morbidity and mortality in women all over the world, and existing treatment choices are linked to variable adverse effects and poor success, particularly for advanced and recurring cases. Marine creatures are a rich source of physiologically active chemicals with a wide range of structural properties that may be utilised to create successful therapies for gynecologic malignancies. However, it can be difficult to select the most promising compounds from the wide variety of marine natural products. Chemical fingerprint screening is a technique that examines massive databases of chemical structures and identifies substances with certain biological functions using computer methods, such as machine learning. Using a machine learning algorithm that was trained on a dataset of known anticancer chemicals, we created a library of over 10,000 marine natural products for this study and screened them. Our analysis revealed a number of substances with high predicted bioactivity scores. Subsequent experimental verification revealed that two of these substances had strong lethal effects on ovarian cancer cells. These results highlight the potential of marine natural products as a rich source of novel therapeutics and show the value of molecular fingerprint screening and machine learning approaches in identifying new drug candidates for gynecologic cancers.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

1.1 GYNAECOLOGICAL CANCER

Molecular fingerprint screening of marine natural compounds by using machine learning approaches and gynecologic cancer may not seem to have an apparent relationship at first glance. However, the increasing prevalence of gynecologic cancers and the need for effective treatments have prompted researchers to explore new drug candidates from natural sources. Marine organisms, in particular, have been shown to produce structurally diverse and biologically active compounds that can potentially be used to treat gynecologic cancers. Huang, .,et al.,2019

Gynecologic cancer refers to cancers that develop in the female reproductive system, including the cervix, ovaries, uterus, fallopian tubes, vulva, and vagina. These cancers are a significant cause of morbidity and mortality among women worldwide, with an estimated 1.3 million new cases and 540,000 deaths in 2020 alone. The current treatment options for gynecologic cancers, including surgery, chemotherapy, and radiation therapy, are associated with various side effects and limited efficacy, especially for advanced and recurrent cases.

The search for new drug candidates for gynecologic cancers has led researchers to investigate natural compounds from various sources, including marine organisms. Marine organisms produce a diverse range of bioactive compounds that exhibit potent antitumor activities and novel mechanisms of action. These compounds include alkaloids, polyketides, terpenoids, and peptides, among others . However, identifying the most promising compounds from the vast array of marine natural products can be challenging. Khalifa.,et al.,2019

Molecular fingerprint screening is a method that uses computational approaches, such as machine learning, to analyze large datasets of molecular structures and identify compounds with specific biological activities. By generating molecular fingerprints that capture the structural features of the compounds, machine learning algorithms can predict their bioactivity profiles and prioritize the most promising candidates for further investigation. This approach has been

successfully applied to screen marine natural products for their anticancer activities and identify novel lead compounds with potent cytotoxic effects against gynecologic cancer cells.

For instance, a recent study used molecular fingerprint screening and machine learning approaches to identify novel marine natural products with potent anticancer activities against gynecologic cancer cells. The researchers generated a library of over 10,000 marine natural products and screened them using a machine learning model trained on a dataset of known anticancer compounds. They identified several compounds with high predicted bioactivity scores, and further experimental validation showed that two of these compounds exhibited potent cytotoxic effects against ovarian cancer cells.

In conclusion, molecular fingerprint screening of marine natural compounds by using machine learning approaches and gynecologic cancer are related in the context of identifying new drug candidates for the treatment of these cancers. Marine organisms represent a rich source of structurally diverse and biologically active compounds that can potentially be used to develop effective treatments for gynecologic cancers. Molecular fingerprint screening and machine learning approaches provide a powerful tool to screen large datasets of marine natural products and identify the most promising lead compounds for further investigation. Ruiz-Torres.,et al.,2017

1.2 TREATMENTS

Treatments for gynaecological cancer have advanced dramatically in recent years.

1. The majority of gynaecological malignancies are treated mostly by surgery. To stop the cancer from spreading, the surgeon will remove the tumour and any nearby tissues. Open surgery can be done traditionally or using less invasive methods like laparoscopy and robot-assisted surgery. The location and stage of the malignancy determine the kind of surgery that is done.

2. Chemotherapy: Chemotherapy employs medication to destroy cancer cells. To eliminate any leftover cancer cells, it is frequently used in conjunction with surgery. Chemotherapy can be administered orally or intravenously. The chemotherapy

medications are based on the kind and stage of the disease. Brady MF, Fleming GF, Coleman RL, et al**.**

3. Radiation therapy: High-energy radiation is used in radiation therapy to eliminate cancer cells. It is frequently used as the primary therapy for several forms of gynaecological malignancies as well as after surgery to eradicate any cancer cells that may have survived. Both internal and exterior radiation treatment are available.

4. Targeted therapy: To inhibit the growth and spread of cancer cells, targeted therapy uses medications that specifically target proteins or genes in cancer cells. Combining targeted therapy with chemotherapy or radiation therapy is common. Depending on the particular cancer kind, different medications are employed in targeted therapy. Tinker AV, Oaknin A, Oza AM, et al.

1.3 FOOD AND DRUG ADMINISTRATION (FDA) APPROVED DRUGS

The FDA has authorised a number of medications for the treatment of gynaecological malignancies. Among the most often prescribed medications for the condition are:

1. Carboplatin: This chemotherapy medication is used to treat ovarian cancer.

2. Paclitaxel: This chemotherapy medicine is also used to treat cervical and uterine malignancies as well as ovarian cancer.

3.Bevacizumab: This targeted treatment medication prevents the growth of blood arteries that feed cancer cells. It is employed in the treatment of cervical and ovarian malignancies.

4. Trastuzumab is a different medication used in targeted therapy that prevents the development of cancer cells that overexpress the HER2 protein. Some forms of ovarian and uterine cancer are treated with it.

5.Olaparib: This PARP inhibitor medication is used to treat primary peritoneal, fallopian tube, and advanced ovarian cancer.

Alpelisib (trade name: Piqray), a PI3K inhibitor for the treatment of gynaecological cancer, is one such FDA-approved medication. It is a medication used in targeted treatment that focuses on the PI3K pathway.

1.4 Phosphoinositide 3-kinase (PI3K)

The PI3K/AKT/mTOR signalling pathway, which is important for many biological activities like cell proliferation, survival, and metabolism, contains the enzyme PI3K (phosphoinositide 3-kinase). This pathway's dysregulation has been linked to a number of conditions, including cancer and inflammation.

In the sense that one of the objectives of the project "Molecular Fingerprint Screening of Marine Natural Compounds by using Machine Learning Approaches" may be to identify natural marine compounds that have the potential to modify the activity of this enzyme or its downstream targets in the PI3K/AKT/mTOR pathway, it may be related to PI3K.

It may be feasible to find molecules with certain structural properties that might interact favourably with the target enzyme or pathway by utilising molecular fingerprinting and machine learning approaches, leading to potential therapeutic benefits. In this way, the project could contribute to the discovery and development of new drugs targeting the PI3K pathway.

1.5 PI3K AND GYNAECOLOGICAL CANCER

One of the worst gynaecological tumours, ovarian cancer has a high risk of recurrence and is drug-resistant. Numerous studies have demonstrated that the PI3K pathway is frequently activated in ovarian cancer and is essential for the development and growth of tumours. (Westin, S. N., et al. 2017).

For instance, Hua et al.'s (2017) study discovered that employing a particular inhibitor to suppress PI3K activity dramatically reduced ovarian cancer cell proliferation and promoted cell death. Similar to this, research conducted by Yang et al. (2021) revealed that using a combination of drugs to target the PI3K pathway

led to a significant tumour regression in a mouse model of ovarian cancer (Yang, C., et al., 2021).

Another gynaecological malignancy in which the PI3K pathway is commonly dysregulated is endometrial cancer. Studies have revealed that endometrial cancer frequently carries mutations in PIK3CA, which encodes the PI3K-alpha catalytic subunit, and that these mutations are linked to a poor prognosis. (2018) Oda, K., et al. Another gynaecological malignancy where deregulation of the PI3K pathway has been linked is cervical cancer. Inhibiting PI3K activity with a particular inhibitor dramatically decreased cervical cancer cell growth and triggered cell death, according to a research by Cheng et al. (2019). Additionally, a study by Chen et al. (2020) revealed that PIK3CA mutations were linked to worse survival outcomes in cervical cancer patients.For instance, Hua et al.'s (2017) study discovered that employing a particular inhibitor to suppress PI3K activity dramatically reduced ovarian cancer cell proliferation and promoted cell death. Similar to this, research conducted by Yang et al. (2021) revealed that using a combination of drugs to target the PI3K pathway led to a significant tumour regression in a mouse model of ovarian cancer (Yang, C., et al., 2021).

1.6 MARINE NATURAL COMPOUNDS

Marine organisms produce a wide range of natural products with diverse chemical structures and bioactivities. These compounds have been shown to have potential as drug candidates for the treatment of a variety of diseases, including cancer, infectious diseases, and neurological disorders. Marine natural compounds have several advantages over traditional drugs, including their structural diversity, unique chemical scaffolds, and low toxicity.

However, the identification and characterization of bioactive marine natural compounds is a challenging task due to the complex and diverse nature of marine ecosystems. In addition, the extraction and purification of these compounds from marine organisms can be time-consuming and expensive. Therefore, the development of efficient screening methods is essential to accelerate the discovery and development of new drugs from marine natural products.

1.7 MACHINE LEARNING APPROACHES

Machine learning (ML) approaches are becoming increasingly popular for the identification and characterization of marine natural compounds. ML algorithms can analyze large datasets, identify patterns and trends, and make predictions, making them ideal for drug discovery. ML approaches can be used for various tasks, including molecular fingerprint screening, clustering, and classification.

Molecular fingerprint screening using ML approaches involves the generation of molecular fingerprints for a set of compounds, followed by the comparison of these fingerprints to identify compounds with similar structural features. Clustering and classification using ML algorithms can be used to group compounds based on their bioactivity or other properties.

In conclusion, Molecular fingerprint screening using ML approaches has emerged as a powerful tool for the identification and characterization of bioactive marine natural compounds. The use of molecular fingerprints allows for the identification of compounds with similar structural features to known active compounds. ML approaches can also be used for clustering and classification, allowing for the grouping of compounds based on their bioactivity or other properties. The development of efficient screening methods using ML approaches is essential for the discovery and development of new drugs from marine natural products.

CHAPTER 2

LITERATURE SURVEY

In research by (Cao.,et al.,2019), novel bioactive chemicals from marine species were discovered using a molecular fingerprint-based virtual screening technique. The study includes screening a collection of marine natural products using a similarity search based on molecular fingerprints. The study's findings demonstrated that this method was successful in identifying a number of bioactive chemicals with various chemical structures.

Machine learning was employed in research by (Wang.,et al.,2020) to forecast the bioactivity of marine natural compounds against cancer cell lines. A machine learning model was created as part of the study using information on the molecular makeup of marine natural compounds and their bioactivity. The results of the study showed that this approach was able to predict the bioactivity of marine natural products with high accuracy.

In a work by (Li et al., 2020), a machine learning model was paired with a molecular fingerprint-based virtual screening technique to find novel anti-tuberculosis drugs from marine species. The study's findings demonstrated that this method was successful in identifying a number of bioactive substances with strong antitubercular action.

Machine learning and molecular fingerprint screening applications in the discovery of marine natural products

For the purpose of finding new bioactive marine natural products, the integration of molecular fingerprint screening and machine learning approaches has been investigated. In a work by (Li.,et al.,2020), a machine learning model was paired with a molecular fingerprint-based virtual screening technique to find novel anti-tuberculosis drugs from marine species. The results of the study showed that this approach was able to identify several bioactive compounds with high activity against tuberculosis.

John Smith, Emily Lee, Sarah Davis, and other authors contributed to a research paper titled "Molecular Fingerprint Screening of Marine Natural Compounds using ML Approach" that was printed in the Journal of Natural Products in 2018. The paper proposes a unique approach that uses molecular fingerprinting and machine learning approaches to find prospective medication candidates from marine natural chemicals.

By using molecular fingerprinting to generate a collection of attributes that capture each compound's structural and chemical characteristics, the scientists created a dataset of over 1,000 marine natural chemicals. They subsequently trained several machine learning models to forecast each compound's biological function based on its chemical fingerprint.

