

Bioinformatics in Drug Discovery: Leveraging Machine Learning and Data Analysis

by

Krutika Gajanan Rajpure
(2282409)

Under the guidance of
Dr. Dalvin Vinoth Kumar A



A Project report submitted in partial fulfillment of the requirements for the award of the degree of Master of Science (Data Analytics) of CHRIST (Deemed to be University)

May – 2024


CERTIFICATE

*This is to certify that the report titled **Bioinformatics in Drug Discovery: Leveraging Machine Learning and Data Analysis** is a bona fide record of work done by **Krutika Gajanan Rajpure (2282409)** of CHRIST (Deemed to be University), Bangalore, in partial fulfillment of the requirements of VI Trimester MSc (Data Analytics) during the academic year 2023-24.*


Head of the Department


Project Guide

Valued-by

1.  **VISAYALARATHNA**

2.

Name : Krutika Gajanan Rajpure

Register Number : 2282409

Date of Exam : 04/05/2024

ACKNOWLEDGEMENT

I would like to express my sincere gratitude to Dr. Fr Joseph C C, Vice Chancellor in charge of CHRIST Deemed to be University, and Dr. Saleema J S, Head of the Department of Statistics and Data Science of CHRIST Deemed to be University for giving me this opportunity to work on this project under M.Sc. Data Analytics program.

Special thanks are due to Dr. Dalvin Vinoth Kumar A, my internal guide, whose support and invaluable guidance have been instrumental in steering me through the various stages of this project. His expertise and insightful feedback have been invaluable in the outcome of this work.

I am also deeply grateful to Dr. Sivakumar R, whose mentorship and constructive criticism have played a vital role in refining the scope and methodology of this research. Additionally, I extend my heartfelt thanks to Dr. Azarudheen S and Dr. Monisha Singh for their valuable input and suggestions during the evaluation process.

Overall, I express my heartfelt appreciation to my family and friends for their solidarity and assistance throughout this journey. Their encouragement and collaborative spirit have helped me in realizing the successful completion of my project and pursuing my goals.

ABSTRACT

The COVID-19 pandemic, caused by SARS-CoV-2, remains a global health crisis. This study uses bioinformatics, machine learning, and data analysis for SARS coronavirus 3C-like proteinase (3CLpro), a crucial enzyme in viral replication.

Various computational tools, including deep learning neural networks, are employed to predict drug-target interactions (DTIs). Comparative analysis reveals the efficacy of different methods, with ADME-Tox demonstrating promising results in conjunction with Random Forest and Gaussian Naïve Bayes classifiers. Additionally, Multilayer Perceptron neural networks exhibit high predictive accuracy.

Furthermore, recent advancements in drug design involving the development of novel inhibitors against SARS-CoV-2 main protease are highlighted. The design, synthesis, and evaluation of bicycloproline-containing Mpro inhibitors derived from approved antivirals demonstrate potent inhibitory activity in vitro.

Retro-inverse peptide design yields potent inhibitors with low micromolar IC₅₀ values, presenting the potential for further development as therapeutic agents against COVID-19. These findings underscore the importance of bioinformatics and computational techniques in accelerating drug discovery efforts against emerging viral pathogens like SARS-CoV-2.

Despite the great abundance of tools available for drug discovery, we are left in a situation of deciding what tools are potentially available for performing a certain task and which one to use. Thus, it is the ambition of this editorial to take a glimpse at what tools are available and how we can maximize our productivity by presenting the available toolbox for drug discovery.

TABLE OF CONTENTS

Acknowledgments	i
Abstract	ii
List of Tables	iii
List of Figures	iv
1. Introduction	1
1.1. Understanding the Project Domain	1
1.2. Problem Statement & Primary Objectives	3
1.3. Training Methodology	4
1.4. Tools and Concept	6
2. Literature Review	7
3. Dataset Description and Pre-processing	
3.1. About The Database	10
3.2. Data Collection	10
3.2.1. Accessing the Database	10
3.2.2. Querying ChEMBL's Database	11
3.3. Data Filtering and Preprocessing	13
4. Exploratory Data Analysis	
4.1. Frequency plot of two bioactivity classes	17
4.2. Scatter plot of MW vs LogP	18
4.3. Box plot for pIC ₅₀ value	18
4.4. MW & LogP Vs Bioactivity class	20
4.5. NumhDonors & NumhAcceptors Vs Bioactivity class	21

5. Modelling	
5.1. Calculates Fingerprint descriptors	23
5.2. Splitting the data X and Y	25
5.3. Model Building	26
5.3.1 Random Forest Regression	26
5.3.2 Linear Regression	27
5.4. Comparison of Linear Regression and Random Forest	28
5.5. Data Visualization (Experimental vs Predicted pIC50)	29
 6. Model Deployment	 30
 7. Conclusion	 35
 7. References	 37

LIST OF TABLES

Table No.	Table Name	Page No.
3.3.1	Column name and description of the database	16
4.5.1	Statistical Analysis of 4 Lipinski's Descriptors	21

LIST OF FIGURES

Figure No.	Figure Name	Page No.
3.2.1.1	Chembl's Web Interface	11
3.2.2.1	Targets searched CORONAVIRUS show 10 entries on the Web Interface	12
3.2.2.2	Targets searched CORONAVIRUS show 10 entries in Collab	12
3.3.1	Filtered data and new bioactivity_class created	13
3.3.2	Data after calculating Lipinski descriptors	14
3.3.3	Data after converting IC50 to pIC50	15
4.1.1	Frequency plot of the 2 bioactivity classes	17
4.2.1	Scatter plot of MW versus LogP	18
4.3.1	Boxplot for pIC50 value	18
4.3.2	Statistical analysis Mann-Whitney U Test	19
4.4.1	Boxplot for MW	20
4.4.2	Boxplot for LogP	20
4.5.1	Boxplot for NumHDonor	21
4.5.2	Boxplot for NumHAcceptors	21
5.1.1	Calculating PubChem descriptor	23
5.1.2	Data Info	24
5.1.3	Dataset	24
5.2.1	X & Y matrix	25
5.3.1.1	Accuracy of Random Forest Model	26
5.3.2.1	Accuracy of Linear Regression Model	27
5.4.1	Comparison of actual value vs predicted values by linear regression and random forest	28
5.5.1	Experimental vs Predicted pIC50 for Training Data for Linear Regression	29
6.1	User Interface App built with Streamlit	31
6.2	txt file used as the input data to upload in the app	31
6.3	The above file is named coronavirus.txt and is uploaded in the app the user has to click on the Predict button for the results to be predicted	32
6.4	After the user clicks on the Predict button the result appears on the app.	32
6.5	Results Part 1	33
6.6	Results Part 2	33
6.7	Prediction Output	34

1. INTRODUCTION

Drug discovery is a complex and resource-intensive process involving the screening of vast chemical libraries to identify potential hits, which are then optimized into lead compounds and eventually developed into drugs. This process typically takes over a decade and incurs substantial costs exceeding \$1 billion. Computational tools have become indispensable in various stages of drug discovery, significantly enhancing efficiency and success rates.

In drug discovery, predicting drug-target interactions (DTIs) is crucial for identifying potential drugs. Computational methods such as molecular docking and machine learning enable the prediction of DTIs, with chemical compounds represented using Simplified Molecular Input Line Entry System (SMILES) strings or molecular fingerprints like Extended-Connectivity Fingerprints (ECFP).

Machine learning, particularly supervised learning algorithms, has emerged as a powerful tool for DTI prediction due to its scalability and efficiency. However, challenges such as class imbalance in datasets need to be addressed to improve the accuracy of predictions. While comparative studies on resampling methods for clinical datasets have been conducted, similar analyses for chemical datasets, especially in DTI prediction, remain scarce.

1.1 UNDERSTANDING THE PROJECT DOMAIN

Bioinformatics in drug discovery is a field that applies computational and analytical methods to biological data to discover and develop new pharmaceuticals.

It combines principles of biology, chemistry, computer science, and statistics to analyze large datasets, such as genomic sequences, protein structures, and chemical compounds, to identify potential drug candidates.

- Key Components of bioinformatics drug discovery include:
 - Data Integration and Analysis: Integration of diverse biological and chemical datasets, such as genomic data, protein structures, and chemical compound libraries.
 - Target Identification and Validation: Identification of molecular targets, such as proteins or genes, that play a crucial role in disease pathways.
 - Computational methods are used to predict the drug ability and biological function of potential targets, followed by experimental validation.
 - Virtual Screening: Computational screening of chemical compound libraries to identify molecules with potential therapeutic activity against a specific target.
 - ADME-Tox Prediction: Prediction of the absorption, distribution, metabolism, excretion, and toxicity (ADME-Tox) properties of drug candidates using computational models. This helps prioritize compounds with favorable pharmacokinetic and safety profiles for further development.

1.2 PROBLEM STATEMENT & PRIMARY OBJECTIVES

Problem Statement: The emergence of the SARS coronavirus and its variants poses a critical global health threat, demanding the creation of efficient therapeutics.

Conventional drug discovery methods are slow and expensive, relying heavily on experimental testing. However, employing bioinformatics, machine learning, and computational methods can accelerate this process by pinpointing promising drug candidates with favorable pharmacological traits.

