



# Bioactivity Prediction against AmpC Beta-lactamase through Machine Learning

**Under the guidance of:**

**Dr. PRATIMA GUPTA**

Associate Professor

National Institute of Technology, Raipur

**PRESENTED BY:**

BONELA SPANDANA (18112015)

PADMANABHAN M (18112033)

PREM SINGH ANANT (18112037)

SHIVANI PARIHAR (18112046)

KARTIK KUMAR SINGH (17112005)

# **CONTENTS:-**

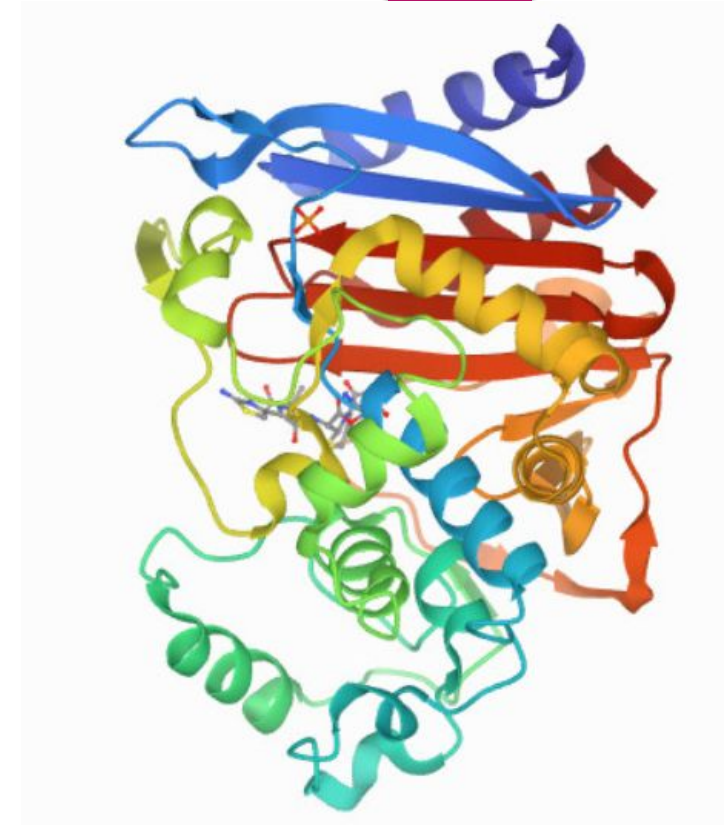
- **OBJECTIVE**
- **INTRODUCTION**
- **TOOLS USED**
- **LIBRARIES AND PACKAGES USED**
- **METHODOLOGY**
- **GENERATING MOLECULAR FINGERPRINTS**
- **X AND Y DATA MATRICES**
- **DATA SPLITTING**
- **BUILDING MODEL: RANDOM FOREST REGRESSION**
- **MODEL PREDICTION**
- **MODEL PERFORMANCE**
- **DATA VISUALIZATION**
- **RESULT**

# Objective

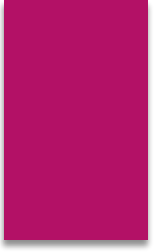
- Antimicrobial resistance (AMR) threatens the effective prevention and treatment of an ever-increasing range of infections caused by bacteria, parasites, viruses and fungi. AMR is an increasingly serious threat to global public health that requires action.
- Bacteria may manifest resistance to antibacterial drugs through a variety of mechanisms like production of beta lactamase.
- Hence it is essential to find the bioactivity of compounds against the enzyme beta-lactamase.
- In recent years, artificial intelligence (AI) has shown significant performance in AMR control.
- For example, collecting clinical data to build clinical decision support systems could help physicians monitor trends in AMR to increase the rational use of antibiotics.
- Furthermore, AI applications are widely used for designing new antibiotics and synergistic drug combination investigations.
- We have tried to develop a Machine Learning Model using Random Forest Regression which can Predict the Bio-Activity of Certain Compounds against AmpC Beta-Lactamase which is a key enzyme of Antimicrobial Resistance in Bacteria.