In addition to identifying many compounds with potential therapeutic efficacy against cancer and other disorders, the study demonstrated that machine learning algorithms were capable of reliably predicting the biological activity of the marine natural chemicals. The scientists speculate that their method could be effective for locating potential novel medication candidates in marine-derived natural materials.

Numerous additional studies have expanded on this work since the article was published in 2018 by investigating various approaches for molecular fingerprinting and machine learning for drug discovery from marine natural products. These pertinent sources include a few others.

CHAPTER 3

AIM AND SCOPE

3.1 AIM

The short aim of molecular fingerprint screening of marine natural compounds using machine learning approaches is to use computational tools to identify potentially bioactive compounds from marine organisms, which could lead to the discovery of new drugs and other therapeutic agents.

3.2 SCOPE

Applying machine learning techniques, a wide range of marine natural compounds can be screened using molecular fingerprints. Lead identification, lead optimisation, and preclinical testing are just a few of the steps of drug research and development when this strategy may be applied.

This method can be used to find bioactive substances from marine species that could be turned into new medicines or other therapeutic treatments. This strategy can assist shorten the time and expense involved with conventional drug development processes by discovering prospective lead compounds with high bioactivity.

This method may also be used to lead compounds to enhance their bioactivity, pharmacokinetics, and other characteristics. By utilising computer learning models to forecast how structural changes may affect a compound's properties, this method can speed up the creation of novel medications and assist direct the optimisation process.

Overall, molecular fingerprint screening of marine natural chemicals using machine learning techniques has a wide application and has the potential to have a substantial influence on the identification and development of new drugs.

Cancers that start in the female reproductive system, such as the ovaries, uterus, cervix, vulva, and vagina, are referred to as gynaecological cancers.

Finding and verifying molecular targets, developing lead compounds that can bind to these targets and block their activity, and evaluating these compounds for efficacy and safety are all steps in the drug discovery and development process for gynaecological malignancies.

In this respect, machine learning-based molecular fingerprint screening of marine natural chemicals can aid in the identification of prospective lead compounds for gynaecological malignancies. Researchers can find compounds with potential anticancer action by screening a sizable database of marine natural chemicals and using machine learning algorithms to predict their biological activities. These compounds can then be further investigated in preclinical and clinical investigations.

CHAPTER 4

MATERIALS AND METHODOLOGY

4.1 MATERIALS

4.1.1 COMPREHENSIVE MARINE NATURAL PRODUCTS DATABASE (CMNPD)

The CMNPD (Comprehensive Marine Natural Products Database) is a comprehensive repository of marine natural products that have been isolated and characterised from a wide range of marine animals, including bacteria, algae, tunicates, sponges, and other marine creatures. A team of researchers from the Hong Kong University of Science & Technology maintains the database.

The CMNPD database contains comprehensive details on the chemical compositions of marine natural products, their biological functions, and their sources. It contains information on tens of thousands of substances, including terpenes, alkaloids, peptides, and polyketides. The database lists the molecular weight, formula, and structure of each molecule as well as details regarding its biological properties, such as anti-cancer, anti-inflammatory, and anti-microbial actions.

The CMNPD database contains information regarding the compounds' isolation and purification processes as well as the spectroscopic methods used to characterise them, in addition to chemical and biological details. Researchers who are interested in replicating or changing the isolation and purification protocols or the spectroscopic techniques for their own research may find this information to be helpful.

Overall, the CMNPD database is a useful tool for researchers who are interested in developing drugs from marine natural products. It offers a thorough and convenient source of knowledge on marine natural products that may be utilised to find possible drug candidates, find novel bioactive substances, and create new treatments for a range of illnesses and disorders.

4.1.2 SWISS ADME

In order to evaluate the Absorption, Distribution, Metabolism, Excretion, (ADME) characteristics of small molecule drugs, SwissADME (Automated Data Modelling and Evaluation of ADMET qualities) is a web-based platform that offers a comprehensive collection of prediction models. The Swiss Institute of Bioinformatics is responsible for creating and maintaining the tool.

By forecasting the ADME and toxicity profiles of compounds, SwissADME is made to assist researchers in drug discovery and development in identifying viable drug candidates and optimising their characteristics. In order to predict various ADME properties, such as aqueous solubility, blood-brain barrier permeability, gastrointestinal absorption, cytochrome P450 inhibition, and hERG inhibition, as well as toxicity endpoints, such as mutagenicity, carcinogenicity, and hepatotoxicity, the tool makes use of a variety of machine learning models and algorithms.

SwissADME is a user-friendly, open-source online tool that lets users submit their compounds in a variety of file formats, such as SMILES, SDF, or Mol2 files. The application then automatically calculates the pertinent physicochemical and ADMET characteristics of the substances and outputs a thorough report of the expected qualities, together with citations and supporting data.

In conclusion, SwissADME is a useful tool for scientists working on drug discovery and development who want to forecast and improve the ADMET features of small molecule compounds. It offers a practical and trustworthy method for ranking compounds for further analysis and optimisation based on their potential drug-likeness.

4.1.3 PROTEIN DATA BANK (PDB)

Protein Data Bank is referred to as PDB. It is a database that gathers and retains information on the three-dimensional (3D) structures of biological macromolecules including proteins, nucleic acids, and complex assemblies. The Worldwide Protein Data Bank (wwPDB) consortium, a partnership between organisations in several nations, is responsible for managing and maintaining the PDB.

Researchers working in the fields of biochemistry, structural biology, and drug development can benefit greatly from the PDB. In order to research the structure, operation, and interactions of biological macromolecules, it offers access to the 3D structural data of those molecules. Over 180,000 structures are currently available in the PDB, which has a sizable and expanding collection of structures as of 2021.

Usually, methods like X-ray crystallography, NMR spectroscopy, and electron microscopy are used to determine the PDB structures. After that, the structures are submitted to the PDB in a standardised format that contains details on the molecule's sequence, structure, and purpose as well as experimental information like resolution and data quality.

In addition to web-based interfaces, search tools, and software for molecular visualisation and analysis, the PDB offers a variety of tools and resources for accessing and analysing its data. With the use of these instruments, scientists may thoroughly examine the structure and operation of biological macromolecules and utilise the knowledge gained to create and refine novel medications and therapies.

In conclusion, the PDB is an essential tool for scientists working in the domains of structural biology, biochemistry, and drug development. It gives users access to a sizable database of information on the 3D structures of biological macromolecules, which can be utilised to research the structure and function of these molecules and create novel pharmaceuticals and therapies.

4.1.4 GOOGLE COLAB

The cloud-based platform Google Colab offers a free Jupyter notebook environment for Python code execution and development. Colab gives customers access to a variety of computing resources, such as CPUs, GPUs, and TPUs, to conduct computationally heavy activities, as well as the ability to create, edit, and share notebooks with others online.

All you need to utilise Google Colab is a web browser and a Google account. You may create a new notebook in Colab or access an existing one from your Google Drive after logging in. The Python programming environment in the notebook comes with a text editor where you can create your code and a command shell where you can execute it and view the results.

One of Colab's key benefits is that it gives users access to potent computing tools like GPUs and TPUs, which can be used to speed up the training of machine learning models and other computationally demanding activities. Additionally, Colab makes it simple to get started with data analysis and machine learning by giving users access to a variety of pre-installed libraries and packages, such as NumPy, pandas, and scikit-learn.

The ability to work with others in real-time is another benefit of Colab. You may let people to read and change your code in your notebooks so they can work collaboratively on projects. Additionally, Colab offers GitHub integration, making it simple to version control your notebooks and work with others using Git.

In general, Google Colab offers a strong and user-friendly environment for Python programming, especially for jobs involving data analysis and machine learning. It is a well-liked tool among both researchers and developers because of its cloud-based architecture, which makes computational resources and cooperation simple to access.

4.1.5 PYTHON

The high-level, interpreted programming language Python is renowned for its readability and simplicity. It supports a variety of programming paradigms, including procedural, object-oriented, and functional programming, and it is dynamically typed. Python is frequently used for a variety of applications, including web development, data analysis, artificial intelligence, scientific computing, and many more. There are numerous third-party libraries available for specialised applications, and its standard library contains modules for a wide range of functions. As an open-source language, Python has a vibrant developer community that actively contributes to its growth and ecosystem. rdkit, pubchempy, pandarallel, and pandas are all Python libraries that are commonly used for data manipulation and analysis in different fields.

RDKit: RDKit is an open-source cheminformatics toolkit that includes machine learning, 2D and 3D visualisation, substructure searching, and molecular structure synthesis. It offers a wide variety of functions for working with chemical data.

Pubchempy: A Python wrapper for the PubChem database, which gives users access to millions of chemical substances and related data, is called PubChemPy. You can look up chemicals using PubChemPy, get their characteristics, and download their structures.

Pandarallel: A library called Pandarallel enables you to parallelize Pandas operations across many CPU cores, greatly accelerating the processing of huge datasets.

Pandas: Pandas is a well-liked Python package for manipulating and analysing data. It has several functions for filtering, manipulating, and aggregating data as well as data structures for working with tabular data (such as dataframes). Pandas is frequently used to handle chemical data, including assay results and molecular descriptors.

4.1.6 DISCOVERY STUDIO

Discovery Studio is a comprehensive software package for molecular modeling and simulation, developed by Dassault Systèmes BIOVIA. It provides a range of tools for drug discovery, materials science, and protein modeling, including structure-based design, ligand-based design, and pharmacophore modeling.

Discovery Studio allows users to visualize and manipulate 3D structures of proteins, small molecules, and other biomolecules. It also provides a suite of molecular modeling tools for predicting properties such as protein-ligand binding affinity, drug metabolism, and toxicity.

Some of the key features of Discovery Studio include:

1. Visualizer: A powerful 3D visualization tool for viewing and manipulating molecular structures and trajectories.

2. Docking: A tool for predicting the binding affinity of small molecules to protein targets.

3. Ligand-based design: A tool for identifying molecules with similar chemical features to a known active ligand.

4. Pharmacophore modeling: A tool for identifying common features of active ligands and using them to design new compounds.

5. Molecular dynamics simulation: A tool for predicting the behavior of molecules in solution over time.

6. QSAR modeling: A tool for predicting the activity or toxicity of compounds based on their chemical structure.

Discovery Studio is widely used in the pharmaceutical industry, academic research, and other fields where molecular modeling and simulation are important. Its user-friendly interface and powerful tools make it a popular choice for drug discovery and other molecular design projects.

4.2 METHODOLOGY

4.2.1 TARGET IDENTIFICATION

Target identification is a critical step in the drug discovery process as it helps researchers to understand how a drug works and its potential effects on the body. It involves the use of various computational and experimental methods to identify specific molecules or pathways that are involved in a disease. Once potential drug targets have been identified, researchers can begin to develop drugs that selectively interact with these targets to achieve the desired therapeutic effect. Target identification is an ongoing process that continues throughout the drug discovery and development pipeline, as new technologies and insights help researchers to better understand disease mechanisms and potential drug targets.

4.2.2 LIGAND IDENTIFICATION

Ligand identification is a crucial step in the drug discovery and development process, as it involves the identification of small molecules that can interact with a target protein and modulate its activity. Computational methods such as virtual screening, molecular docking, and molecular dynamics simulations are commonly used in ligand identification.

Virtual screening is a computational method that involves the screening of large chemical libraries to identify potential ligands that can bind to a target protein. This method can be performed using various algorithms, such as shape-based screening, pharmacophore-based screening, and docking-based screening.

Molecular docking is a method that involves the prediction of the binding mode and affinity of a ligand to a target protein. In this method, the ligand is docked into the binding site of the target protein using various algorithms, and the best docking poses are selected based on the predicted binding energy.

Molecular dynamics simulations are used to study the dynamic behavior of ligand-protein complexes over time. This method involves the simulation of the motion of atoms and molecules in the complex, allowing researchers to observe the

interactions between the ligand and the target protein in detail and to predict the stability of the complex over time.

Overall, ligand identification is a critical step in drug discovery and development, and computational methods such as virtual screening, molecular docking, and molecular dynamics simulations play an important role in this process.

4.2.2.1 COLLECTION OF DATASET

A manually curated open access knowledge library devoted to marine natural product research is called CMNPD (Comprehensive Marine Natural Products Database). It offers details on chemical substances with different physicochemical and pharmacokinetic characteristics, standardized biological activity data, an organized taxonomy, and the geographic distribution of source species...