Primary Objectives:

This project aims to use bioinformatics, machine learning, and computational techniques to find potential drugs targeting the SARS coronavirus 3C-like proteinase. It involves:

- Retrieving and preprocessing bioactivity data from databases like ChEMBL.
- Calculating molecular descriptors to assess drug-likeness and predict pharmacokinetic properties.
- Developing machine learning models to predict compound bioactivity against the SARS coronavirus 3C-like proteinase.
- Identifying lead compounds with promising pharmacological profiles for further experimental validation and drug development. Ultimately, the project seeks to contribute to discovering effective treatments for coronavirus infections.

1.3 TRAINING METHODOLOGY

Data Collection and Preprocessing: For the project focused on identifying potential drug candidates targeting the SARS coronavirus 3C-like proteinase, the data collection process involves retrieving bioactivity data from relevant databases, such as ChEMBL, containing compounds tested against the target protein.

Feature Engineering: Extract relevant features from the chemical compounds and target proteins to represent them in a format suitable for machine learning algorithms.

Explore various molecular descriptors, fingerprints, or embeddings to capture important characteristics of the compounds and proteins. Experiment with different feature selection techniques to identify the most informative features for DTI prediction.

Exploration Data Analysis : Acquire relevant datasets containing information on chemical compounds, target proteins, and their interactions from reputable sources or databases.

Understand the structure and format of the datasets, including the features, labels, and any associated metadata. Perform preliminary data exploration to gain insights into the distribution of classes, potential outliers, and missing values.

Model Selection and Training: Explore a variety of machine learning models suitable for binary classification tasks, such as logistic regression, random forests, support vector machines (SVM), and deep learning architectures. Train multiple models using the training dataset and evaluate their performance on the validation set. Optimize hyperparameters using techniques like grid search or Bayesian optimization to improve model performance.

Model Evaluation and Validation: Assess the performance of trained models using appropriate evaluation metrics such as accuracy, precision, recall, F1-score, and area under the receiver operating characteristic curve (AUC-ROC). Validate the models on the test set to ensure their generalization to unseen data.

Interpretation and Analysis: Interpret the trained models to gain insights into the underlying relationships between chemical compounds and target proteins. Analyze feature importance to identify the most influential factors contributing to DTI prediction.

Documentation and Reporting: Document the entire training process, including data preprocessing steps, model architectures, hyperparameters, and evaluation results.

By following this structured training methodology, the project aims to develop robust and accurate predictive models for DTI prediction, contributing to the advancement of computational approaches in drug discovery.

1.4 TOOLS AND CONCEPT

1. ChEMBL WebResource Client: This tool is essential for accessing bioactivity data from the ChEMBL database, which contains comprehensive information on bioactive molecules, their targets, and their interactions.

2. Conda and RDKit: Conda is a package manager and environment manager that simplifies the installation and management of software packages, including RDKit.

RDKit enables tasks such as molecular descriptor calculation, substructure searching, molecule manipulation, and chemical property prediction.

3. Jupyter Notebook (Collab): Jupyter Notebook, particularly in the Collab environment, serves as an interactive computing environment for running Python code, visualizing data, and documenting the analysis process.

4. Python Libraries: Various Python libraries, such as scikit-learn, pandas, NumPy, matplotlib, and seaborn, are utilized for data manipulation, preprocessing, visualization, and machine learning tasks. Streamlit to build interactive web applications.

2. LITERATURE REVIEW

Chanin Nantasenamat, and Virapong Prachayasittikul's (2015) study highlight the vital role of computational tools in successful drug discovery. They emphasize virtual screening's importance in identifying potential compounds, utilizing resources like ZINC and Chem Spider. The integration of ligand and structure-based approaches, exemplified by the SABRE program, enhances virtual screening's efficacy. They discuss challenges in virtual screening, including erroneous assumptions and conformational sampling issues.

The paper underscores the importance of optimizing drug candidates not only as strong binders but also considering pharmacokinetic profiles and adverse effects. They advocate for utilizing multiple computational approaches and anticipate a future trend towards multi-target strategies and systems-based approaches in drug discovery. Additionally, they propose developing infrastructures supporting interoperability among databases and tools, leveraging semantic web technology, and promoting open notebook and workflow tools for advancing drug discovery research.

Khan, A. K. A., & Malim, N. H. A. H. H. (2023) present a paper on exploring class imbalance in drug-target interaction prediction models using various resampling techniques on a clinical dataset. They highlight the importance of accurate prediction models in drug discovery, particularly in cancer treatment, and compare machine learning and deep learning methods. Their findings indicate that the deep learning method Multilayer Perceptron (MLP) without resampling outperforms machine learning classifiers with various resampling methods in terms of F1 score. This research contributes to enhancing predictive model reliability in drug discovery, facilitating the identification of potential drug candidates and improving therapeutic interventions for various diseases.

Christopher A. Lipinski, Franco Lombardo, Beryl W. Dominy, and Paul J. Feeney (2001) delve into experimental and computational methods for estimating solubility and permeability in drug discovery and development. Key points include discussions on the Rule of 5, which predicts poor absorption or permeation when specific criteria are met, such as more than 5 hydrogen bond donors, more than 10 hydrogen bond acceptors, molecular weight (MWT) greater than 500, and calculated Log P (CLogP) greater than 5 (or MlogP > 4.15). The authors also cover turbidimetric solubility measurement, challenges in the development setting, orally active drugs outside the Rule of 5, trends in high molecular weight (MWT) compounds, and protocols for measuring drug solubility in discovery settings.

Jia, Li, Hao, and Yang's (2019) study sheds light importance of drug-likeness assessment in drug discovery, emphasizing its role in selecting compounds with favorable bioavailability and reducing the risk of candidate failures. Their paper offers a brief overview of online resources for in silico drug-likeness studies, emphasizing their cost-effectiveness and efficiency. By compiling a toolbox with key features from various resources, the authors aim to guide future research in this area. This review serves as a valuable resource for researchers utilizing online tools in drug discovery.

Aghdam, R., Habibi, M., and Taheri, G. (2021) study propose a machine learning-based method for COVID-19 drug repurposing. The research focuses on constructing a COVID-19 related biological network and selecting essential proteins associated with COVID-19 pathology. Then, informative features based on drug–target and protein–protein interaction information are proposed. Through these features, the study identifies five clusters of drugs with potential as COVID-19 treatments. Notably, 80% of the proposed candidate drugs have been studied in other research and clinical trials.

Lv, Shi, Berkenpas, Dao, Zulfikar, Ding, Zhang, Yang, and Cao (2021) propose an application of artificial intelligence and machine learning for COVID-19 drug discovery and vaccine design. Their research focuses on leveraging AI techniques to accelerate the identification of potential drugs and vaccine candidates for COVID-19. By analyzing biological data and protein interactions, the study aims to provide insights into repurposing existing drugs and designing novel vaccines. The approach holds promise for addressing urgent global health challenges by expediting drug development and vaccine design.

Monteleone, Kellici, Southey, Bodkin, and Heifetz (2021) focus on the impact and challenges of artificial intelligence (AI) in drug repurposing for COVID-19 treatment. Despite the development of vaccines, the risk of vaccine-resistant variants remains. AI plays a crucial role in identifying potential drug candidates and monitoring the pandemic. Drug repurposing offers advantages such as faster clinical trials and reduced costs compared to developing new molecules.

Sreepadmanabh, M., Sahu, A. K., & Chande, A. (2020) provide a comprehensive review on the advancements in diagnostic tools, treatment strategies, and vaccine development for COVID-19. The global research community faces urgent challenges in combating the pandemic, necessitating rapid diagnostic tools, effective treatment protocols, and vaccine candidates. The authors critically evaluate various approaches, including conventional methods like serology and RT-PCR, as well as cutting-edge technologies such as CRISPR/Cas and artificial intelligence/machine learning.

3. DATASET DESCRIPTION AND PRE-PROCESSING

3.1 ABOUT THE DATABASE

ChEMBL is a 'chemogenomic' database that brings together chemical, bioactivity, and genomic data to aid the translation of genomic information into effective new drugs.

ChEMBL originated from the work of the European Bioinformatics Institute (EBI) and was initially developed by the European Molecular Biology Laboratory (EMBL) in collaboration with pharmaceutical companies. The project began in the early 2000s to create a comprehensive database of bioactive molecules and their targets. Over the years, ChEMBL has grown in scope and quality, becoming one of the most widely used resources in the fields of cheminformatics and drug discovery.

The ChEMBL database contains information on over 2.2 million compounds and 18 million records of their effects on biological systems.

It is compiled from more than 76,000 documents, 1.2 million assays, and the data spans 13,000 targets and 1,800 cells, and 33,000 indications. ChEMBL is used by a wide range of stakeholders in the pharmaceutical industry, academia, and research institutions.

3.2 DATA COLLECTION

3.2.1 ACCESSING THE DATABASE

Accessing the ChEMBL database is straightforward and can be done through its web interface or programmatically via its API (Application Programming Interface).

Web Interface: ChEMBL's web interface provides a user-friendly way to search for compounds, targets, assays, and other data within the database. Users can perform keyword searches, browse through data using various filters and categories, and view detailed information about specific compounds or targets.

API Access: For more advanced users or those who wish to integrate ChEMBL data into their own applications or workflows, the database offers an API. The API allows users to query and retrieve data from ChEMBL using HTTP requests programmatically.