# Introduction

- The ability of microorganisms like bacteria to develop mechanisms against the treatment is becoming a concern globally, This is referred as Anti microbial resistance aka AMR.
- Microorganisms like bacteria use a myriad of mechanisms to resist the effects of antimicrobial agents.
- Some bacteria may acquire the genes which encode enzymes such as *β-lactamases* which can nullify the effects of the antimicrobial agents used.
- ***β-lactamases*** are a diverse family of microbial enzymes that hydrolyze the cyclic amide bond of susceptible β-lactam antibiotics.
- Recent advances in molecular biology allow classification which are based on sequence schemes and they cannot be distorted by mutations or kinetic activity.



Crystal Structure of AmpC  
beta-lactamase from E. coli

- 
- Sequencing currently reveals four classes of  $\beta$ -lactamases, designated A to D, With Class A and Class C being the most clinically Important.
  - The class C beta-lactamases (AmpC) are clinically significant because Gram-negative-bacteria which confer AmpC are also resistant to penicillin, cephalosporin, cephameycin and monobactam groups.
  - Statistical data on the bioactivity of certain compounds against AmpC beta-lactamases can give an understanding of how effective the compound is.
  - The bioactivity of a compound refers to the beneficial or adverse effects of it on a biomolecule. Standardized reference points such as IC50 value, pIC50 value, pChEMBL value are used to describe bioactivity of a compound.
  - In this project, we have developed a Random Forest Regression model in machine learning 1 to predict the bioactivity of certain compounds like flavonoids and terpenoids against AmpC beta-lactamase.

# TOOLS USED

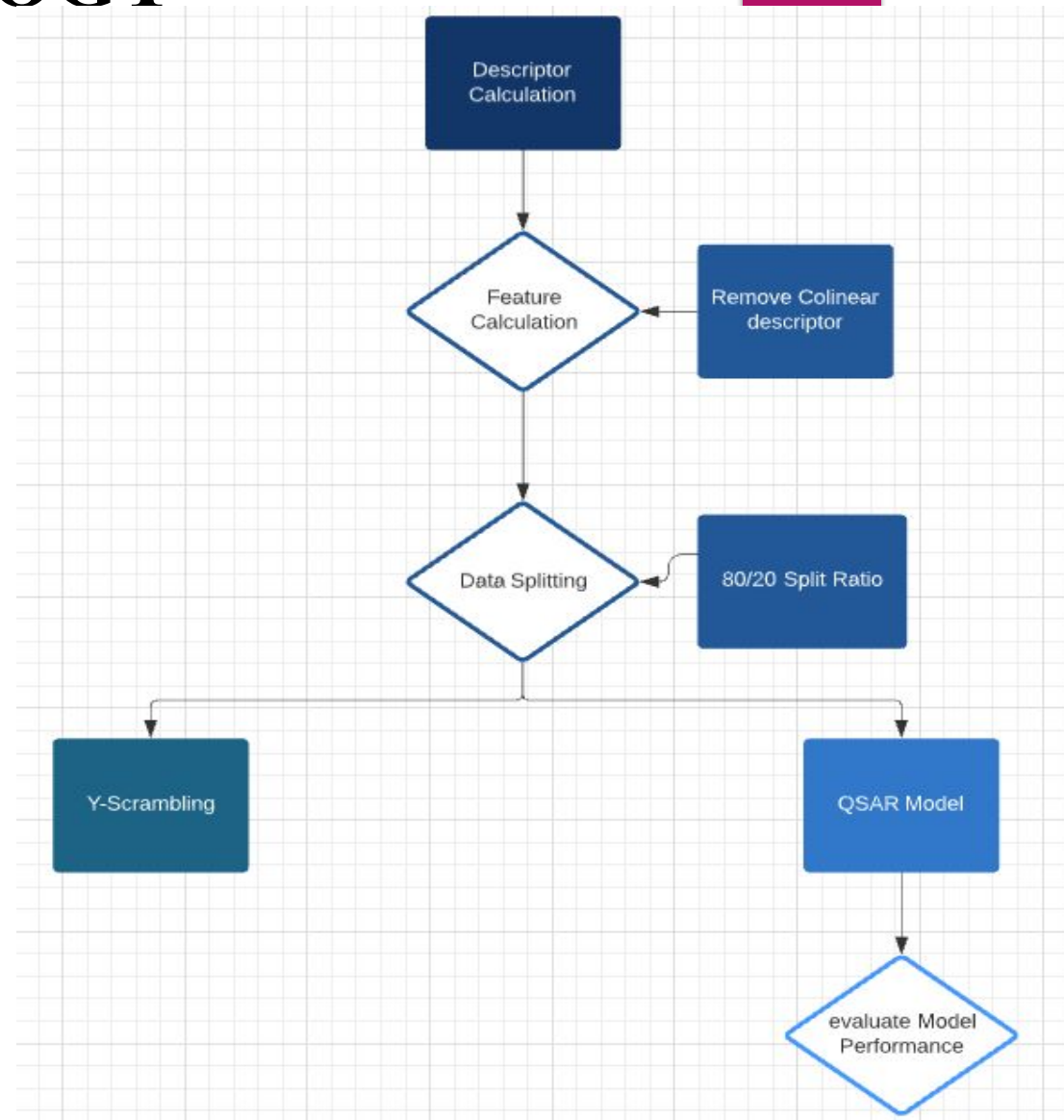
- **PaDEL-Descriptor** : software for calculating molecular descriptors and fingerprints.
- **PubChem fingerprints**: PubChem System creates a binary substructure fingerprint.
- PubChem uses these fingerprints for similarity and neighbouring searching.
- **ChEMBL Database**: A large-scale bioactivity database which is developed to provide vital data during drug discovery.
- **MODEL USED**: QSAR (Random Forest Regression ML Model)