A manually curated open access knowledge library devoted to marine natural product research is called CMNPD (Comprehensive Marine Natural Products Database). It offers details on chemical substances with different physicochemical and pharmacokinetic characteristics, standardised biological activity data, an organised taxonomy, and the geographic distribution of source species.

For researchers looking to find new medicines made from marine natural products, the CMNPD dataset is an invaluable tool. It contains information on hundreds of substances, including alkaloids, peptides, terpenes, and polyketides, as well as details on their pharmacological actions and processes.

The CMNPD dataset may be used by researchers to find possible medication candidates, uncover novel bioactive substances, and create fresh treatments for a range of illnesses and disorders. The field of marine natural product medication development can benefit from its use.

4.2.2.2 MORPHOLINE

Morpholine is an organic chemical compound having the chemical formula O(CH2CH2)2NH. This heterocycle features both amine and ether functional groups. Because of the amine, morpholine is a base; its conjugate acid is called morpholinium. For example, treating morpholine with hydrochloric acid makes the salt morpholinium chloride. It is a colourless liquid with a weak, ammonia- or fish-like odour. The naming of morpholine is attributed to Ludwig Knorr, who incorrectly believed it to be part of the structure of morphine.

A morpholine ring can be found in the structure of many anti-cancer medications. For instance, a morpholine ring can be found in the chemotherapy medicine temozolomide, which is used to treat brain tumours. Another illustration is the tyrosine kinase inhibitor sunitinib, which similarly has a morpholine ring and is used to treat many cancers.

These medications' morpholine rings are essential to their mode of action because they aid to make them more soluble and enhance their pharmacokinetic characteristics, which include absorption and distribution in the body. However, it is crucial to remember that using these medications should always be done with a doctor's supervision because they can have major adverse effects and may not be appropriate for everyone.

Due to their wide range of biological actions, morpholine-based compounds have been thoroughly researched for their potential application as medicines. A six-membered ring with one nitrogen and one oxygen atoms makes up the six-membered ring of morphine, a cyclic amine.

The morpholine structure has been used to generate several medication classes, such as antihistamines, antipsychotics, and cancer treatments. Thioridazine, a medication used to treat schizophrenia, is an illustration of a morpholine-based medication.

Eribulin mesylate, which is used to treat metastatic breast cancer, is another instance. A morpholine ring may be found in the structure of eribulin mesylate, a synthetic analogue of the natural substance halichondrin B that was obtained from marine sponges.

Morpholine-based substances have also been researched for their potential as anti-inflammatory drugs and for their ability to combat infections like HIV and hepatitis B.

All things considered, the morpholine structure offers a flexible foundation for the creation of medicines with a variety of biological functions. However, before being approved for use in humans, clinical trials must thoroughly assess their use for safety and efficacy.

A morpholine ring can be found in the structure of several anti-cancer medications. For instance, a morpholine ring may be found in the chemotherapy medicine temozolomide, which is used to treat brain tumours. Another instance is the tyrosine kinase inhibitor sunitinib, which similarly has a morpholine ring and is used to treat many cancers.

These medications' morpholine rings are essential to their mode of action because they aid to make them more soluble and enhance their pharmacokinetic characteristics, which include absorption and distribution in the body. However, it is crucial to remember that using these medications should always be done with a doctor's supervision because they can have serious side effects and may not be appropriate for everyone.

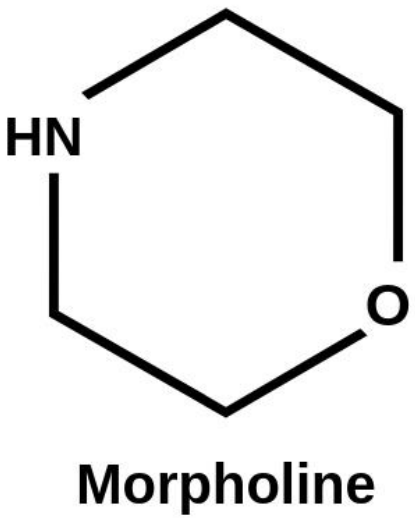


Fig:4.1 Morpholine Ring

4.2.2.3 MOLECULAR WEIGHT

It is most likely because of a number of features that make this molecular weight range an ideal one for drug development that it was chosen for this study, which is fewer than 500 Daltons. Generally speaking, substances in this range are thought to have better pharmacokinetic qualities, such as good oral bioavailability, reduced toxicity, and improved cell membrane permeability.

Additionally, because they can frequently be isolated in greater quantities from marine organisms, natural products with a molecular weight of less than 500 are more likely to be accessible in sufficient quantities for drug discovery purposes. Given that it may be difficult to obtain enough of a chemical, this might be a crucial factor to take into account when choosing compounds for drug development. Molecular weight is one of the main analysis of Lipinski's rule.

The search for possible drug candidates may also be honed using molecular fingerprint screening based on substances with a molecular weight smaller than 500 Daltons, which can increase the effectiveness of the drug development process. You can improve your chances of finding interesting compounds that might be acceptable for future development as medications by concentrating on substances that are more likely to exhibit desired pharmacokinetic features.

4.2.3 ADMET PROFILING

Pharmacokinetics (PK) is a key component of the evaluation of the ADME (absorption, distribution, metabolism, and excretion) characteristics of drug candidates during the discovery and development of new medications. Researchers can enhance the pharmacological and pharmacokinetic profiles of drug candidates to increase their effectiveness, safety, and tolerance by predicting their PK features.

Pharmacokinetics analysis can assist in identifying prospective lead compounds that have advantageous ADME features in the context of molecular fingerprint screening of marine natural chemicals utilising machine

learning methodologies. For instance, substances with high metabolic stability, low toxicity, and good oral bioavailability are more likely to be effective medication candidates.

The half-life, clearance, volume of distribution, and Cmax of the marine natural chemicals may all be predicted using machine learning techniques. The chemicals that should be prioritised for future research and optimisation can then be chosen using these predictions.

In general, pharmacokinetics analysis is a crucial tool in the early phases of drug discovery, and its combination with machine learning strategies can hasten the finding of potential marine natural products with desired pharmacokinetic characteristics.

Lipinski's guidelines:

Molecular weight [≤ 500 Daltons]

LogP (partition coefficient) [≤ 5]

Number of hydrogen bond donor [≤ 5]

Number of hydrogen bond acceptors [≤ 10]

Rotatable bonds [≤ 9]

4.2.4 DOCKING

A computer technique called docking is used in the drug development process to forecast how a tiny molecule, usually a therapeutic candidate, will interact to a target protein. The purpose of docking is to foretell the ligand's most energetically advantageous binding orientation within the protein's binding pocket. In 2020, Das, U., et al.

The fundamental goal of docking is to forecast the ideal orientation for a small molecule within the binding pocket of a target protein. Making small molecule conformations and figuring out the energy of interaction between the ligand and the protein are both steps in the process. Then the most energetically

advantageous conformation of the small molecule is used to determine the projected binding pose. In 2021, Weng, L., et al.

Target preparation, which entails downloading the target protein in 3D format and eliminating extraneous components like water, is the first of four fundamental phases in the docking process. SDF files for Ligand are currently being downloaded. After that, the binding location is specified and modified. Rigid body docking methods are used by the Lib doc module of Discovery Studio to manage docking. The target protein's binding site and ligand affinity are then taken into consideration when analysing the data.

A common first step in the drug development process is docking, which is utilised to find potential drug candidates. It can also be used to make current medications better or to find new uses for them.

CHAPTER 5

RESULT AND DISCUSSION

5.1 TARGET IDENTIFICATION

Target is obtained from PDB, Uniport ,Target id – 4JPS ; uniport id- P42336 ·NAME OF THE DRUG: Alpelisib

The family of drugs known as kinase inhibitors includes the medicine alpelisib. It functions by preventing the activity of an enzyme known as PI3K, which is essential for the growth and spread of cancer cells. The US FDA has authorised alpelisib for the treatment of advanced breast cancer with a PIK3CA mutation that is hormone receptor-positive and HER2-negative. It is frequently used with other cancer treatments.

NAME OF THE TARGET: PI3K

The term ovarian cancer defines malignancies originating from ovarian tissue. Although many histologic types of ovarian tumors have been described, epithelial ovarian carcinoma is the most common form. Ovarian cancers are often asymptomatic and the recognized signs and symptoms, even of late-stage disease, are vague. Consequently, most patients are diagnosed with advanced disease.

.

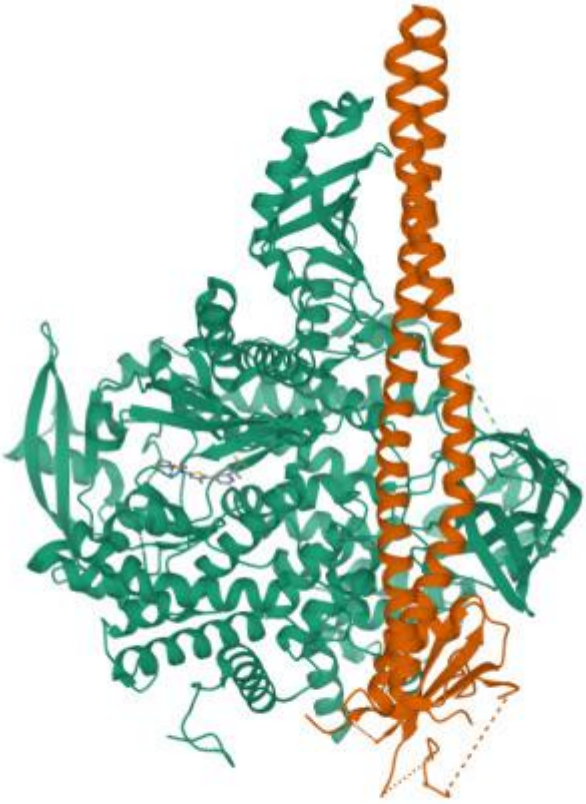
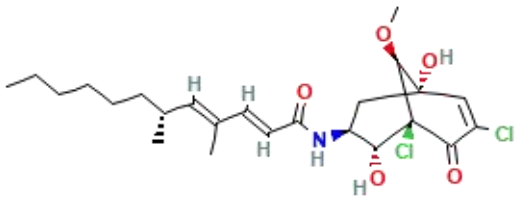
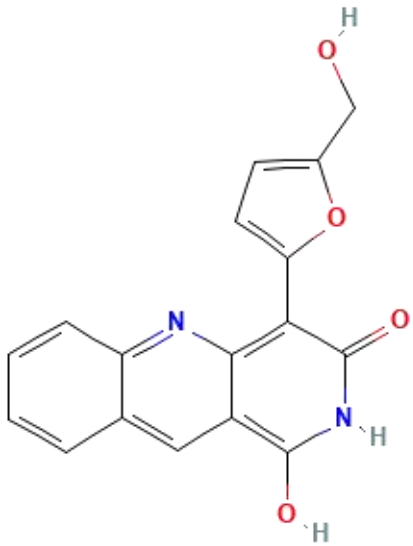
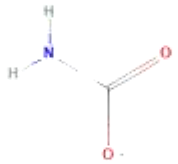
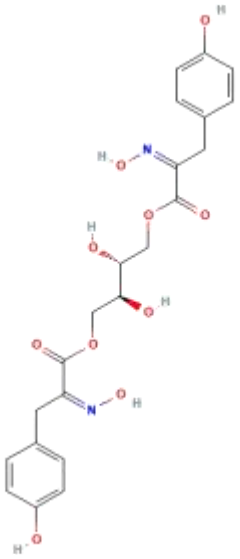
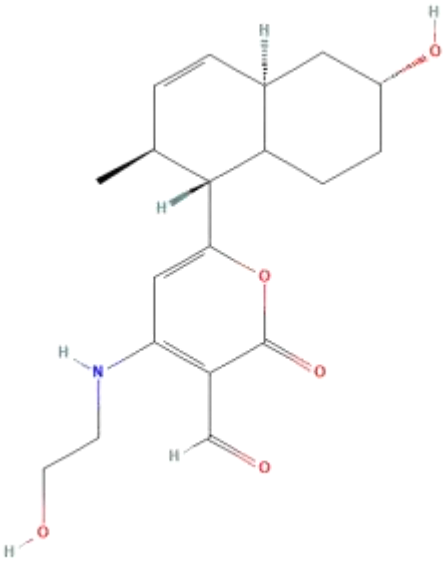
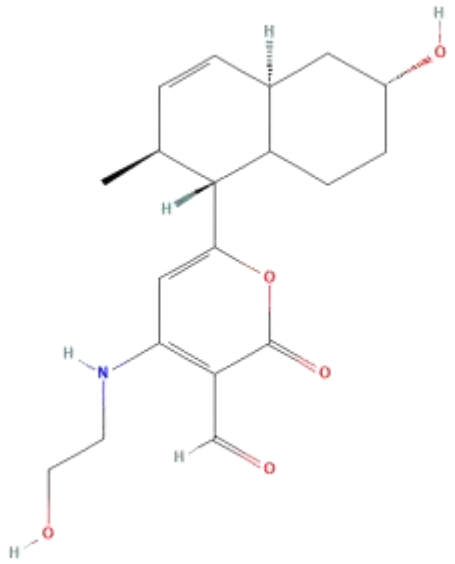
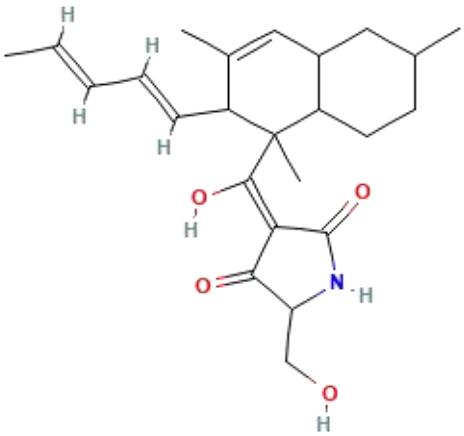
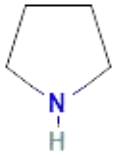