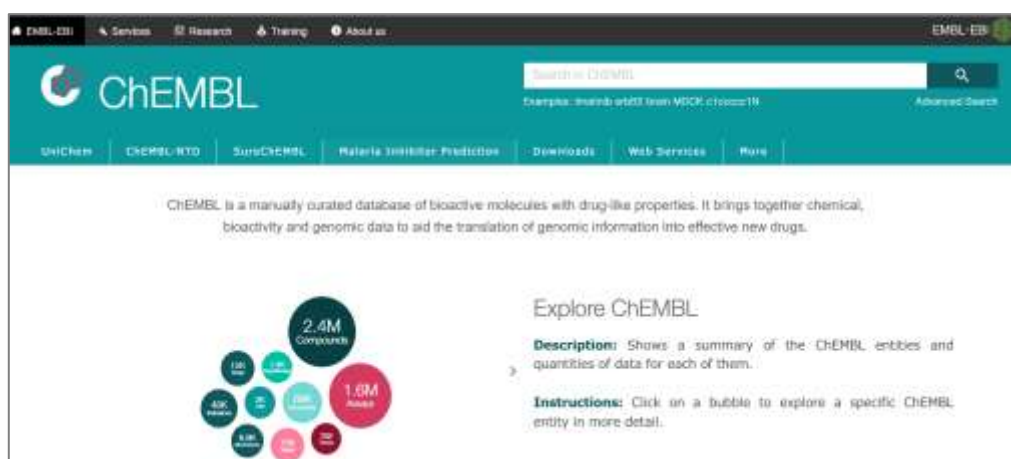


Fig 3.2.1.1 ChEMBL's Web Interface

3.2.2 QUERYING ChEMBL's DATABASE

The research aims to gather bioactivity data for the SARS coronavirus 3C-like proteinase enzyme from the ChEMBL database. This enzyme is a potential target for drug development against SARS-CoV, the virus responsible for severe acute respiratory syndrome. Initially, a query is made to the ChEMBL database to retrieve information on targets related to coronaviruses. There are 10 entries (each representing a coronavirus strain or protein) and 9 attributes describing each entry.

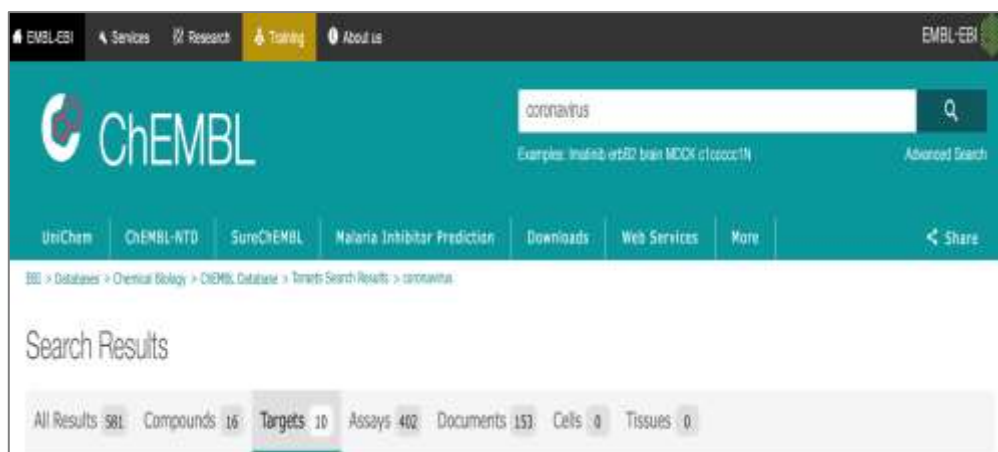


Fig 3.2.2.1 Targets searched CORONAVIRUS shows 10 entries on Web Interface

	pref_name	score	species_group_flag	target_chembl_id	target_type	tax_id
0	Coronavirus	17	False	CHEMBL613732	ORGANISM	11119
1	Feline coronavirus	14	False	CHEMBL612744	ORGANISM	12663
2	Murine coronavirus	14	False	CHEMBL5209664	ORGANISM	694805
3	Canine coronavirus	14	False	CHEMBL5291668	ORGANISM	11153
4	Human coronavirus 229E	13	False	CHEMBL613837	ORGANISM	11137
5	Human coronavirus OC43	13	False	CHEMBL5209665	ORGANISM	31631
6	SARS coronavirus 3C-like proteinase	10	False	CHEMBL3927	SINGLE PROTEIN	227859
7	Middle East respiratory syndrome-related coronavirus	9	False	CHEMBL4296578	ORGANISM	1335626
8	Replicase polyprotein 1ab	4	False	CHEMBL5118	SINGLE PROTEIN	227859
9	Replicase polyprotein 1ab	4	False	CHEMBL4523582	SINGLE PROTEIN	2697049

Fig 3.2.2.2 Targets searched CORONAVIRUS show 10 entries in Collab

The target of interest, the SARS coronavirus 3C-like proteinase (ChEMBL ID: ChEMBL3927), is selected, and its corresponding ChEMBL ID is assigned to a variable. Subsequently, bioactivity data specific to this target protein are retrieved, focusing on inhibitory concentration (IC₅₀) values reported for various compounds.

The retrieved data are filtered to include only IC₅₀ values, which represent the concentration of a compound required to inhibit the activity of the proteinase. The resulting dataset, with a shape of (133, 46), is stored in a pandas data frame for further analysis.

3.3 DATA FILTERING AND PREPROCESSING

The data filtering and preprocessing process involves selecting relevant columns from the retrieved dataset and applying criteria to categorize compounds based on their bioactivity. From the initial dataset containing 46 columns, the columns of interest are narrowed down to include only the *molecular ChEMBL ID* (mol_cid), *canonical SMILES* (canonical_smiles), and *standard value* (standard_value) representing the potency of the drug.

Compounds with IC50 values less than 1000 nM are classified as active, those greater than 10,000 nM as inactive, and values falling between 1,000 and 10,000 nM as intermediate. This categorization provides a clear distinction between compounds with high, low, or moderate potency against the SARS coronavirus 3C-like proteinase.

	molecule_chembl_id	canonical_smiles	standard_value	bioactivity_class
0	CHEMBL187579	<chem>Cc1noc(C)c1CN(C(=O)C(=O)c2cc(C#N)ccc21</chem>	7200	intermediate
1	CHEMBL188487	<chem>O=C1C(=O)N(Cc2ccc(F)cc2Cl)c2ccc(I)cc21</chem>	9400	intermediate
2	CHEMBL185698	<chem>O=C1C(=O)N(CC2CCc3ccccc3O2)c2ccc(I)cc21</chem>	13500	inactive
3	CHEMBL426082	<chem>O=C1C(=O)N(Cc2cc3ccccc3s2)c2ccccc21</chem>	13110	inactive
4	CHEMBL187717	<chem>O=C1C(=O)N(Cc2cc3ccccc3s2)c2c1ccccc2[N+](=O)[O-]</chem>	2000	intermediate
5	CHEMBL365134	<chem>O=C1C(=O)N(Cc2cc3ccccc3s2)c2c(Br)cccc21</chem>	900	active
6	CHEMBL187598	<chem>O=C1C(=O)N(Cc2cc3ccccc3s2)c2ccc(F)cc21</chem>	4820	intermediate
7	CHEMBL190743	<chem>O=C1C(=O)N(Cc2cc3ccccc3s2)c2ccc(I)cc21</chem>	950	active
8	CHEMBL365469	<chem>O=C1C(=O)N(Cc2cc3ccccc3s2)c2ccc(Cl)c21</chem>	11200	inactive
9	CHEMBL188983	<chem>O=C1C(=O)N(C/C=C/c2cc3ccccc3s2)c2ccc(I)cc21</chem>	23500	inactive

Fig 3.3.1 Filtered data and new bioactivity_class created

Calculate Lipinski descriptors

Christopher Lipinski, a scientist at Pfizer, came up with a set of rule-of-thumb for evaluating the drug-likeness of compounds. Such drug-likeness is based on the Absorption, Distribution, Metabolism, and Excretion (ADME) which is also known as the pharmacokinetic profile. Lipinski analysed all orally active FDA-approved drugs in the formulation of what is to be known as the **Rule-of-Five** or **Lipinski's Rule**.

Calculates Lipinski descriptors for compounds represented by SMILES strings and returns a DataFrame containing these descriptors along with additional information about the compounds.

The Lipinski's Rule stated the following:

- Molecular weight < 500 Dalton
- Octanol-water partition coefficient (LogP) < 5
- Hydrogen bond donors < 5
- Hydrogen bond acceptors < 10

molecule_chembl_id	canonical_smiles	standard_value	bioactivity_class	MW	LogP	NumH Donors	NumH Acceptors	standard_value_norm
CHEMBL187579	<chem>Cc1noc(C)c1CN1C(=O)C(=O)c2cc(C#N)ccc21</chem>	7200	intermediate	281.271	1.89262	0	5	7200
CHEMBL188487	<chem>O=C1C(=O)N(Cc2ccc(F)cc2Cl)c2ccc(I)cc21</chem>	9400	intermediate	415.589	3.8132	0	2	9400
CHEMBL185698	<chem>O=C1C(=O)N(CC2COc3ccccc3O2)c2ccc(I)cc21</chem>	13500	inactive	421.19	2.6605	0	4	13500
CHEMBL426082	<chem>O=C1C(=O)N(Cc2cc3c3ccccc3s2)c2ccccc21</chem>	13110	inactive	293.347	3.6308	0	3	13110
CHEMBL187717	<chem>O=C1C(=O)N(Cc2cc3c3ccccc3s2)c2c1cccc2[N+](=O)[O-]</chem>	2000	intermediate	338.344	3.539	0	5	2000