# Libraries and packages used

- **Numpy:** NumPy is a general-purpose array-processing package. It provides high-performance multidimensional array objects and tools to work with the arrays. NumPy is used to process arrays that store values of the same datatype. NumPy facilitates math operations on arrays and their vectorization
- **Pandas:** Pandas is a perfect tool for data wrangling or munging. It is designed for quick and easy data manipulation, reading, aggregation, and visualization. Pandas take data in a CSV or TSV file or a SQL database and create a Python object with rows and columns called a data frame.
- **Scikit learn:** Scikit Learn is a robust machine learning library for Python. It features ML algorithms like SVMs, random forests, k-means clustering, spectral clustering, mean shift, cross-validation and more. Classification, clustering, Regression, dimensionality reduction ,model selection and pre-processing can be done with this.
- **Rdkit:** RDKit is a collection of cheminformatics and machine-learning software written in C++ and Python.
- **Matplotlib:** Matplotlib is one of the most popular and oldest plotting libraries in Python which is used in Machine Learning. In Machine learning, it helps to understand the huge amount of data through different visualisations.

# METHODOLOGY

- Data Extraction: from the ChEMBL database (the bioactivity of 64280 compounds against beta-lactamase AmpC).
- Importing Packages and libraries: Pandas, NumPy, Scikit Learn.
- We have selected pubchem fingerprints and pChEMBL value.
- pChEMBL Value: This value allows a number of roughly comparable measures of half-maximal response concentration/potency/affinity to be compared on a negative logarithmic scale (e.g. an IC<sub>50</sub> measurement of 1 nM has a pChEMBL value of 9). The pChEMBL value is currently defined as follows:  $-\log_{10}$  (molar IC<sub>50</sub>, XC<sub>50</sub>, EC<sub>50</sub>, AC<sub>50</sub>, K<sub>i</sub>, K<sub>d</sub> or Potency)





# Data cleaning

- Data Cleaning : deleting inaccurate, corrupted, improperly formatted, duplicate, or incomplete data from a dataset.