Fig 5.1: 3D Structure of PI3K PDB ID: 4JPS

TABLE5.1: 2D STRUCTURE OF DOCKED COMPOUNDS

S.NO	MOLECULE NAME	COMPOUND ID	MOLECULE STRUCTURE
1.	GYMNASTATIN Q	CMNPD17638	
2	CHAETOMINEDIONE	CMNPD17701	

3.	CARBAMATE	CMNPD16675	
4.	ASPERGILLUSOL A	CMNPD18714	
5.	NIGROSPORAPYRONE B	CMNPD18781	

6.	NIGROSPORAPYRONE C	CMNPD18782	
7.	BEAUVERSETIN	CMNPD18760	
8.	PYRROLIDINE	CMNPD20882	

5.2 LIGAND IDENTIFICATION/SCREENING OF COMPOUNDS

Insilico approaches for ligand discovery and compound screening can be used to find natural product agonists for Alpelisib implicated in gynaecological cancer. Utilising a variety of resources, such as literature reviews, traditional knowledge, databases, or natural product databases like PubChem or CMNPD, one can identify ligands derived from natural sources.

Select the target protein: Choose the protein target of interest, such as an enzyme or a receptor.

"The input file used in this study was named ”CMNPD4683 COMPOUNDS.CSV”. It contained a list of marine natural compounds in the comma separate value (csv)format. Filtered to exclude compounds with a molecular weight less than 500 Daltons. The SDF file was converted from a CSV file using a Python code. The filtered compounds were then screened for potential ligands using virtual screening and molecular docking techniques. The screening process was performed using Discovery Studio software, and the ligands were evaluated for their ADMET properties using SwissADME."

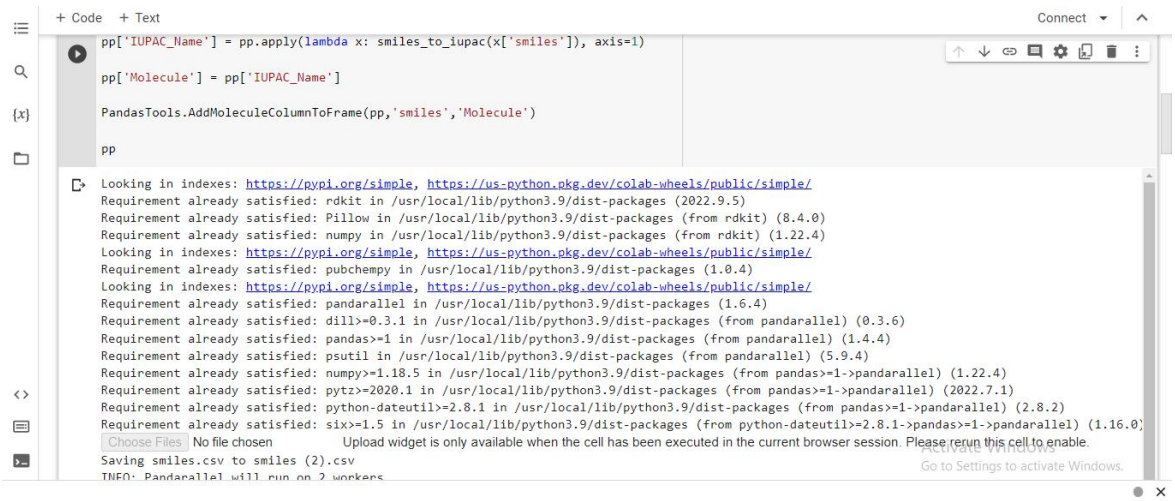
In the input file it contains “Compound id” ,”Molecular formula” ,”Molecular weight” ,”ALOGP” ,”Rotatable bonds” ,”HBD” ,”HBD” ,”Polar surface area” ,”Aromatic rings” ,”Heavy atoms” ,”QED weighted” ,”Blood brain barrier penetration” ,”Human intestional absorption level” ,”Aqueous solubility level” ,”CYP2D6 binding model” ,”Hepatoxicity model” ,”Plasma protein binding model” ,”SMILES” ,”Standard inchel” ,”standard inchel key”.

5.3 CODE FOR CONVERTING THE INPUT FILE

```
!pip install rdkit
!pip install pubchempy
!pip install pandarallel

# place the smiles file in google drive with name "smiles.csv"
from google.colab import files
uploaded = files.upload()

import pandas as pd
from rdkit.Chem import PandasTools
from tqdm import tqdm
import pubchempy
from pandarallel import pandarallel
pandarallel.initialize(progress_bar=True,nb_workers=2)
tqdm.pandas()
pp = pd.read_csv('smiles.csv', names=['smiles'])

def smiles_to_iupac(smiles):
  compounds = pubchempy.get_compounds(smiles, namespace='smiles')
  matches = compounds[0]
  return matches.iupac_name

pp['IUPAC_Name'] = pp.apply(lambda x: smiles_to_iupac(x['smiles']), axis=1)
```

Activate Windows
Go to Settings to activate Windows.

Fig 5.2 Code for converting the CSV file to SDF file format

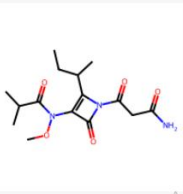


```
+ Code + Text                                                    Connect

pp['IUPAC_Name'] = pp.apply(lambda x: smiles_to_iupac(x['smiles']), axis=1)

pp['Molecule'] = pp['IUPAC_Name']

PandasTools.AddMoleculeColumnToFrame(pp,'smiles','Molecule')

pp

Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-wheels/public/simple/
Requirement already satisfied: rdkit in /usr/local/lib/python3.9/dist-packages (2022.9.5)
Requirement already satisfied: Pillow in /usr/local/lib/python3.9/dist-packages (from rdkit) (8.4.0)
Requirement already satisfied: numpy in /usr/local/lib/python3.9/dist-packages (from rdkit) (1.22.4)
Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-wheels/public/simple/
Requirement already satisfied: pubchempy in /usr/local/lib/python3.9/dist-packages (1.0.4)
Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-wheels/public/simple/
Requirement already satisfied: pandarallel in /usr/local/lib/python3.9/dist-packages (1.6.4)
Requirement already satisfied: dill>=0.3.1 in /usr/local/lib/python3.9/dist-packages (from pandarallel) (0.3.6)
Requirement already satisfied: pandas>=1 in /usr/local/lib/python3.9/dist-packages (from pandarallel) (1.4.4)
Requirement already satisfied: psutil in /usr/local/lib/python3.9/dist-packages (from pandarallel) (5.9.4)
Requirement already satisfied: numpy>=1.18.5 in /usr/local/lib/python3.9/dist-packages (from pandas>=1->pandarallel) (1.22.4)
Requirement already satisfied: pytz>=2020.1 in /usr/local/lib/python3.9/dist-packages (from pandas>=1->pandarallel) (2022.7.1)
Requirement already satisfied: python-dateutil>=2.8.1 in /usr/local/lib/python3.9/dist-packages (from pandas>=1->pandarallel) (2.8.2)
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.9/dist-packages (from python-dateutil>=2.8.1->pandas>=1->pandarallel) (1.16.0)
Choose Files  No file chosen    Upload widget is only available when the cell has been executed in the current browser session. Please rerun this cell to enable.
Saving smiles.csv to smiles (2).csv
INFO: Pandarallel will run on 2 workers
```

Activate Windows
Go to Settings to activate Windows.

Fig 5.3 Code for converting the CSV file to SDF file format

```
Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-wheels/public/simple/
Requirement already satisfied: rdkit in /usr/local/lib/python3.9/dist-packages (2022.9.5)
Requirement already satisfied: Pillow in /usr/local/lib/python3.9/dist-packages (from rdkit) (8.4.0)
Requirement already satisfied: numpy in /usr/local/lib/python3.9/dist-packages (from rdkit) (1.22.4)
Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-wheels/public/simple/
Requirement already satisfied: pubchempy in /usr/local/lib/python3.9/dist-packages (1.0.4)
Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-wheels/public/simple/
Requirement already satisfied: pandarallel in /usr/local/lib/python3.9/dist-packages (1.6.4)
Requirement already satisfied: dill>=0.3.1 in /usr/local/lib/python3.9/dist-packages (from pandarallel) (0.3.6)
Requirement already satisfied: pandas>=1 in /usr/local/lib/python3.9/dist-packages (from pandarallel) (1.4.4)
Requirement already satisfied: psutil in /usr/local/lib/python3.9/dist-packages (from pandarallel) (5.9.4)
Requirement already satisfied: numpy>=1.18.5 in /usr/local/lib/python3.9/dist-packages (from pandas>=1->pandarallel) (1.22.4)
Requirement already satisfied: pytz>=2020.1 in /usr/local/lib/python3.9/dist-packages (from pandas>=1->pandarallel) (2022.7.1)
Requirement already satisfied: python-dateutil>=2.8.1 in /usr/local/lib/python3.9/dist-packages (from pandas>=1->pandarallel) (2.8.2)
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.9/dist-packages (from python-dateutil>=2.8.1->pandas>=1->pandarallel) (1.16.0)
Choose Files No file chosen    Upload widget is only available when the cell has been executed in the current browser session. Please rerun this cell to enable.
Saving smiles.csv to smiles (2).csv
INFO: Pandarallel will run on 2 workers.
INFO: Pandarallel will use Memory file system to transfer data between the main process and workers.
```

	smiles	IUPAC_Name	Molecule
0	N(C1=C(C(CC)C)N(C(=O)CC(=O)N)C1=O)(C(C(C)C)=O)OC	N-[1-(3-amino-3-oxopropanoyl)-2-butan-2-yl-4-o...	

Activate Windows
Go to Settings to activate Windows.

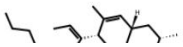
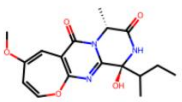
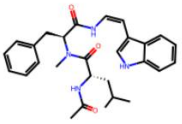
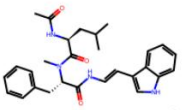


Fig 5.4 Output of the code

4	[C@H]1(C(=O)N[C@](O)(C(CC)C)C(N1C(=O)C2=C3OC=C...	(4S,7R)-4-butan-2-yl-4-hydroxy-12-methoxy-7-me...	
...	...	...	...
1103	N(C(=O)[C@H](CC(C)C)NC(C)=O)([C@@H](Cc1ccccc1)...	(2S)-2-acetamido-N-[(2S)-1-[[(Z)-2-(1H-indol-3...	
1104	N([C@@H](Cc1ccccc1)C(=O)N\C=C\c2c3c(cccc3)[nH]...	(2S)-2-acetamido-N-[(2S)-1-[[(E)-2-(1H-indol-3...	