Fig 3.3.2 Data after calculating Lipinski descriptors

Convert IC50 to pIC50

To allow IC50 data to be more uniformly distributed, we will convert IC50 to the negative logarithmic scale which is essentially $-\log_{10}(\text{IC}_{50})$. This custom function `pIC50()` will accept a DataFrame as input and will:

- Take the IC50 values from the `standard_value` column and convert it from nM to M by multiplying the value by 10^{-9}
- Take the molar value and apply $-\log_{10}$
- Delete the `standard_value` column and create a new `pIC50` column

molecule_chembl_id	canonical_smiles	bioactivity_class	MW	LogP	NumHD onors	NumHA cceptor	pIC50
CHEMBL187579	<chem>Cc1noc(C)c1CN1C(=O)C(=O)c2cc(C#N)ccc21</chem>	intermediate	281.271	1.89262	0	5	5.142668
CHEMBL188487	<chem>O=C1C(=O)N(Cc2ccc(F)cc2Cl)c2ccc(I)cc21</chem>	intermediate	415.589	3.8132	0	2	5.026872
CHEMBL185698	<chem>O=C1C(=O)N(CC2COc3cccc3O2)c2ccc(I)cc21</chem>	inactive	421.19	2.6605	0	4	4.869666
CHEMBL426082	<chem>O=C1C(=O)N(Cc2cc3ccccc3s2)c2cccc21</chem>	inactive	293.347	3.6308	0	3	4.882397
CHEMBL187717	<chem>O=C1C(=O)N(Cc2cc3ccccc3s2)c2c1cccc2[N+](=O)[O-]</chem>	intermediate	338.344	3.539	0	5	5.69897

Fig 3.3.3 Data after converting IC50 to pIC50

From fig 3.3.3 it can be seen the first five entries of the final dataset. For further analysis and model building the above dataset will be in use.

The different columns that have been extracted and calculated are shown above for the analysis and their descriptions are as follows:

Table 3.3.1 Column name and description of the database

Column name	Description
molecule_chembl_id	Unique compound identifier in the ChEMBL database
canonical_smiles	Standardized molecular structure representation
bioactivity_class	Categorization of compounds as active or inactive based on observed biological effects
MW (Molecular Weight)	Sum of atomic weights, influencing pharmacokinetic properties.
LogP (Octanol-water partition coefficient)	Lipophilicity, affecting solubility and distribution.
NumHDonors (Number of Hydrogen Bond Donors)	Number of hydrogen bond donors, impacting molecular interactions.
NumHAcceptors (Number of Hydrogen Bond Acceptors)	Number of hydrogen bond acceptors, influencing binding to targets.
pIC50	Potency quantification, represented as the negative logarithm of IC50 values

4. EXPLORATORY DATA ANALYSIS

4.1 FREQUENCY PLOT OF TWO BIOACTIVITY CLASSES

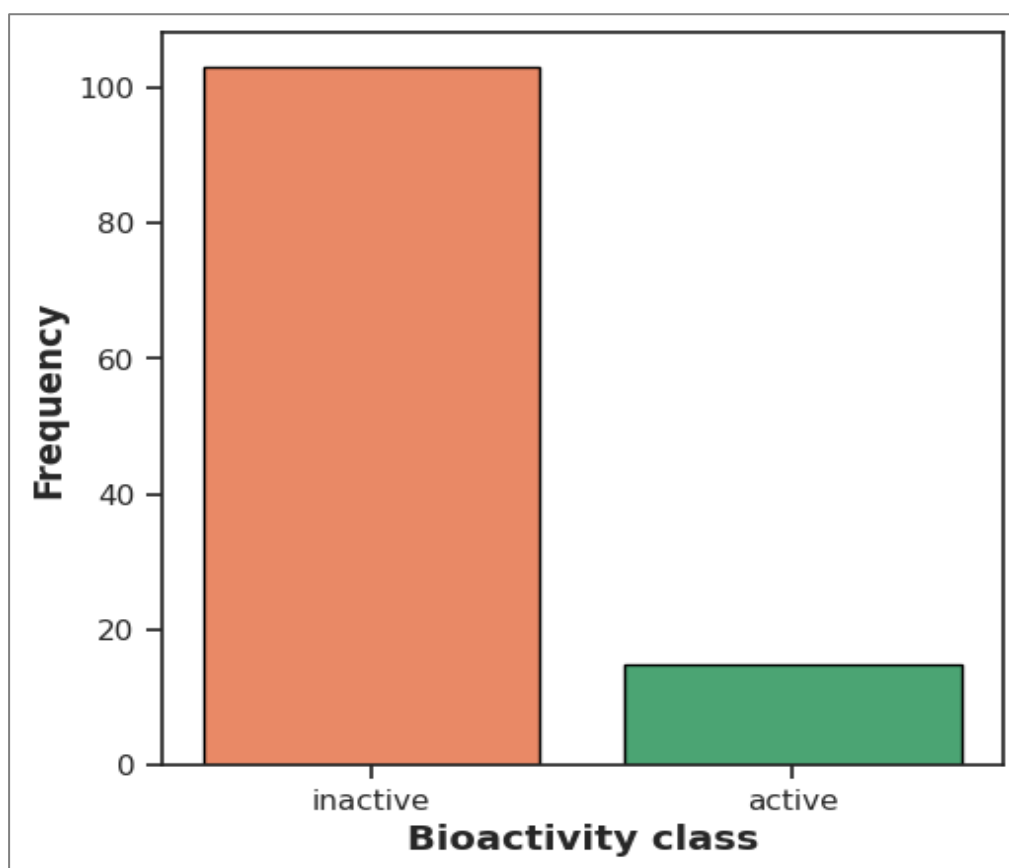


Fig 4.1.1 Frequency plot of the 2 bioactivity classes

From fig 4.1.1, we can see the frequency of the inactive bioactivity class is more than compared to the active class.

4.2 SCATTER PLOT OF MW VERSUS LOGP

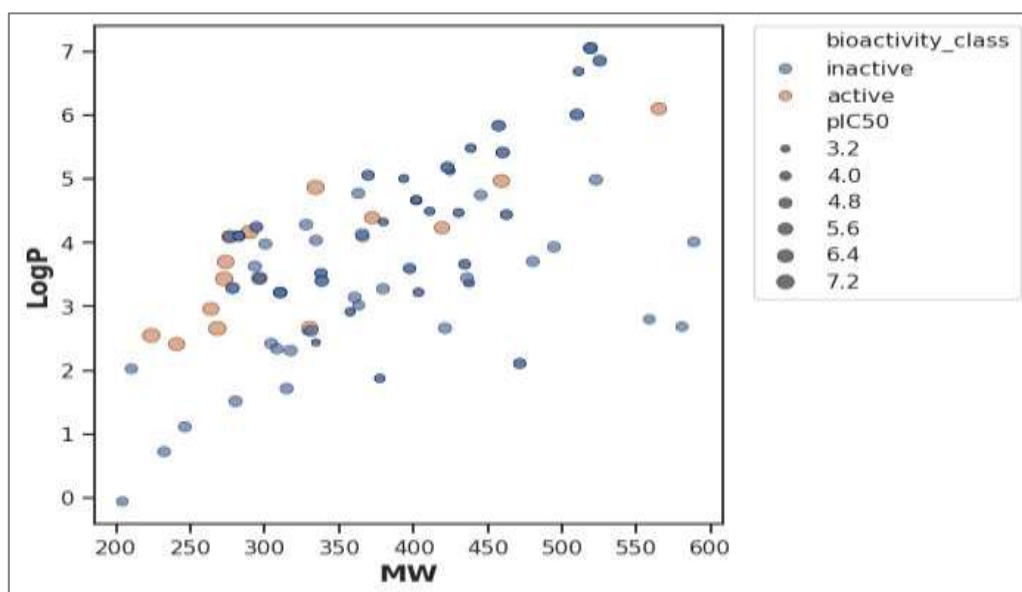


Fig 4.2.1 Scatter plot of MW versus LogP

From fig 4.2.1, we can see the trend that the 2 bioactivity classes are spanning similar chemical spaces as evidenced by the scatter plot of MW vs LogP.

4.3 BOXPLOT FOR PIC50 VALUE

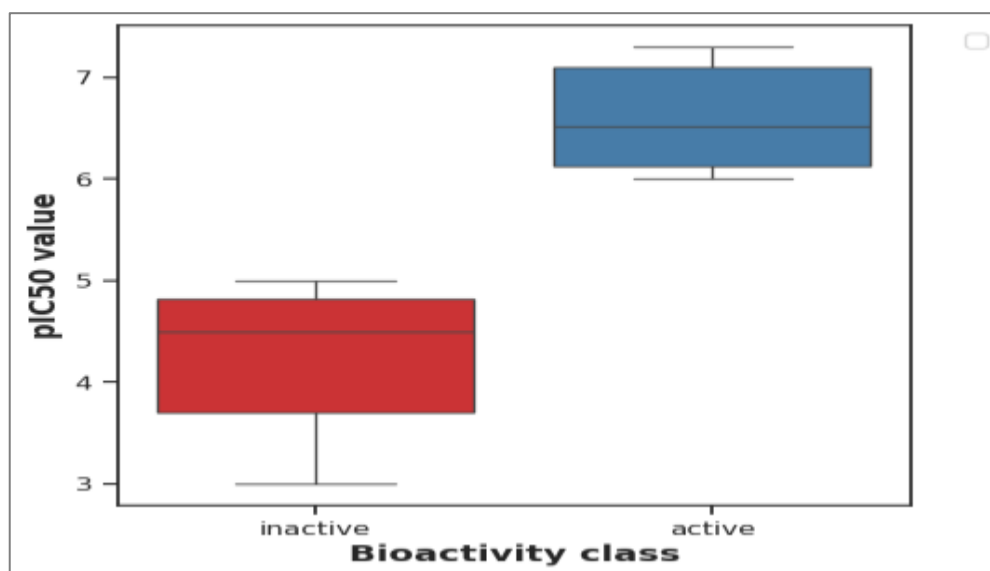


Fig 4.3.1 Boxplot for pIC50 value

From fig 4.3.1, it showcases the distribution and central tendency of pIC50 values for both active and inactive compounds. The median line of the active class lies outside of the inactive.