	molecule_chembl_id	canonical_smiles	standard_value	pchembl_value
1	CHEMBL263746	O=C([O-])C1=CS[C@@H]2/C(=C\c3cnc4n3CCOC4)C(=O)...	5.00	8.30
2	CHEMBL331090	O=C([O-])C1=CS[C@@H]2/C(=C\c3cnc4n3CCNC4)C(=O)...	6.00	8.22
3	CHEMBL124416	O=C([O-])C1=CS[C@@H]2/C(=C\c3cc4n(n3)CCC4)C(=O)...	2.00	8.70
4	CHEMBL404	C[C@]1(Cn2ccnn2)[C@H](C(=O)O)N2C(=O)C[C@H]2S1(...	84000.00	4.08
5	CHEMBL122450	O=C([O-])C1=CS[C@@H]2/C(=C\c3cn4c5c(sc4n3)CCC5...	2.00	8.70
...	...	...	...	...
62475	CHEMBL4443633	CC(=O)OCC1=C(C(=O)O)N2C(=O)[C@@H](NC(=O)Cc3ccc...	4.78	NaN
62476	CHEMBL4442878	CCCC(=O)N[C@@H]1C(=O)N2C(C(=O)O)=C(COC(C)=O)CS...	0.64	NaN
62477	CHEMBL4448907	CC(=O)OCC1=C(C(=O)O)N2C(=O)[C@@H](NC(=O)c3cccc...	0.52	NaN
62478	CHEMBL4434858	CC(=O)OCC1=C(C(=O)O)N2C(=O)[C@@H](NC(=O)Cc3ccc...	7.33	NaN
62479	CHEMBL4436556	CC(=O)OCC1=C(C(=O)O)N2C(=O)[C@@H](NC(=O)[C@H](...	0.87	NaN

# GENERATING MOLECULAR FINGERPRINTS

- Generating pubchem fingerprints: Using PaDel-Descriptor.

	PubchemFP0	PubchemFP1	PubchemFP2	PubchemFP3	PubchemFP4	PubchemFP5	PubchemFP6	PubchemFP7	PubchemFP8	PubchemFP9	...
0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	...
1	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	...
2	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	...
3	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	...
4	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	...
...	...	...	...	...	...	...	...	...	...	...	...
62030	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	...
62031	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	...
62032	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	...
62033	1.0	1.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	1.0	...
62034	1.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	...

62034 rows × 882 columns

# X and Y Data Matrices

- Preparing the X and Y Data Matrices: pubchem fingerprints in the X variable and pchembl\_value in the Y variable.

	PubchemFP2	PubchemFP12	PubchemFP15	PubchemFP16	PubchemFP19	PubchemFP20	PubchemFP23	PubchemFP33
0	0.0	0.0	1.0	0.0	1.0	1.0	0.0	1.0
1	0.0	0.0	1.0	1.0	1.0	0.0	0.0	1.0
2	0.0	0.0	1.0	1.0	1.0	1.0	0.0	1.0
3	0.0	0.0	1.0	0.0	1.0	0.0	0.0	1.0
4	0.0	0.0	1.0	0.0	1.0	0.0	0.0	1.0
...	...	...	...	...	...	...	...	...
62030	1.0	1.0	0.0	0.0	1.0	1.0	0.0	0.0
62031	1.0	1.0	0.0	0.0	1.0	1.0	0.0	0.0
62032	1.0	1.0	0.0	0.0	1.0	1.0	0.0	0.0
62033	0.0	0.0	1.0	1.0	1.0	1.0	1.0	1.0
62034	1.0	1.0	1.0	0.0	1.0	1.0	0.0	0.0

62034 rows × 243 columns

**X variable**

Y	
0	8.30
1	8.22
2	8.70
3	4.08
4	8.70
...	...
62030	4.24
62031	5.27
62032	5.34
62033	5.19
62034	10.30
Name: pchembl_value, Length: 62034, dtype: float32	

**Y variable**

# DATA SPLITTING

- Data splitting: 80% goes to the training of the model and the rest 20% for testing.

```
X_train.shape, Y_train.shape
```

```
((49627, 243), (49627,))
```

```
X_test.shape, Y_test.shape
```

```
((12407, 243), (12407,))
```

# Building Model: Random Forest Regression

- Random forest is a supervised learning technique that can be used to classify and predict data. However, it is mostly employed to solve classifying issues.
- A forest, as we all know, is made up of trees, and more trees equals a more strong forest. Similarly, the random forest method constructs decision trees from data samples, extracts predictions from each, and then votes on the best option.
- It's an ensemble method that's superior than a single decision tree because it averages the results to reduce overfitting.