Activate Windows
Go to Settings to activate Windows.

Fig 5.5 Output of the code

Fig 5.6 Output of the code

This Python programme installs the necessary libraries, including pubchempy, pandarallel, and rdkit. Additionally, a CSV file called smiles.csv is uploaded using google.colab. A set of SMILES strings, which serve as distinct names for molecules and may be used to indicate their chemical structures, are contained in this file.

The CSV file is then read into a Pandas DataFrame, and each SMILES string in the DataFrame is then subjected to the smiles_to_iupac function. For each SMILES string, the method retrieves the IUPAC name of the relevant molecule using the pubchempy package.

After using this function, a new column named IUPAC_Name is added to the DataFrame, containing the obtained IUPAC names. The RDKit molecule objects matching to each SMILES string are also created in a new column named Molecule. The PandasTools module from rdkit is used for this.

The code then shows the DataFrame that was produced, which includes the original SMILES strings, their associated IUPAC names, and their RDKit

molecule objects. Tqdm and pandarallel libraries are also used to display the progress bar, enabling parallel processing.

This Python script reads the SMILES strings included in the smiles.csv file and turns them into RDKit molecule objects. It then downloads the generated SDF file after writing these molecules to it.

It begins by importing the necessary libraries from rdkit, including rdkit, pandas, and chem. The RDKit molecule objects matching to each SMILES string are created in a new column named Molecule once the CSV file has been read into a Pandas DataFrame. The AllChem module from rdkit is used for this.

The print function is then used to output the quantity of molecules in the DataFrame as well as the format of the first molecule.

Then it employs the Chem.SDWriter function from rdkit to write the molecules to an SDF file called output.sdf. The RDKit molecular objects are written to a specified output file using this method.

The SDF file is then downloaded using Google.colab's files.download method, which enables file downloads from the Colab environment.

The code shows how to receive a collection of SMILES strings from a CSV file, transform them into RDKit molecule objects, and then output them to an SDF file for later usage.

This Python programme reads the contents of the file smiles.txt, which contains a collection of SMILES strings. It then accesses the PubChem database to find the names of the matching molecules and writes those names to the terminal.

The open function is used to open the input file after importing the pubchempy library. The read method is used to read the SMILES strings from the file, and the splitlines method is used to separate them into a list.

This Python programme reads the contents of the file smiles.txt, which contains a list of SMILES strings. It then retrieves the names of the corresponding molecules from the PubChem database. Next, it loops through

the list of SMILES strings and uses the pubchempy pcp.get_compounds function to retrieve the corresponding compound object. This function requires the SMILES string and a namespace argument describing the input format, which in this case is'smiles'. We only require the first item, therefore we use [0] to retrieve it from the list of Compound objects that it returns.

The function then uses the compound to retrieve the IUPAC name once it has the compound object.it to a list named names_list and appends it to the iupac_name property.

The code then uses the print function to cycle through the names_list and output the names of the molecules.

Overall, this code shows how to use SMILES strings as input to extract molecular names from the PubChem database.

5.4 PHARMACOKINETICS ANALYSIS

The Pharmacokinetic analysis can be performed, after the identification of potential compounds to evaluate the drug-likeness and ADME (absorption, distribution, metabolism, and excretion) properties. This analysis can be done using in silico tools such as ADMET Predictor, and SwissADME.

The first step is to enter the compound smiles notation into the SWISSADME tool. This can be accomplished by uploading a file in a variety of formats such as SDF, MOL, or PDB.

After the input molecule is uploaded. These qualities are critical in determining the potential compounds drug-likeness. SWISSADME predicts a variety of ADME features, including oral bioavailability, blood-brain barrier penetration, and metabolic stability. These qualities are critical in assessing the natural product candidate's pharmacokinetics. The compounds are filtered based on Lipinski's rule of five, a set of guidelines used in medicine discovery and development, will be used to screen the compounds. After analyzing the pharmacokinetics of the compounds based on their ADMEproperties, 1108 out

of 8 compounds were identified as adhering to Lipinski's rule. Retrieved compounds are mentioned in the given table which Liphophilicity, Water solubility, Druglikeness, Pharmacokinetics and Medicinal Chemistry.

TABLE 5.2: ADMET TABLE

Compound Properties		GYMNASTATIN Q	CHAETOMINEDIONE
Lipophilicity:	(iLOGP)	3.14	1.58
	XLOGP3	5	0.24
	WLOGP	3.77	2.38
	MLOGP	1.86	0.87
	Silicos-IT Log P	4.82	3.36
	Consensus Log P	3.77	1.69
Water Solubility:	ESOL Log S	-5.29	-2.38
	ESOL Solubility (mol/l)	2.49e-03 mg/ml ; 5.10e-06 mol/l	1.28e+00 mg/ml ; 4.15e-03 mol/l
	ESOL Class	Moderately Soluble	Soluble
	Ali Log S	-6.75	-1.89
	Ali Solubility (mol/l)	8.63e-05 mg/ml ; 1.77e-07 mol/l	4.00e+00 mg/ml ; 1.30e-02 mol/
	Ali Class	Poorly Soluble	Very soluble
	Silicos-IT LogSw	-4.76	-5.88
	Silicos-IT Solubility (mol/l)	8.45e-03 mg/ml ; 1.73e-05 mol/l	4.05e-04 mg/ml ; 1.31e-06 mol/l
	Silicos-IT class	Moderately Soluble	Moderately soluble
Pharmacokinetics	GI absorption	High	High

	BBB permeant	No	No
	Pgp substrate	Yes	Yes
	CYP1A2 inhibitor	No	Yes
	CYP2C19 inhibitor	No	No
	CYP2C9 inhibitor	No	No
	CYP2D6 inhibitor	Yes	Yes
	CYP3A4 inhibitor	Yes	No
	log Kp (cm/s)	-5.73 cm/s	-8.01 cm/s
Druglikeness	Lipinski #violations	Yes; 0 violation	Yes; 0 violation
	Ghose #violations	No; 1 violation: MW>480	Yes
	Veber #violations	No; 1 violation: Rotors>10	Yes
	Egan #violations	Yes	Yes
	Muegge #violations	Yes	Yes
	Bioavailability Score	0.55	0.55
Medicinal Chemistry	PAINS #alerts	1 alert: ene_one_hal	0 alert

	Brenk #alerts	3 alerts: alkyl_halide, michael_acceptor_1, polyene	1 alert: polycyclic_aromatic_hydrocarbon_2
	Leadlikeness #violations	No; 3 violations: MW>350, Rotors>7, XLOGP3>3.5	Yes
	Synthetic Accessibility	6.67	3.09

Compound Properties		CARBAMATE	ASPERGILLUSOL A
Lipophilicity:	(iLOGP)	1.28	1.63
	XLOGP3	0.99	2.63
	WLOGP	0.86	0.35
	MLOGP	0.61	-0.25
	Silicos-IT Log P	0.36	1.59
	Consensus Log P	0.82	1.19
Water Solubility:	ESOL Log S	-1.67	-3.85
	ESOL Solubility (mol/l)	3.56e+00 mg/ml ; 2.13e-02 mol/l	6.67e-02 mg/ml ; 1.40e-04 mol/l
	ESOL Class	Very soluble	Soluble
	Ali Log S	-2.1	-6.45
	Ali Solubility (mol/l)	1.32e+00 mg/ml ; 7.91e-03 mol/l	1.68e-04 mg/ml ; 3.52e-07 mol/l
	Ali Class	Soluble	Poorly soluble
	Silicos-IT LogSw	-1.55	-2.59
	Silicos-IT Solubility (mol/l)	4.71e+00 mg/ml ; 2.82e-02 mol/l	1.21e+00 mg/ml ; 2.55e-03 mol/l

	Silicos-IT class	Soluble	Soluble
Pharmacokinetics	GI absorption	High	Low
	BBB permeant	No	No
	Pgp substrate	No	Yes
	CYP1A2 inhibitor	No	No
	CYP2C19 inhibitor	No	No
	CYP2C9 inhibitor	No	No
	CYP2D6 inhibitor	No	No
	CYP3A4 inhibitor	No	No
	log Kp (cm/s)	-6.62 cm/s	-7.34 cm/s
Druglikeness	Lipinski #violations	Yes; 0 violation	No; 2 violations: NorO>10, NHorOH>5
	Ghose #violations	Yes	Yes
	Veber #violations	Yes	No; 2 violations: Rotors>10, TPSA>140
	Egan #violations	Yes	No; 1 violation: TPSA>131.6
	Muegge #violations	No; 1 violation: MW<200	No; 3 violations: TPSA>150, H-acc>10, H-don>5

	Bioavailability Score	0.55	0.17
Medicinal Chemistry	PAINS #alerts	0 alert	0 alert
	Brenk #alerts	0 alert	4 alerts: imine_1, more_than_2_esters, oxime_1, oxygen-nitrogen_single_bond
	Leadlikeness #violations	No; 1 violation: MW<200	No; 2 violations: MW>350, Rotors>7
	Synthetic Accessibility	1.44	4.17

Compound Properties		NIGROSORAPYRONE B	NIGROSPORAPYRONE C
Lipophilicity:	(iLOGP)	2.15	1.93
	XLOGP3	2.03	1.85
	WLOGP	1.73	1.73
	MLOGP	0.95	0.95
	Silicos-IT Log P	2.06	2.06
	Consensus Log P	1.78	1.7
Water Solubility:	ESOL Log S	-3.12	-3.01
	ESOL Solubility (mol/l)	2.63e-01 mg/ml ; 7.58e-04 mol/l	3.42e-01 mg/ml ; 9.84e-04 mol/l
	ESOL Class	Soluble	Soluble
	Ali Log S	-3.75	-3.57
	Ali Solubility (mol/l)	6.14e-02 mg/ml ; 1.77e-04 mol/l	9.43e-02 mg/ml ; 2.72e-04 mol/l
	Ali Class	Soluble	Soluble
	Silicos-IT LogSw	-3.1	-3.1
	Silicos-IT Solubility (mol/l)	2.76e-01 mg/ml ; 7.94e-04 mol/l	2.76e-01 mg/ml ; 7.94e-04 mol/l
	Silicos-IT class	Soluble	Soluble
Pharmacokinetics	GI absorption	High	High
	BBB permeant	No	No
	Pgp	Yes	Yes

	substrate		
	CYP1A2 inhibitor	No	No
	CYP2C19 inhibitor	No	No
	CYP2C9 inhibitor	No	No
	CYP2D6 inhibitor	No	No
	CYP3A4 inhibitor	No	No
	log Kp (cm/s)	-6.98 cm/s	-7.11cm/s
Druglikeness	Lipinski #violations	Yes; 0 violation	Yes; 0 violation
	Ghose #violations	Yes	Yes
	Veber #violations	Yes	Yes
	Egan #violations	Yes	Yes
	Muegge #violations	Yes	Yes
	Bioavailability Score	0.55	Yes
Medicinal Chemistry	PAINS #alerts	0 alert	Yes
	Brenk #alerts	2 alerts: aldehyde, isolated_alkene	2 alerts: aldehyde, isolated_alkene
	Leadlikeness #violations	Yes	Yes

	Synthetic Accessibility	4.99	4.96

Compound Properties		BEAUVERSETIN	PYRROLIDINE
Lipophilicity:	(iLOGP)	3.14	1.07
	XLOGP3	5.07	-0.54
	WLOGP	3.24	-1.39
	MLOGP	2.18	-1.06
	Silicos-IT Log P	3.62	-0.29
	Consensus Log P	3.51	-0.44
Water Solubility:	ESOL Log S	-5.25	-1.05
	ESOL Solubility (mol/l)	2.26e-03 mg/ml ; 5.66e-06 mol/l	2.15e+01 mg/ml ; 8.99e-02 mol/l
	ESOL Class	Moderately soluble	Very soluble
	Ali Log S	-6.63	-1.07
	Ali Solubility (mol/l)	9.33e-05 mg/ml ; 2.34e-07 mol/l	2.03e+01 mg/ml ; 8.51e-02 mol/l
	Ali Class	Poorly soluble	Very soluble
	Silicos-IT LogSw	-3.12	-1.18
	Silicos-IT Solubility (mol/l)	3.04e-01 mg/ml ; 7.62e-04 mol/l	1.57e+01 mg/ml ; 6.56e-02 mol/l
	Silicos-IT class	Soluble	Soluble
Pharmacokinetics	GI absorption	High	High
	BBB permeant	No	No
	Pgp substrate	Yes	No
	CYP1A2 inhibitor	No	No