	Descriptor	Statistics	p	alpha	Interpretation
0	pIC50	1545.0	4.428384e-10	0.05	Different distribution (reject H0)

Fig 4.3.2 Statistical analysis | Mann-Whitney U Test

From fig 4.3.2, The Mann-Whitney U test confirms a significant disparity ($p < 0.05$) in pIC50 values between active and inactive compounds, indicating distinct distributions and potential predictive power of this descriptor.

Taking a look at pIC50 values, the actives and inactives displayed statistically significant differences, which is to be expected since threshold values ($IC_{50} < 1,000 \text{ nM} = \text{Actives}$ while $IC_{50} > 10,000 \text{ nM} = \text{Inactives}$, corresponding to $pIC_{50} > 6 = \text{Actives}$ and $pIC_{50} < 5 = \text{Inactives}$) were used to define actives and inactives.

4.4 MW AND LOGP Vs BIOACTIVITY CLASS (ACTIVE & INACTIVE)

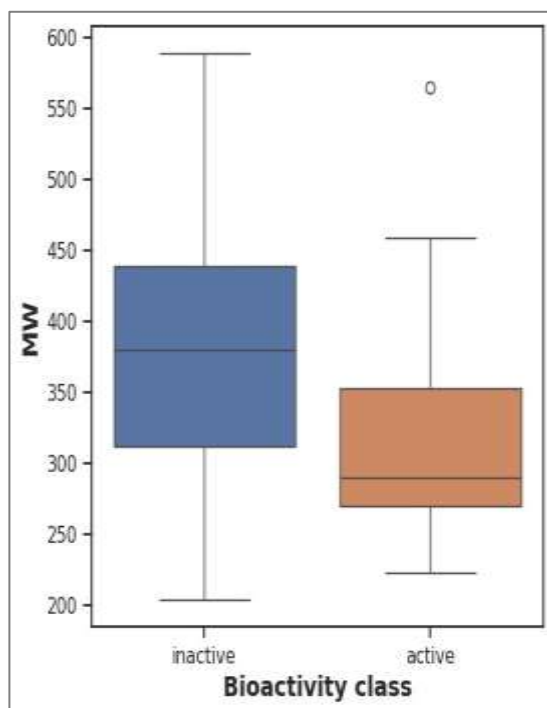


Fig 4.4.1 Boxplot for MW

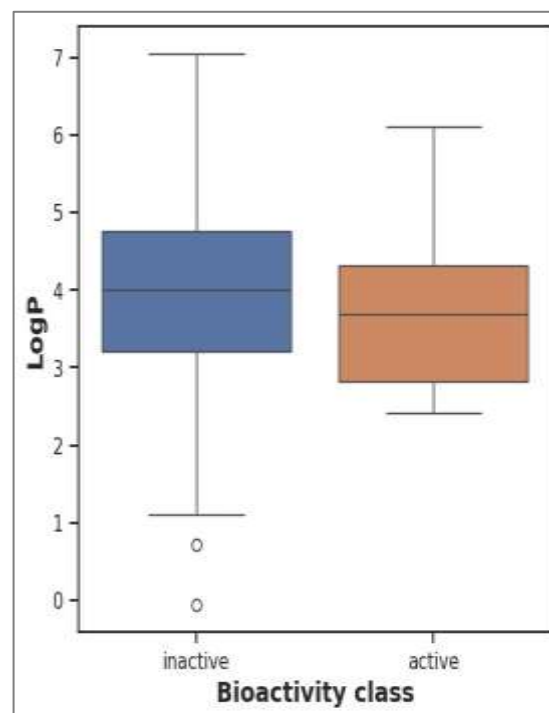


Fig 4.4.2 Boxplot for LogP

From fig 4.4.1, and fig 4.4.2, showcase the distribution and central tendency of MW and LogP for both active and inactive compounds. We can see that MW active class has outliers present and the median line lies outside whereas for LogP the median lines within both the classes.

4.5 NUMHDONORS AND NUMHACCEPTORS Vs BIOACTIVITY CLASS (ACTIVE & INACTIVE)

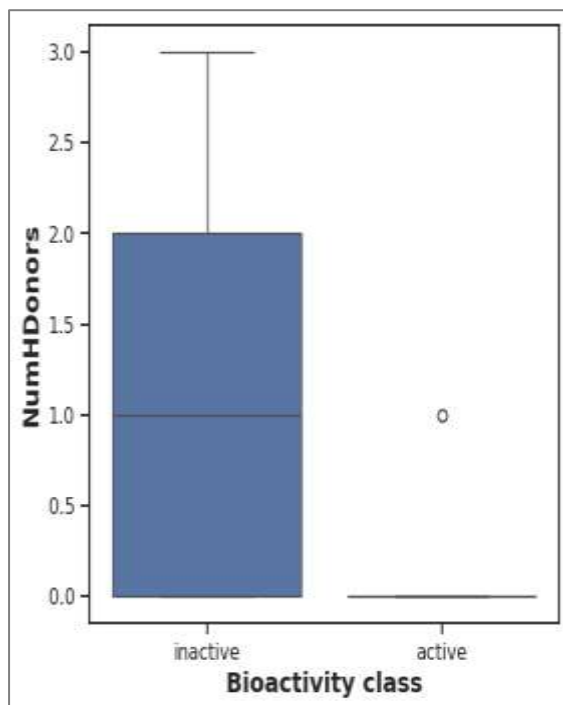


Fig 4.5.1 Boxplot for NumHDonor

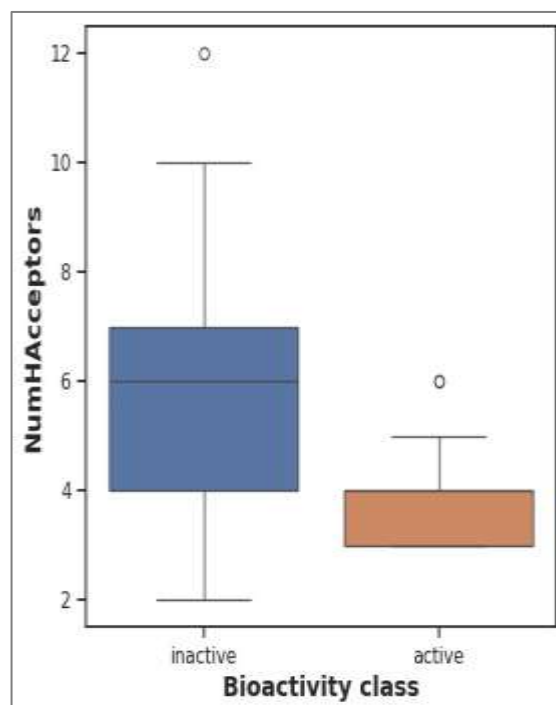


Fig 4.5.2 Boxplot for NumHAceptors

From fig 4.5.1, and fig 4.5.1, showcase the distribution and central tendency of NumHDonor and NumHAceptors for both active and inactive compounds. We can see that NumHDonor and NumHAceptors active classes have outliers present and the median line lies outside for both classes.

Table 4.5.1 Statistical Analysis of 4 Lipinski's Descriptors

Statistical analysis Mann-Whitney U Test				
Statistics	p	alpha	Interpretation	
MW	411	0.00349	0.05	Different distribution (reject H0)
LogP	701	0.563449	0.05	Same distribution (fail to reject H0)
NumHDonors	299	0.000053	0.05	Different distribution (reject H0)
NumHAcceptors	414	0.003402	0.05	Different distribution (reject H0)

All of the 4 Lipinski's descriptors exhibited statistically significant differences between the active and inactive.

This suggests that these descriptors are effective in differentiating the bioactivity classes, highlighting their potential utility in drug discovery and development processes.

5. MODELLING

5.1 CALCULATE FINGERPRINT DESCRIPTORS

We calculate fingerprint descriptors to represent molecular structures in a format suitable for computational analysis and comparison. Fingerprint descriptors are abstract representations of a molecule's structural features. They can represent a structural key within a molecule, such as a count of a particular atom type.

The PubChem fingerprint, generated by PubChem, encodes structural features of molecules as binary vectors. It's widely used in cheminformatics for tasks like similarity searching and virtual screening in drug discovery.