## Building a Regression Model using Random Forest.

```
model = RandomForestRegressor(n_estimators=100)
model.fit(X_train, Y_train)
r2 = model.score(X_test, Y_test)
r2
```

```
-0.10102670869263553
```

# Model prediction

- Model prediction: Getting the Array of predicted pChEMBLvalue.

```
array([5.0955    , 5.01000001, 5.03403334, ..., 5.38475    , 5.01042501,  
       5.11425004])
```

# Model performance

```
print('Mean squared error (MSE): %.2f'  
      % mean_squared_error(Y_test, Y_pred))  
print('Coefficient of determination (R^2): %.2f'  
      % r2_score(Y_test, Y_pred))
```

Mean squared error (MSE): 0.75

## Mean Square Error(MSE): 0.75

- MSE is calculated by taking the average of the square of the difference between the original and predicted values of the data. The lower the MSE, the better a model fits a dataset.

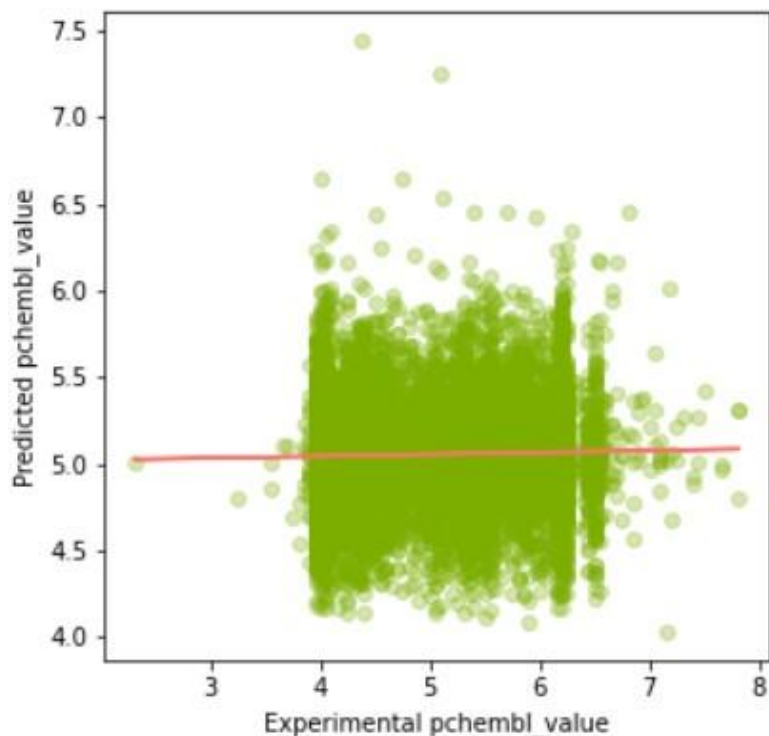
# Data Visualization

```
import matplotlib.pyplot as plt
import numpy as np
```

```
plt.figure(figsize=(5,5))
plt.scatter(x=Y_test, y=Y_pred, c="#7CAE00", alpha=0.3)
```

```
z = np.polyfit(Y_test, Y_pred, 1)
p = np.poly1d(z)
```

```
plt.plot(Y_test, p(Y_test), "#F8766D")
plt.ylabel('Predicted pchembl_value')
plt.xlabel('Experimental pchembl_value')
```