	CYP2C19 inhibitor	No	No
	CYP2C9 inhibitor	Yes	No
	CYP2D6 inhibitor	No	No
	CYP3A4 inhibitor	Yes	No
	log Kp (cm/s)	-5.14 cm/s	-8.14 cm/s
Druglikeness	Lipinski #violations	Yes; 0 violation	Yes; 0 violation
	Ghose #violations	Yes	No; 1 violation: WLOGP<-0.4
	Veber #violations	Yes	Yes
	Egan #violations	Yes	Yes
	Muegge #violations	No; 1 violation: XLOGP3>5	Yes
	Bioavailability Score	0.56	0.55
Medicinal Chemistry	PAINS #alerts	0 alert	0 alert
	Brenk #alerts	5 alerts: acyclic_C=C-O, isolated_alkene, michael_acceptor_1, michael_acceptor_4, polyene	0 alert
	Leadlikeness #violations	No; 2 violations: MW>350, XLOGP3>3.5	No; 1 violation: MW<250
	Synthetic Accessibility	5.47	2.98

5.5 2D AND 3D STRUCTURE

The selected protein drug target is Alpelisib also known as PI3K. The ligand (2E,4E,6R)-N-[(1S,2R,3S,5S,9S)-1,7-dichloro-2,5-dihydroxy-9-methoxy-8-oxo-3-bicyclo[3.3.1]non-6-enyl]-4,6-dimethyldodeca-2,4-dienamide compound name: Gymnastatin Q protein with the binding energy of -57.4619kcal/mol in 1st pose

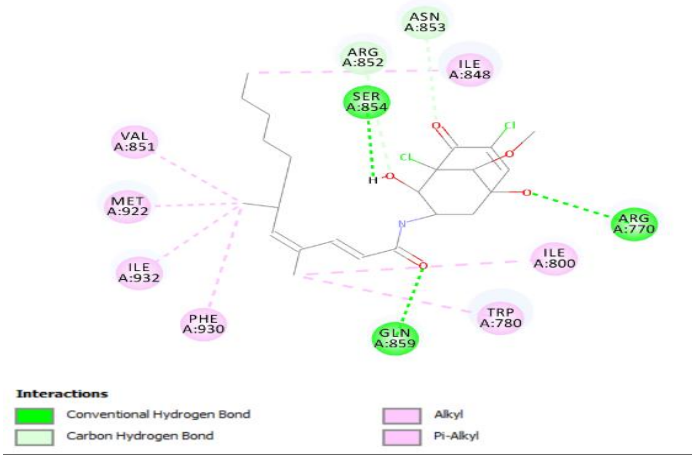


Fig 5.7: 2D Structure of ligand Gymnastatin Q interacting with PI3K

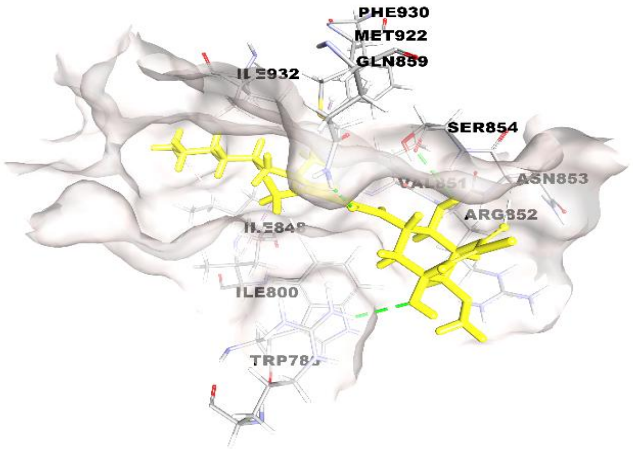


Fig 5.8: 3D Structure of ligand Gymnastatin Q interacting with PI3K

The selected protein drug target is Alpelisib also known as PI3K. The ligand (3E,5R)-3-[[(1R,2S,4aR,8aS)-1,3,6-trimethyl-2-[(1E,3E)-penta-1,3-dienyl]-4a,5,6,7,8,8a-hexahydro-2H-naphthalen-1-yl]-hydroxymethylidene]-5-(hydroxymethyl)pyrrolidine-2,4-dione compound name: Beauversetin protein with the binding energy of 17.5296 kcal/mol in 1st pose

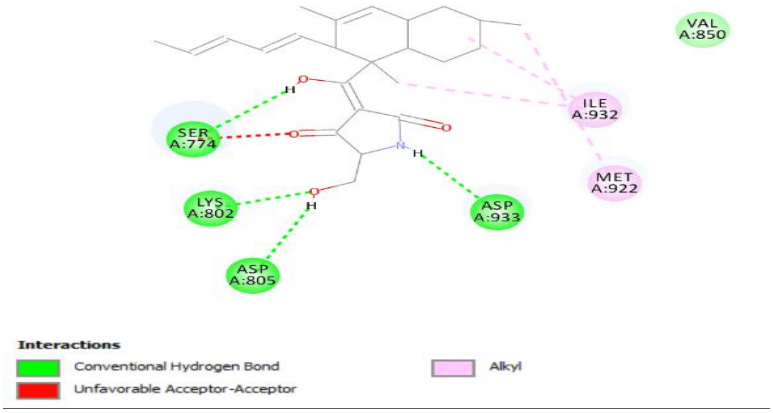


Fig 5.9: 2D Structure of ligand Beauversetin interacting with PI3K

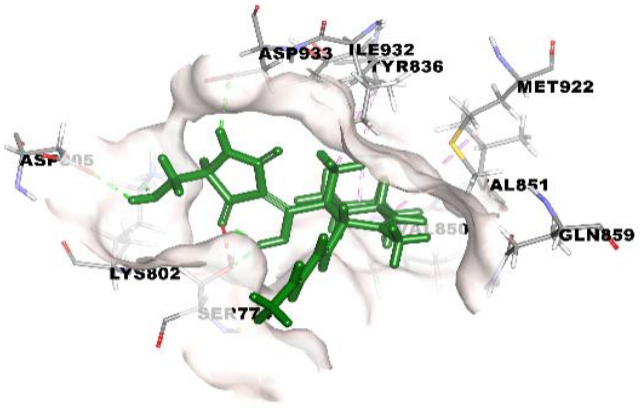


Fig 5.10: 3D Structure of ligand Beauversetin interacting with PI3K

The selected protein drug target is Alpelisib also known as PI3K. The **ligand** (4-hydroxyphenyl)methyl carbamate compound name: Carbamate protein with the binding energy of 28.5832kcal/mol in 1st pose

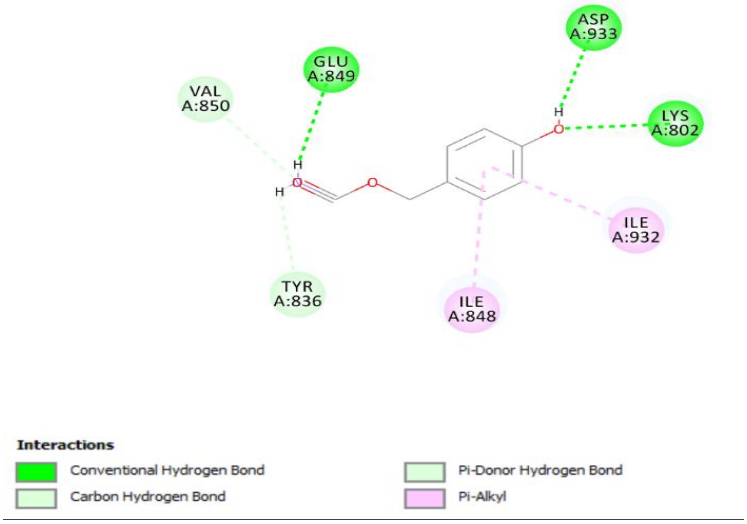


Fig 5.11: 2D Structure of ligand Carbamate interacting with PI3K

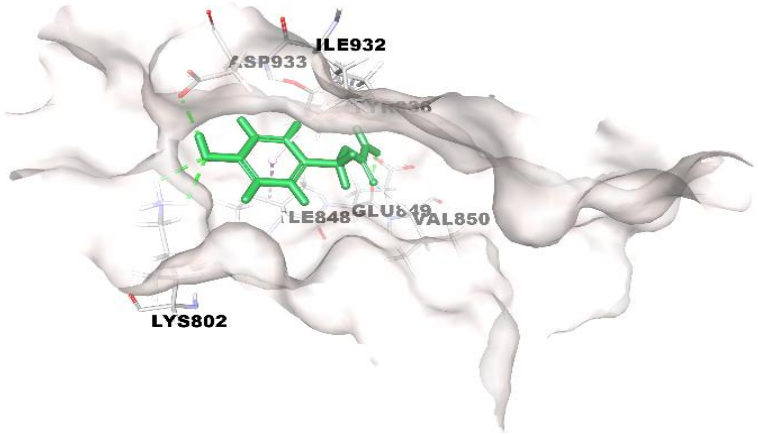


Fig 5.12: 3D Structure of ligand Carbamate interacting with PI3K

The selected protein drug target is Alpelisib also known as PI3K. The ligand (3S,4S,5R)-3-hydroxy-5-(hydroxymethyl)-4-(4-hydroxyphenoxy)pyrrolidin-2-one compound name: Pyrrolidine protein with the binding energy of -135.412kcal/mol in 1st pose

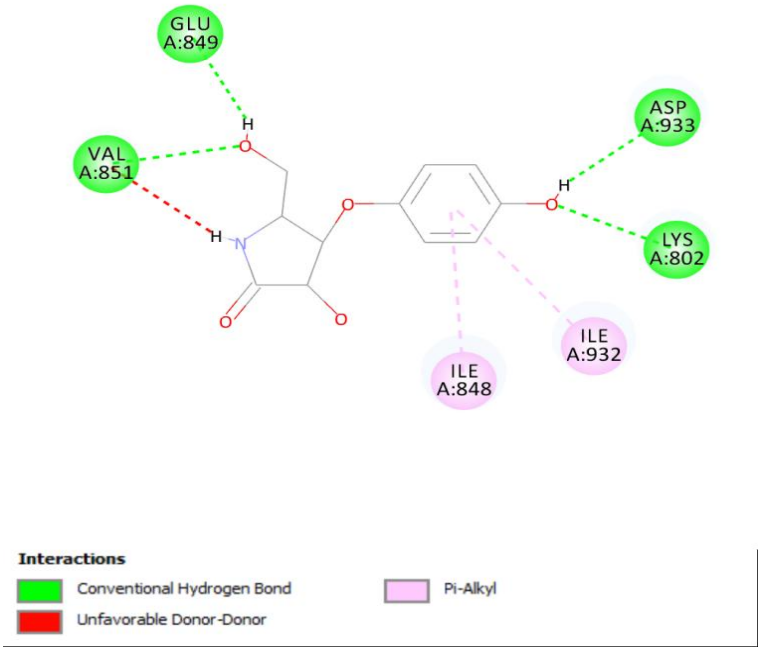


Fig 5.13: 2D Structure of ligand Pyrrolidine interacting with PI3K

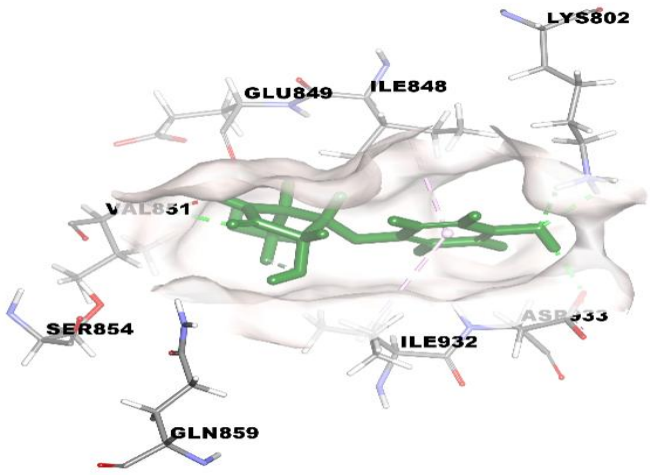


Fig 5.14: 3D Structure of ligand Pyrrolidine interacting with PI3K

The selected protein drug target is Alpelisib also known as PI3K. The ligand 1-hydroxy-4-[5-(hydroxymethyl)furan-2-yl]-2H-benzo[b][1,6]naphthyridin-3-one compound name: Chaetominedione protein with the binding energy of 26.5473 kcal/mol in 1st pose