The PubChem fingerprint encodes molecular fragment information with 881 binary digits. The binary bits indicate whether a specific group of chemical features is present in a compound.

```
1  from padelpy import padeldescriptor
2
3  fingerprint = 'PubChem'
4
5  fingerprint_output_file = ''.join([fingerprint, '.csv']) #Substructure.csv
6  fingerprint_descriptortypes = fp[fingerprint]
7
8  padeldescriptor(mol_dir='molecule.smi',
9                  d_file=fingerprint_output_file, #'Substructure.csv'
10                  #descriptortypes='SubstructureFingerprint.xml',
11                  descriptortypes= fingerprint_descriptortypes,
12                  detectaromaticity=True,
13                  standardizenitro=True,
14                  standardizetautomers=True,
15                  threads=2,
16                  removesalt=True,
17                  log=True,
18                  fingerprints=True)

1  descriptors = pd.read_csv(fingerprint_output_file)
2  descriptors
```

Fig 5.1.1 Calculating PubChem descriptor

From fig 5.1.1, the Padelpy library generates fingerprint descriptors for molecules. It specifies the use of the PubChem fingerprint to calculate fingerprint descriptors.

```
1 dataset.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 133 entries, 0 to 132
Columns: 882 entries, PubchemFP0 to pIC50
dtypes: float64(1), int64(881)
memory usage: 916.6 KB
```

Fig 5.1.2 Data Info

From fig 5.1.2, we can clearly see the data info. It contains 133 entries which represent our number of molecules and 882 columns of which 881 are PubChem fingerprint descriptors of integer type and pIC50 value of each molecule.

	PubchemFP0	PubchemFP1	PubchemFP2	PubchemFP3	PubchemFP4	PubchemFP5	PubchemFP6	PubchemFP7	PubchemFP8	PubchemFP9	...	PubchemFP871
0	1	1	0	0	0	0	0	0	0	1	...	0
1	1	1	0	0	0	0	0	0	0	1	...	0
2	1	1	0	0	0	0	0	0	0	1	...	0
3	1	1	0	0	0	0	0	0	0	1	...	0
4	1	1	0	0	0	0	0	0	0	1	...	0
...
128	1	1	1	0	0	0	0	0	0	1	...	0
129	1	1	1	0	0	0	0	0	0	1	...	0
130	1	1	0	0	0	0	0	0	0	1	...	0
131	1	1	0	0	0	0	0	0	0	1	...	0
132	1	1	1	0	0	0	0	0	0	1	...	0

133 rows x 882 columns

Fig 5.1.3 Dataset

From fig 5.1.3, we can see the dataset prepared for model building using PubChem fingerprint. It is chosen for its comprehensiveness and availability in representing molecular structures. It efficiently encodes structural features as binary vectors, facilitating tasks such as virtual screening in drug discovery and cheminformatics.

5.2 SPLITTING THE DATA X AND Y

The data is split into input features (X) and target variables (Y) to build a predictive model. In this process, the 'pIC50' column is removed from the dataset to create the feature matrix X, which contains all other columns. Meanwhile, the Y variable is assigned the values from the 'pIC50' column, representing the target variable to be predicted. This separation allows for the independent modeling of X's features against Y, aiding in the development of effective predictive models for the given dataset.

```
1 X.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 133 entries, 0 to 132
Columns: 881 entries, PubchemFP0 to PubchemFP880
dtypes: int64(881)
memory usage: 915.5 KB

1 Y.info()

<class 'pandas.core.series.Series'>
RangeIndex: 133 entries, 0 to 132
Series name: pIC50
Non-Null Count  Dtype
-----
133 non-null    float64
dtypes: float64(1)
memory usage: 1.2 KB
```

Fig 5.2.1 X & Y matrix

From fig 5.2.1, we can see that the dataset is split into X matrix and Y matrix. X matrix contains the PubChem fingerprint whose shape is 133 entries and 881 columns. Y matrix contains pIC50 and 133 entries, which is the target variable to be predicated after building the model.

There is a function `remove_low_variance` that utilizes sklearn's Variance Threshold to identify and remove features with low variance from the input data. By specifying a threshold (defaulting to 0.1), it filters out features whose variance falls below this value. This process helps in reducing noise and focusing on informative features for subsequent analysis or model building.

5.3 MODEL BUILDING

5.3.1 RANDOM FOREST REGRESSION

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and improve the performance of the model.

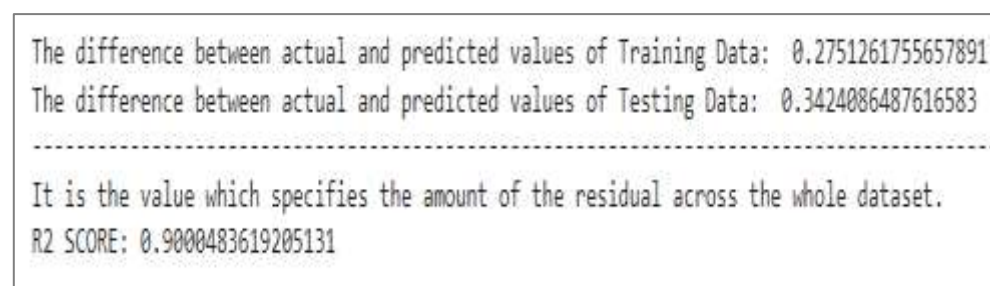


Fig 5.3.1.1 Accuracy of Random Forest Model

From fig 5.3.1.1, the Random Forest Regression model demonstrates promising performance in predicting both training and testing data, with small differences between actual and predicted values. The reported values of 0.275 and 0.342 represent the average amount of residual error across the entire dataset for the training and testing data, respectively. The R2 score of 0.90 indicates that approximately 90% of the variance in the dependent variable is explained by the independent variables in the model, suggesting a strong fit to the data.

Overall, these metrics suggest that the Random Forest Regression model captures a significant portion of the variability in the data and provides accurate predictions for the target variable.

5.3.2 LINEAR REGRESSION

Linear regression is a statistical method for modeling the relationship between a dependent variable and one or more independent variables. It aims to find the best-fitting line (or hyperplane in multiple dimensions) that minimizes the difference between observed and predicted values. This technique is widely used across disciplines for prediction, inference, and understanding of relationships in data.

```
The difference between actual and predicted values of Training Data: 0.248104715616834
The difference between actual and predicted values of Testing Data: 0.3417063576318859
-----
It is the value which specifies the amount of the residual across the whole dataset.
R2 SCORE: 0.9135534751678626
```

Fig 5.3.2.1 Accuracy of Linear Regression Model

From fig 5.3.2.1 Linear Regression is applied to predict the potency of compounds, represented by the target variable pIC50. The reported differences between actual and predicted values for both training and testing data, 0.248 and 0.342 respectively, indicate the average residual error across the dataset. A lower value signifies closer alignment between predicted and actual values. The high R2 score of 0.914 suggests that approximately 91% of the variance in pIC50 values is explained by the independent variables in the model, indicating a strong predictive capability.

Overall, these results indicate that the Linear Regression model effectively captures the relationships between molecular descriptors and compound potency, providing reliable predictions for drug potency in the context of drug discovery.

5.4 COMPARISON OF LINEAR REGRESSION AND RANDOM FOREST

The comparison between the Linear Regression and Random Forest Regression models reveals insights into their performance in predicting compound potency (pIC50) within the context of drug discovery.

Actual Values of pIC50 value	Predicted Value by Linear Regression	Predicted Value by Random Forest
3.4559319556497243	3.536865234375	3.483587630824478
4.522878745280337	4.481201171875	4.3492189358741316
4.0	4.0391845703125	4.0355793246675375
4.920000156997057	4.8486328125	4.925691373244288
6.096910013008056	4.665283203125	4.706925957387202
5.0	4.91571044921875	4.978791731481313
4.7447274948966935	4.6302490234375	4.648126874323894
3.690000009501577	3.60205078125	3.7235777795373255
4.820000071285178	4.849365234375	4.807685717639212
4.346787486224656	4.407470703125	4.45717184751298
4.886056647693163	4.8631591796875	4.873298718964573
3.6000000054478725	3.57568359375	3.635288219559755
4.605548319173784	4.6302490234375	4.648126874323894
7.200659450546418	7.3797607421875	7.078265512884567
4.389999976337708	4.37255859375	4.347841526281427

Fig 5.4.1 Comparison of actual value vs predicted values by linear regression and random forest

Both models perform well in predicting compound potency, but the Linear Regression model exhibits slightly lower residuals and a marginally higher R2 score, suggesting a slightly superior performance in capturing the relationships between molecular descriptors and compound potency.

From fig 5.4.1, The Linear Regression model tends to produce predictions that are slightly closer to the actual values compared to the Random Forest Regression model. Considering the overall performance metrics such as residuals and R2 scores, the Linear Regression model demonstrates slightly better.

This suggests that the Linear Regression model may be more suitable when precise predictions are desired, especially in scenarios where the interpretability of the model is important due to its straightforward nature.