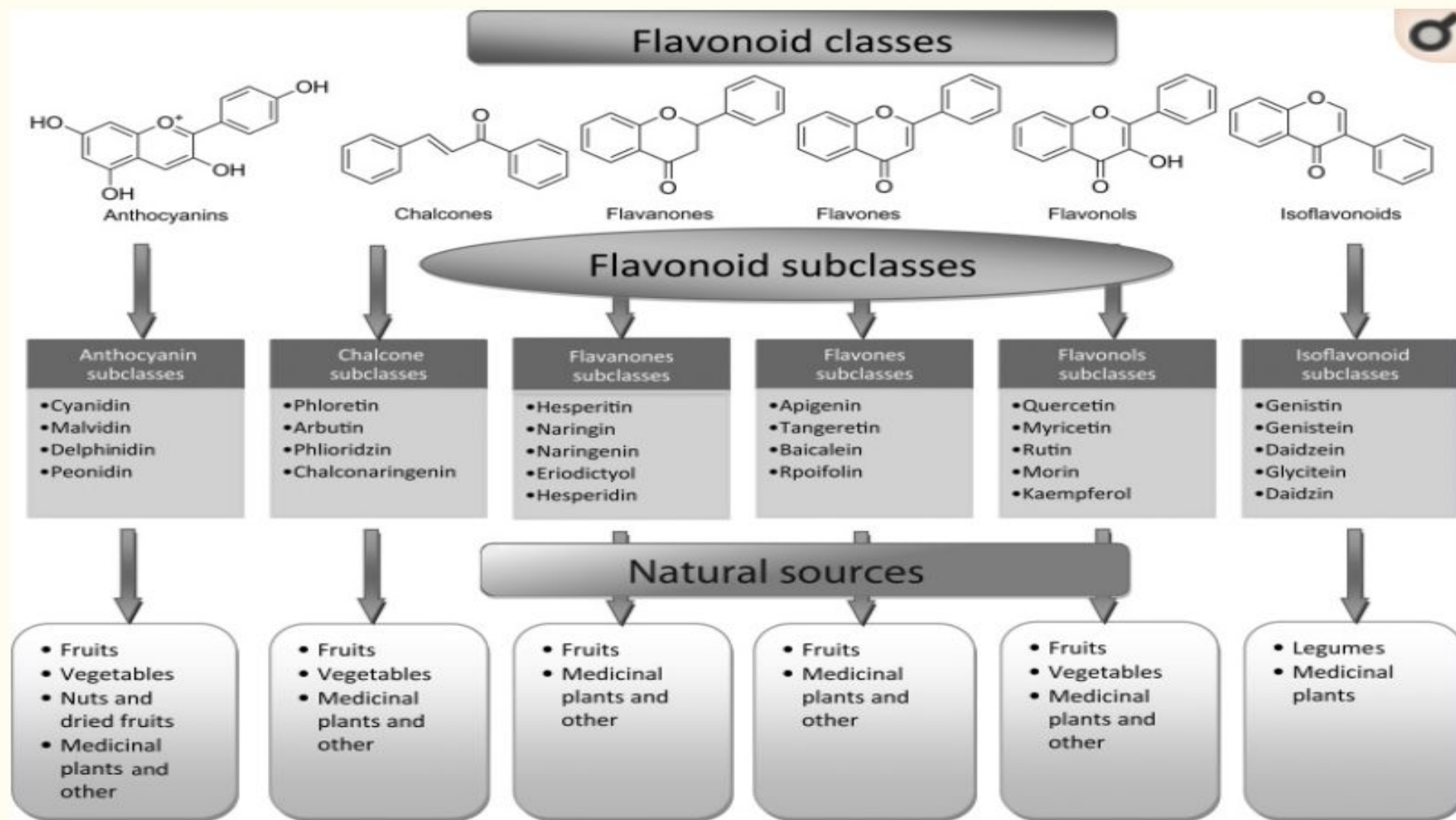




# Compounds for bioactivity prediction

## FLAVANOIDS:

Flavonoids are an important class of natural products; particularly, they belong to a class of plant secondary metabolites having a polyphenolic structure, widely found in fruits, vegetables and certain beverages.





## **TERPENOIDS**

- Naturally occurring organic compound derived from the 5-carbon compound isoprene
- Also known as isoprenoids, are the most numerous and structurally diverse natural products found in many plants.
- 
- Several studies, in vitro, preclinical, and clinical have confirmed that this class of compounds displays a wide array of very important pharmacological properties.
- 
- The diverse collection of terpenoid structures and functions have provoked increased interest in their commercial use resulting in some with established medical applications being registered as drugs on the market.

# Result

- The threshold pchembl value is decided as 5. So, as we can see from the result there are 62 potential compounds whose pchembl value is greater than 5.