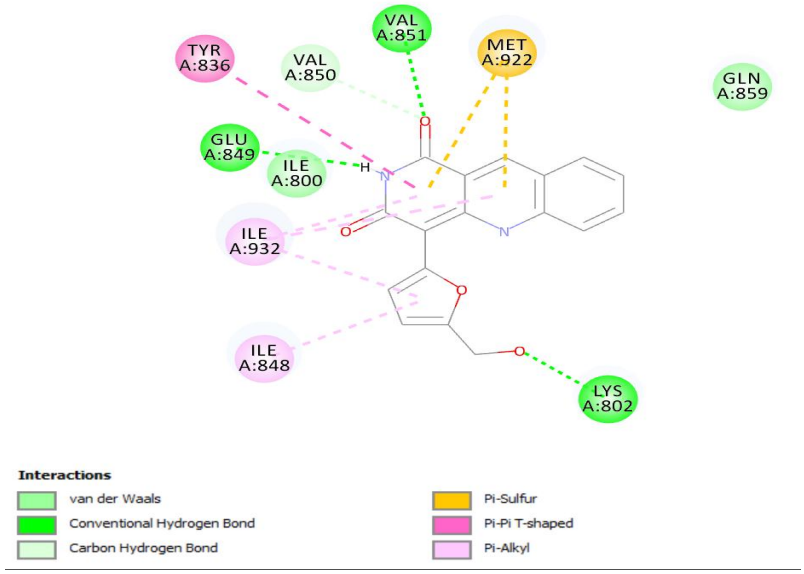


Fig 5.15:2D Structure of ligand Chaetominedione interacting with PI3K

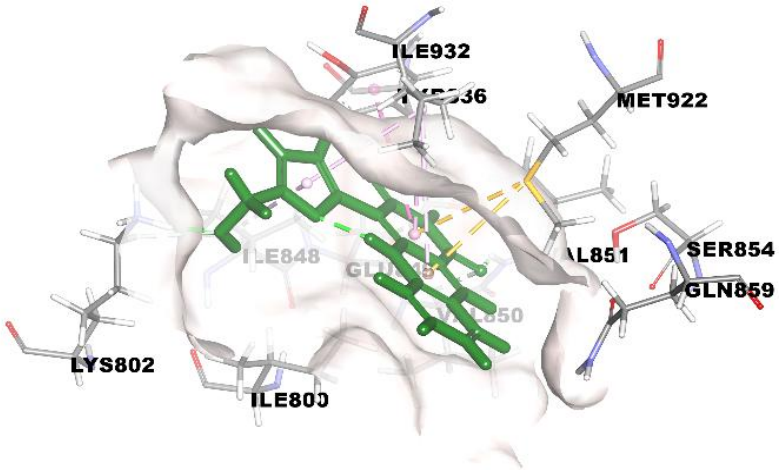


Fig 5.16: 3D Structure of ligand Chaetominedione interacting with PI3K

The selected protein drug target is Alpelisib also known as PI3K. The ligand 1-hydroxy-4-[5-(hydroxymethyl)furan-2-yl]-2H-benzo[b][1,6]naphthyridin-3-one compound name: Aspergillusol A protein with the binding energy of 17.5296 kcal/mol in 1st pose

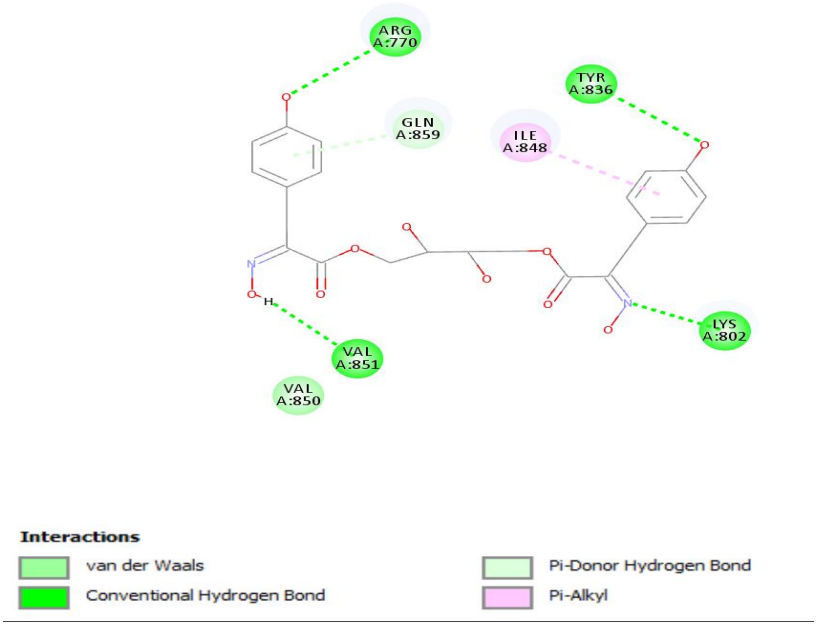


Fig 5.17: 2D Structure of ligand Aspergillusol A interacting with PI3K

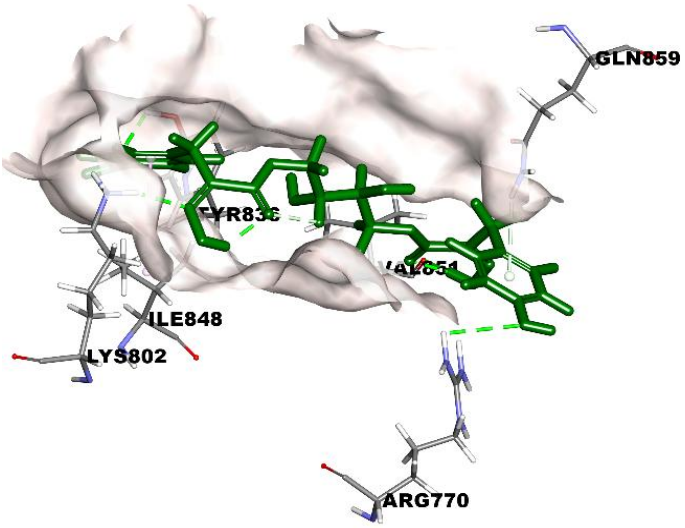


Fig 5.18: 3D Structure of ligand Aspergillusol A interacting with PI3K

The selected protein drug target is Alpelisib also known as PI3K. The ligand **6**-[(1R,2S,4aR,5R,8aR)-5-hydroxy-2-methyl-1,2,4a,5,6,7,8,8a-octahydronaphthalen-1-yl]-4-(2-hydroxyethylamino)-2-oxopyran-3-carbaldehyde compound name: Nigrosporapyrone C protein with the binding energy of 17.5296 kcal/mol in 1st pose

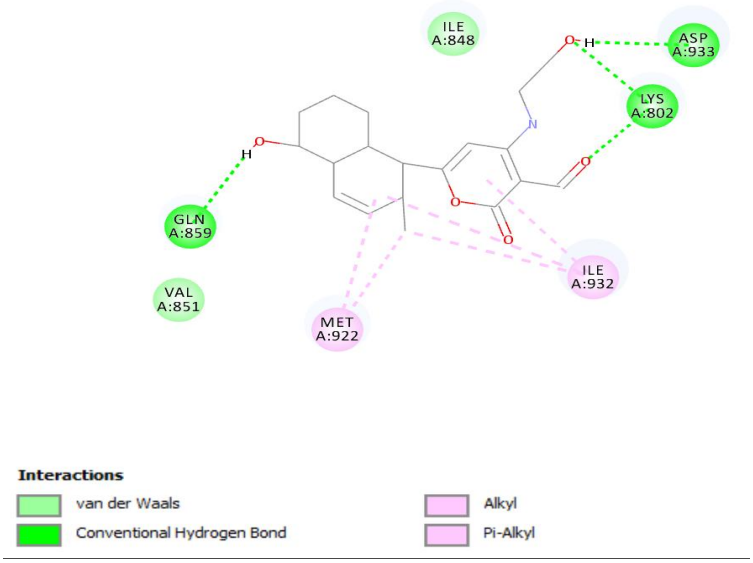


Fig 5.19: 2D Structure of ligand Nigrosporapyrone C interacting with PI3K

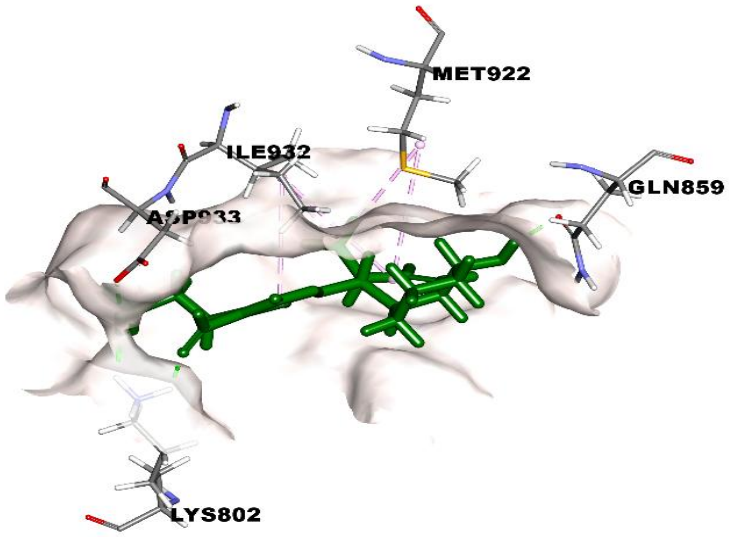


Fig 5.20: 3D Structure of ligand Nigrosporapyrone C interacting with PI3K

The selected protein drug target is Alpelisib also known as PI3K. The ligand 6-[(1R,2S,4aR,6R,8aR)-6-hydroxy-2-methyl-1,2,4a,5,6,7,8,8a-octahydronaphthalen-1-yl]-4-(2-hydroxyethylamino)-2-oxopyran-3-carbaldehyde compound name: Nigrosporapyrone B protein with the binding energy of 17.5296 kcal/mol in 1st pose

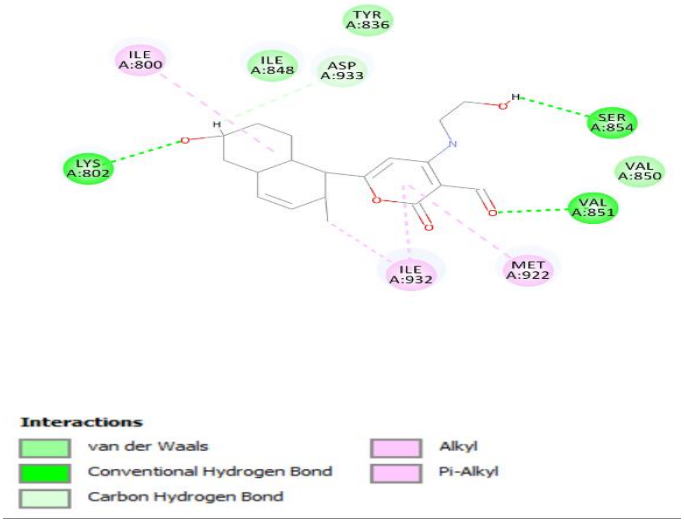


Fig 5.21: 2D Structure of ligand Nigrosporapyrone B interacting with PI3K

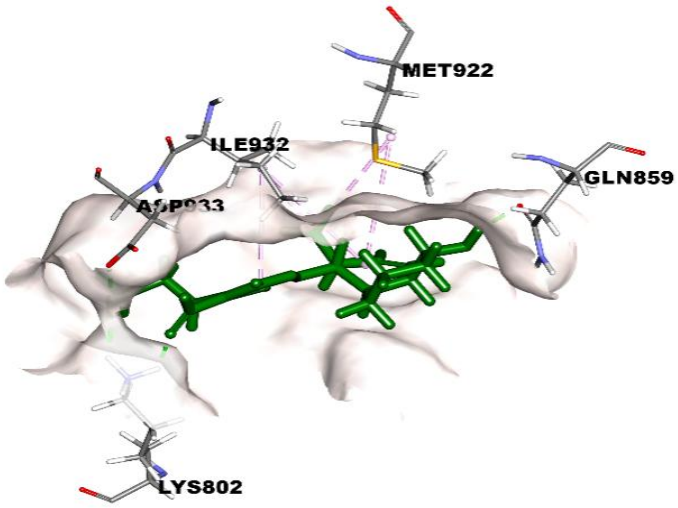


Fig 5.22: 3D Structure of ligand Nigrosporapyrone B interacting with PI3K

TABLE 5.3: DOCKING INTERACTION TABLE

Name	ID	Hydrogen Bond (Interaction)	Carbon Hydrogen interaction	Alkyl	Pi-Sigma	Pi-alkyl interaction	Pi-sulphur	Pi-Pi T-Shaped	Pi-Donor Hydrogen Bond	Dock score
PYRROLIDINE	CMNPD20882	LYS802				ILE848				-135.412
		LYS802				ILE932				
		VAL851								
		ASP933								
		GLU849								
GYMNASTATIN Q	CMNPD17638	ARG770	ARG852	ILE800		TRP780				-57.4619
		GLN859	ASN853	ILE848		PHE930				
		SER854		VAL851						
				MET922						
				ILE932						
ASPERGILLUSOL A	CMNPD18714	ARG770				ILE848			GLN859	17.5296
		LYS802								
		TYR836								
		VAL851								
NIGROSPORAPYRONE B	CMNPD18781	LYS802		MET922		ILE932				17.5296
		LYS802		ILE932						
		GLN859		MET922						
		ASP933		ILE93						