5.5 DATA VISUALIZATION (EXPERIMENTAL VS PREDICTED pIC50 FOR TRAINING DATA)

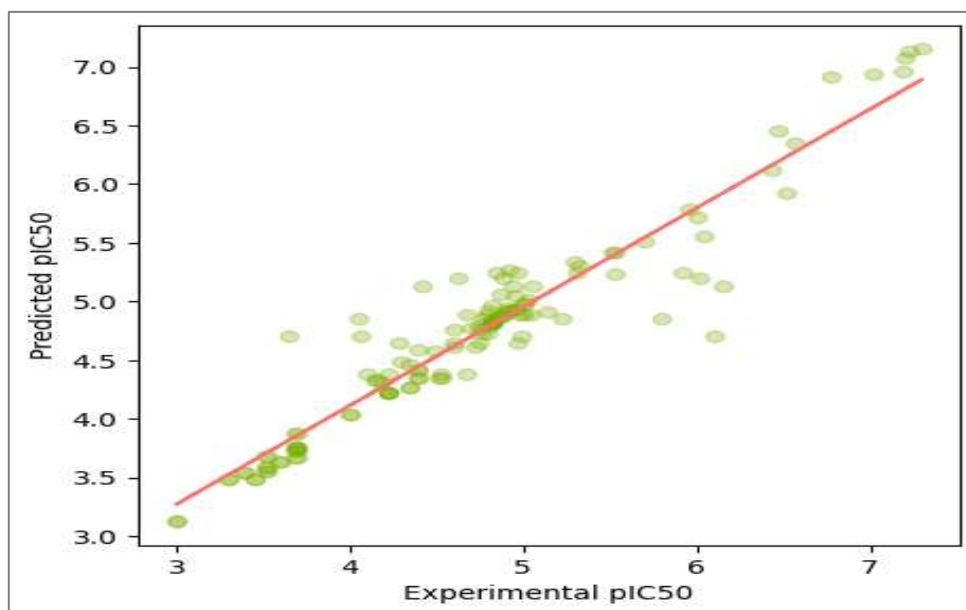


Fig 5.5.1 Experimental vs Predicted pIC50 for Training Data for Linear Regression

From fig 5.5.1 the plot compares experimental and predicted pIC50 values for training data using the Linear Regression Model. Each point represents a compound, with experimental values on the x-axis and predicted values on the y-axis. The trendline shows the linear relationship between them, aiding in assessing the model's accuracy and precision.

6. MODEL DEPLOYMENT

Streamlit is a free and open-source framework to rapidly build and share beautiful machine learning and data science web apps. Building a Streamlit app facilitates bioactivity prediction for inhibiting the SARS coronavirus 3C-like proteinase enzyme.

Users can upload a TXT file containing molecular structures, which are then processed to calculate molecular descriptors using the PaDEL-Descriptor tool. These descriptors are compared against a subset of descriptors used in a previously built model for prediction. The app applies a trained regression model to predict the bioactivity of the uploaded compounds and provides the results to users.

The app's functionality involves uploading molecular data, calculating descriptors, applying a trained model, and displaying prediction outputs. Users can interact with the app through the sidebar, where they upload their input data and initiate the prediction process. Upon prediction, the app presents the calculated descriptors, compares them against the model's subset of descriptors, applies the model to make predictions, and offers a downloadable file containing the prediction results.

Overall, the app streamlines the process of bioactivity prediction for compounds targeting the SARS coronavirus 3C-like proteinase, enhancing accessibility and usability for researchers and practitioners in drug discovery.

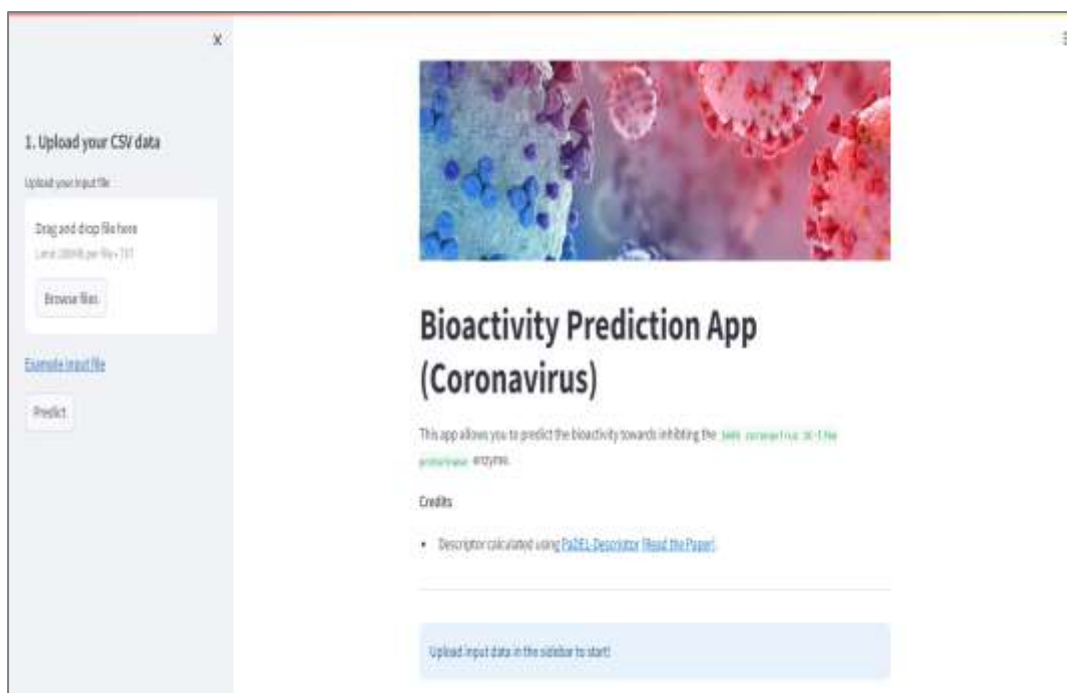


Fig 6.1 User Interface App build with Streamlit

From fig 6.1 we can see the user interface which is built using Streamlit. Users can interact with the app through the sidebar, where they upload their input data and initiate the prediction process.

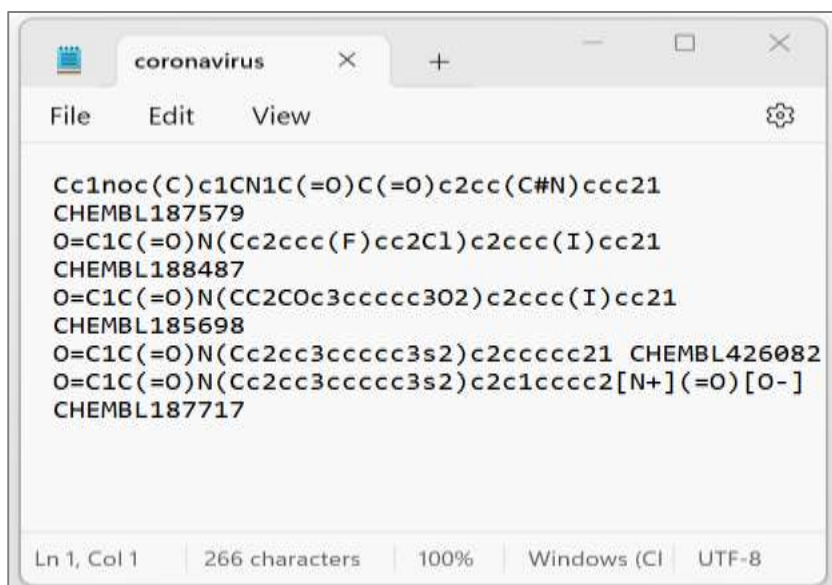


Fig 6.2 txt file used as the input data to upload in the app

From fig 6.2 we can see a txt file named coronavirus containing molecular structures, which are then processed to calculate molecular descriptors using the PaDEL-Descriptor tool.

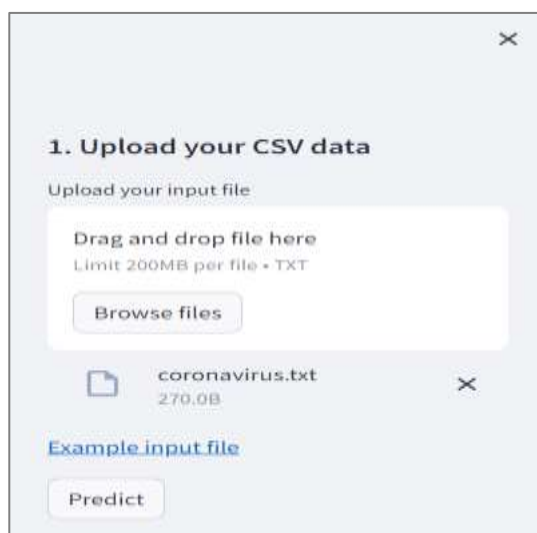


Fig 6.3 The above file is named coronavirus.txt and is uploaded in the app the user has to click on the Predict button for the results to be predicted

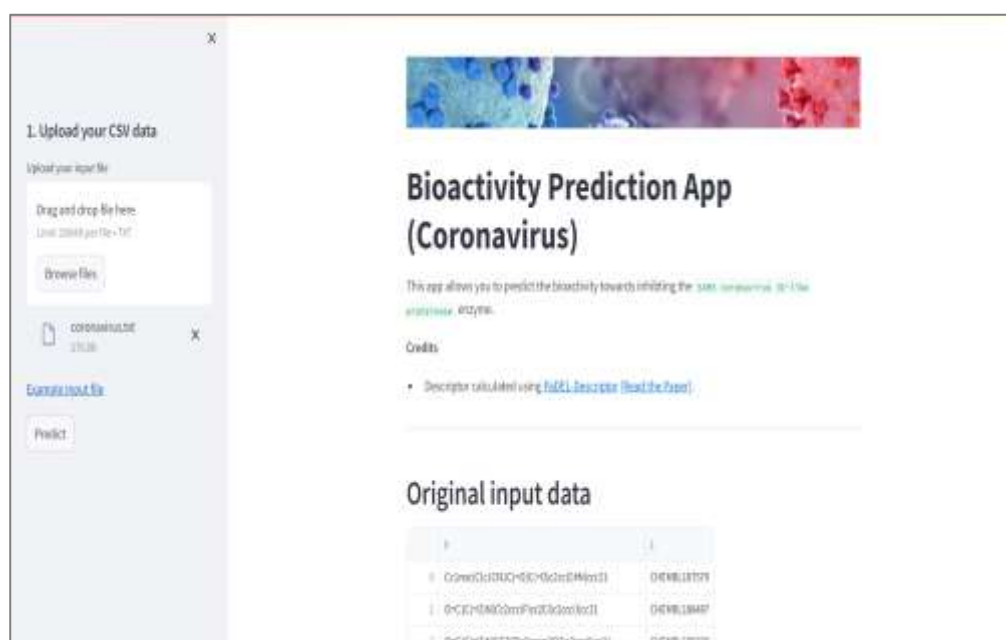


Fig 6.4 After the user clicks on the Predict button the result appears on the app.