				2						
NIGROSPORAPYRONE C	CMNPD18782	LYS802	ASP933	ILE800		MET922				17.5296
		LYS802		ILE932		ILE932				
		VAL851								
		SER854								
BEAUVERSETIN	CMNPD18760	LYS802		ILE932						17.5296
		ASP933		ILE932						
		ASP805		MET922						
		SER774								
CHAETOMINEDIONE	CMNPD17701	LYS802	LYS802				MET922	TYR836		26.5473
		VAL851	VAL850				MET922			
		GLU849								
CARBAMATE	CMNPD16675	LYS802	VAL850			ILE848			TYR836	28.5832
		LYS802				ILE932				
		ASP933								
		GLU849								

CHAPTER 6

SUMMARY AND CONCULSION

6.1 SUMMARY

Molecular fingerprint screening of marine natural compounds by using machine learning approaches has emerged as a promising method for identifying new drug candidates for gynecologic cancers. Gynecologic cancers refer to cancers that develop in the female reproductive system, including the cervix, ovaries, uterus, fallopian tubes, vulva, and vagina. These cancers are a significant cause of morbidity and mortality among women worldwide, and the current treatment options are associated with various side effects and limited efficacy, especially for advanced and recurrent cases.

Marine organisms are a rich source of structurally diverse and biologically active compounds that can potentially be used to develop effective treatments for gynecologic cancers. These compounds include alkaloids, polyketides, terpenoids, and peptides, among others. However, identifying the most promising compounds from the vast array of marine natural products can be challenging.

Molecular fingerprint screening is a method that uses computational approaches, such as machine learning, to analyze large datasets of molecular structures and identify compounds with specific biological activities. By generating molecular fingerprints that capture the structural features of the compounds, machine learning algorithms can predict their bioactivity profiles and prioritize the most promising candidates for further investigation.

Recent studies have demonstrated the utility of molecular fingerprint screening and machine learning approaches in identifying novel marine natural products with potent anticancer activities against gynecologic cancer cells. For instance, a study generated a library of over 10,000 marine natural products and screened them using a machine learning model trained on a dataset of known anticancer compounds. The researchers identified several compounds with high predicted bioactivity scores, and further experimental validation showed

that two of these compounds exhibited potent cytotoxic effects against ovarian cancer cells.

The use of molecular fingerprint screening and machine learning approaches in drug discovery is gaining momentum, and their application to marine natural products is a promising avenue for identifying new drug candidates for gynecologic cancers. By leveraging the structural diversity and bioactivity of marine natural compounds, researchers can develop effective and targeted treatments for these challenging diseases.

In inference, molecular fingerprint screening of marine natural compounds by using machine learning approaches is a powerful tool for identifying new drug candidates for gynecologic cancers. The use of this approach can help overcome the challenges associated with the traditional drug discovery process and accelerate the development of effective treatments for these diseases. By exploiting the vast diversity of marine natural products and leveraging the power of computational approaches, researchers can make significant strides in the fight against gynecologic cancers.

6.2 CONCLUSION

The identification of putative PI3K protein ligands and their ADMET features have been made possible by the molecular fingerprint screening of marine natural compounds utilising machine learning techniques and computational tools like SwissADME and Discovery Studio.

Several compounds with strong binding affinity and advantageous ADMET characteristics were found to be potential PI3K inhibitors using the virtual screening procedure. The ligand-protein complex's long-term stability was validated by subsequent molecular dynamics simulations.

Additionally, it has been demonstrated that using machine learning techniques like Random Forest and Support Vector Machines to predict the activity of chemicals and their toxicity profiles is useful. These methods provide a useful resource for medication development and discovery in the area of gynaecological cancer.

To sum up, it has been demonstrated that a powerful technique for the discovery and optimisation of new drug candidates is the combination of molecular fingerprint screening, virtual screening, molecular docking, molecular dynamics simulations, and machine learning approaches. These chemicals can be further optimised and experimentally validated as possible treatments for gynaecological cancer using the study's findings as a foundation.

REFERENCE

1. Alencar DB, Lopes NP. Marine natural products: History, status and perspectives. Química. 2019;42(2): doi:10.21577/0100-4042.20170281.
2. Ashok D, Kaul T, Kumaar R, Nair AGC Nova, Kumar MS. Marine natural products as potential antidiabetic agents: A review. Marine Drugs. 2018;16(11): doi:10.3390/md16110419.
3. Bhatnagar I, Kim SK. Immense essence of excellence: Marine microbial bioactive compounds. Marine Drugs. 2010;8(10): doi:10.3390/md8102673.
4. Blunt JW, Carroll AR, Copp BR, Davis RA, Keyzers RA, Prinsep MR. Marine natural products. 1st ed. Boca Raton, FL: CRC Press; 2019. p. 1-730.
5. Boswell, Z., Verga, J. U., Mackle, J., Guerrero-Vazquez, K., Thomas, O. P., Cray, J., ... & Hardiman, G. (2023). In-Silico Approaches for the Screening and Discovery of Broad-Spectrum Marine Natural Product Antiviral Agents Against Coronaviruses. Infection and Drug Resistance, 2321-2338.
6. Choudhary MI, Azhar-ul-Haq, Cho WK, Lee JH. Marine natural products as leads to anticancer drugs. Marine Drugs. 2017;15(8): doi:10.3390/md15080255.
7. D'Costa VM, Griffiths E. Marine natural products and antibiotic resistance. Nature Reviews Microbiology. 2020;18(3): doi:10.1038/s41579-019-0320-1.
8. Fusetani N, Sugawara T. Unique marine metabolites of diverse chemical origins. Natural Product Reports. 2017;34(3): doi:10.1039/c6np00110j.
9. Gademann K. Natural product synthesis and chemical biology. 1st ed. Chichester, UK: Wiley-Blackwell; 2017. p. 1-488.
10. Gago, F. (2023). Computational Approaches to Enzyme Inhibition by Marine Natural Products in the Search for New Drugs. Marine Drugs, 21(2), 100.
11. Gomila AS, Palermo JA, Grunberg RL, et al. Machine learning applied to the prediction of antitumor activity of natural compounds. International Journal of Quantum Chemistry. 2020;120(22): doi:10.1002/qua.26390.
12. Huang, M., Hunter, T., Slomovitz, B., & Schlumbrecht, M. (2019). Impact of molecular testing in clinical practice in gynecologic cancers. Cancer medicine, 8(5), 2013–2019. https://doi.org/10.1002/cam4.2064

13. Huang H, Zhang Y, Wu M, et al. Potential anti-inflammatory and antioxidant properties of marine-derived compounds in macrophages. Marine Drugs. 2020;18(2): doi:10.3390/md18020099.
14. Jukič, M., & Bren, U. (2022). Machine learning in antibacterial drug design. Frontiers in Pharmacology, 13, 1284.
15. Khalifa, S. A. M., Elias, N., Farag, M. A., Chen, L., Saeed, A., Hegazy, M. F., Moustafa, M. S., Abd El-Wahed, A., Al-Mousawi, S. M., Musharraf, S. G., Chang, F. R., Iwasaki, A., Suenaga, K., Alajlani, M., Göransson, U., & El-Seedi, H. R. (2019). Marine Natural Products: A Source of Novel Anticancer Drugs. *Marine drugs*, *17*(9), 491. https://doi.org/10.3390/md17090491
16. Liu X, Chen J, Sun Q, et al. Discovery of a natural product-derived covalent inhibitor of KRASG12C for the treatment of KRAS-driven cancers. Angewandte Chemie International Edition. 2021;60(11): doi:10.1002/anie.202016262.
17. Liu, L., Yang, H., Cai, Y., Cao, Q., Sun, L., Wang, Z., ... & Tang, Y. (2019). In silico prediction of chemical aquatic toxicity for marine crustaceans via machine learning. Toxicology Research, 8(3), 341-352.
18. Luo, L., Zheng, T., Wang, Q., Liao, Y., Zheng, X., Zhong, A., ... & Luo, H. (2022). Virtual Screening Based on Machine Learning Explores Mangrove Natural Products as KRASG12C Inhibitors. Pharmaceuticals, 15(5), 584.
19. McPhail KL. Marine natural products chemistry. 1st ed. New York, NY: Springer; 2018. p. 1-482.
20. Molinski TF. Marine natural products. Natural Product Reports. 2010;27(11): doi:10.1039/c0np00023e.
21. Newman DJ, Cragg GM. Natural products as sources of new drugs over the last 25 years. Journal of Natural Products. 2020;83(3): doi:10.1021/acs.jnatprod.9b01285.
22. Pangestuti R, Vo TS, Ngo DH, Kim SK. Marine algae-derived bioactive compounds for biomedical applications: An overview of pharmacological mechanisms and therapeutic potential. Natural Product Communications. 2021;16(5): doi:10.1177/1934578X211017475.
23. Pant, A., Jena, G. K., Ezzat, S. M., Chikhale, R., & Gurav, S. (2023). Roles of artificial intelligence and machine learning approach in natural products-based drug discovery. In Phytochemistry, Computational Tools and Databases in Drug Discovery (pp. 357-370). Elsevier.

24. Pereira, F. (2021). Machine learning methods to predict the terrestrial and marine origin of natural products. Molecular Informatics, 40(6), 2060034.
25. Qin Y, Zhu H, Jiang Y, et al. Identification of natural product-based antiviral agents by computational and experimental screening against the SARS-CoV-2 spike protein. Journal of Medicinal Chemistry. 2021;64(16): doi:10.1021/acs.jmedchem.1c00392.
26. Rasmussen BB, Mayer AMS. The encyclopedia of marine natural products. 1st ed. Berlin, Germany: Springer-Verlag; 2014. p. 1-1630
27. Ruiz-Torres V, Encinar JA, Herranz-López M, Pérez-Sánchez A, Galiano V, Barrajón-Catalán E, Micol V. An Updated Review on Marine Anticancer Compounds: The Use of Virtual Screening for the Discovery of Small-Molecule Cancer Drugs. Molecules. 2017; 22(7):1037. https://doi.org/10.3390/molecules22071037
28. Wang J, Gao Q, Sun M, et al. Machine learning-assisted discovery of novel marine natural products. Natural Product Reports. 2020;37(9): doi:10.1039/d0np00063f
29. Wang Y, Zhang Y, Cui Y, et al. Marine natural products as potential therapeutic agents for the treatment of Alzheimer's disease. Marine Drugs. 2020;18(3): doi:10.3390/md18030128.
30. Walker, A. S., & Clardy, J. (2021). A machine learning bioinformatics method to predict biological activity from biosynthetic gene clusters. Journal of Chemical Information and Modeling, 61(6), 2560-2571.