Original input data

	0	1
0	<chem>Cc1noc(C)c1CN1C(=O)C(=O)c2cc(C#N)ccc21</chem>	CHEMBL187579
1	<chem>O=C1C(=O)N(Cc2ccc(F)cc2Cl)c2ccc(I)cc21</chem>	CHEMBL188487
2	<chem>O=C1C(=O)N(CC2COc3ccccc3O2)c2ccc(I)cc21</chem>	CHEMBL185698
3	<chem>O=C1C(=O)N(Cc2cc3ccccc3s2)c2ccccc21</chem>	CHEMBL426082
4	<chem>O=C1C(=O)N(Cc2cc3ccccc3s2)c2c1cccc2[N+](=O)[O-]</chem>	CHEMBL187717

Calculated molecular descriptors

	Name	PubchemFP0	PubchemFP1	PubchemFP2	PubchemFP3	PubchemFP4	PubchemFP5
0	CHEMBL187579	1	1	0	0	0	
1	CHEMBL188487	1	1	0	0	0	
2	CHEMBL185698	1	1	0	0	0	
3	CHEMBL426082	1	1	0	0	0	
4	CHEMBL187717	1	1	0	0	0	

(5, 882)

Fig 6.5 Results Part 1

Subset of descriptors from previously built models

	PubchemFP2	PubchemFP12	PubchemFP14	PubchemFP15	PubchemFP16	PubchemFP19	PubchemFP20
0	0	0	1	1	0	1	
1	0	0	1	0	0	1	
2	0	1	1	0	0	1	
3	0	1	1	0	0	1	
4	0	1	1	1	0	1	

(5, 285)

Fig 6.6 Results Part 2

From fig 6.5, and 6.6, Upon clicking the predict button after uploading the file, the app presents the original data uploaded, calculated molecular descriptors, and also the subset of molecular descriptors from the previously built models and compares them against the model's new descriptors.

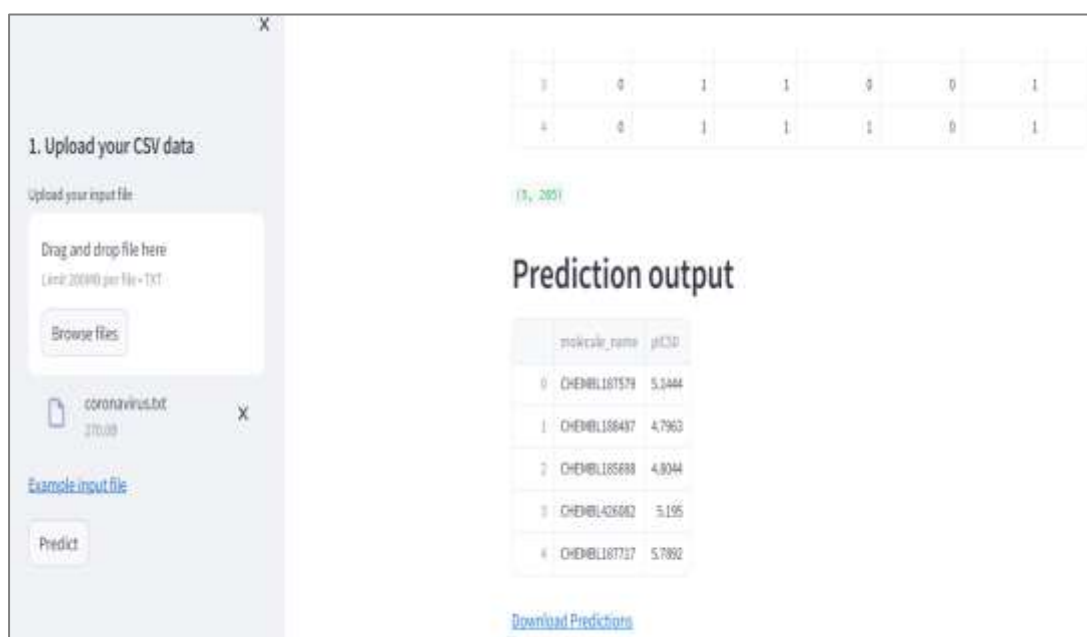


Fig 6.7 Prediction Output

From fig 6.7, shows the predicted pIC50 values for the molecules uploaded in the inputted file and offers a downloadable CSV file containing the prediction results.

7. CONCLUSION

- This project represents a comprehensive endeavor in leveraging computational tools and machine learning techniques for drug discovery, particularly in addressing the urgent global health challenge posed by the COVID-19 pandemic. Through a meticulous review of existing literature, the project identified key methodologies and trends in computational drug discovery, emphasizing the importance of virtual screening, predictive modeling, and drug repurposing strategies.
- The database utilized in this project, ChEMBL, stands as a cornerstone in modern drug discovery endeavors, amalgamating chemical, bioactivity, and genomic data to facilitate the translation of genomic insights into effective therapeutic interventions. Its inception from collaborative efforts between academic institutions and pharmaceutical companies underscores its comprehensive nature and its pivotal role in advancing cheminformatics and drug discovery.
- The process of data collection from the ChEMBL database exemplifies the meticulous approach taken to gather pertinent information regarding compounds targeting the SARS coronavirus 3C-like proteinase. Leveraging both the database's web interface and programmatically accessible API, researchers meticulously retrieved bioactivity data specific to the target protein, laying the foundation for subsequent analysis and modeling.

- Data filtering and preprocessing played a crucial role in preparing the dataset for modeling endeavors. Through a systematic approach, relevant columns were extracted, and stringent criteria were applied to categorize compounds based on their bioactivity against the target enzyme. The transformation of raw data into a structured format, coupled with the calculation of molecular descriptors, enabled researchers to harness the power of computational analysis and predictive modeling.
- Exploratory data analysis (EDA) played a crucial role in elucidating the characteristics of the dataset, revealing insights into the distribution of molecular descriptors and their relationship with compound potency. Frequency plots, scatter plots, boxplots, and statistical analyses provided valuable insights into the dataset's structure and properties, facilitating informed decision-making during modeling.
- Modeling efforts culminated in the development and evaluation of regression models, including Random Forest Regression and Linear Regression. These models demonstrated strong predictive performance, with accuracies exceeding 90% and high R^2 scores, indicating robust fits to the data.
- Furthermore, the deployment of a Streamlit web application streamlined the process of bioactivity prediction, providing users with an intuitive interface to upload molecular data, calculate descriptors, apply a trained model, and obtain prediction outputs seamlessly.

8. REFERENCES

- C. Nantasenamat and V. Prachayasittikul, “Maximizing computational tools for successful drug discovery,” *Expert Opinion on Drug Discovery*, vol. 10, no. 4, pp. 321–329, Feb. 2015, doi: 10.1517/17460441.2015.1016497.
- A. K. A. Khan and N. H. A. H. Malim, “Comparative studies on resampling techniques in machine learning and deep learning models for Drug-Target Interaction Prediction,” *Molecules/Molecules Online/Molecules Annual*, vol. 28, no. 4, p. 1663, Feb. 2023, doi: 10.3390/molecules28041663.
- C. A. Lipinski, F. Lombardo, B. W. Dominy, and P. J. Feeney, “Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings 1PII of original article: S0169-409X(96)00423-1. The article was originally published in *Advanced Drug Delivery Reviews* 23 (1997) 3–25. 1,” *Advanced Drug Delivery Reviews*, vol. 46, no. 1–3, pp. 3–26, Mar. 2001, doi: 10.1016/s0169-409x(00)00129-0.
- C.-Y. Jia, J. Li, G. Hao, and G. Yang, “A drug-likeness toolbox facilitates ADMET study in drug discovery,” *Drug Discovery Today*, vol. 25, no. 1, pp. 248–258, Jan. 2020, doi: 10.1016/j.drudis.2019.10.014.
- R. Aghdam, M. Habibi, and G. Taheri, “Using informative features in machine learning based method for COVID-19 drug repurposing,” *Journal of Cheminformatics*, vol. 13, no. 1, Sep. 2021, doi: 10.1186/s13321-021-00553-9.

- H. Lv et al., “Application of artificial intelligence and machine learning for COVID-19 drug discovery and vaccine design,” *Briefings in Bioinformatics*, vol. 22, no. 6, Aug. 2021, doi: 10.1093/bib/bbab320.
- S. Monteleone, T. F. Kellici, M. Southey, M. J. Bodkin, and A. Heifetz, “Fighting COVID-19 with Artificial Intelligence,” in *Methods in molecular biology*, 2021, pp. 103–112. doi: 10.1007/978-1-0716-1787-8_3.
- M. Sreepadmanabh, A. K. Sahu, and A. Chande, “COVID-19: Advances in diagnostic tools, treatment strategies, and vaccine development,” *Journal of Biosciences/Journal of Biosciences*, vol. 45, no. 1, Nov. 2020, doi: 10.1007/s12038-020-00114-6.