Flavanonols	Column1
molecule_name	pchembl_value
Dihydromorin	5.044025039
Garbanzol	5.036800034
Ampelopsin	5.044025039
Aromadendrin	5.162486679
Fustin	5.044025039
Taxifolin	5.050085027
Pinobanksin	4.989250014

# FUTURE PERSPECTIVE

- ❖ In this project we have worked with single fingerprints.
- ❖ This project has the scope of working with multiple fingerprints
- ❖ Possibility of applying the results into renowned softwares like Schrodinger.
- ❖ Results provide vital data to discover high-quality, novel molecules more rapidly, at a much lower cost.
- ❖ Once the desired compound is identified, it can be tested in laboratories for further efficiency.
- ❖ Developed forms of computational platforms helps in exploring a wider biochemical space.
- ❖ Prediction of the bioactivity and molecular behaviour of a plethora of compounds with high degree of accuracy.

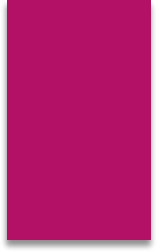
# CONCLUSION

- ❖ In this modern age of genomics more researchers are trying to understand the phenomena of antimicrobial resistance.
- ❖ The evolution and evolutionary drivers of AMR in populations of bacteria are gaining more attraction.
- ❖ Necessity to understand the functioning and bioactivity of compounds against resistant mechanisms such as the production of beta-lactamase by bacteria.
- ❖ A large set of data needs to be processed at a single time for accurate decision making along the development of such compounds.
- ❖ Computational science like machine learning can help in this process.
- ❖ Machine learning allows a user to feed the computer algorithm an immense amount of data.
- ❖ Data is analysed by the computer and then data driven decisions and recommendations are made only on the basis of the input data.
- ❖ A collaborative effort using computational science and biochemistry will lead to novel discoveries that will help humanity to better tackle obstacles such as antimicrobial resistance.

# Literature

- Many bioinformatic methods have been developed for predicting AMR phenotypes from whole-genome sequences and AMR genes.
- There were many Machine Learning Algorithms are used for AMR research some of them are naïve Bayes (NB), decision trees (DT), random forests (RF), support vector machines (SVM), and artificial neural networks (ANN).
- Choisy et al. used naïve Bayes to estimate the probabilities of ineffective treatment due to AMR.
- Reynolds et al. employed a decision tree model to estimate healthcare utilization and cost for AMR.
- Her et al. predicted whether E. coli are resistant to antibiotics using SVM models.
- Recently, Stokes et al. identified new antibiotics without any prior assumptions using a deep learning approach.

# Contribution



## REFERENCES: -

1. Fred C. Tenover, Mechanisms of Antimicrobial Resistance in Bacteria, Division of Healthcare Quality Promotion, Centers for Disease Control and Prevention, Atlanta, Georgia, USA, The American Journal of Medicine, Vol 119 (6A), 2006
2. John Osei Sekyere, Jonathan Asante, Emerging mechanisms of antimicrobial resistance in bacteria and fungi: advances in the era of genomic, Faculty of Pharmacy & Pharmaceutical Sciences, Kwame Nkrumah University of Science & Technology, Kumasi, Ghana, Future Microbiol, 2018
3. DM Livermore, DM Livermore, beta-Lactamases in laboratory and clinical resistance, Department of Medical Microbiology, London Hospital Medical College, United Kingdom, Clinical microbiology reviews, ASM Journal, Vol 8 No 4, 1995
4. Ambler, R. P. 1980. The structure of b-lactamases. Philos. Trans. R. Soc. London Ser. B289:321–331.
5. Jacoby GA, Munoz-Price LS. The new beta-lactamases. N Engl J Med. 2005;352:380–91
6. Todeschini, R.; Consonni, V. Handbook of Molecular Descriptors; Wiley-VCH: Weinheim, 2000.
7. Carhartt, R. E.; Smith, D. H.; Venkataraghavan, R. J Chem Inf Comput Sci 1985, 25, 64.
8. Kier, L. B.; Hall, L. H. Molecular Connectivity in Structure-Activity Analysis; Wiley: New York, 1986.
9. Medina-Franco JL, Maggiora GM, Molecular similarity analysis. In: Bajorath J Chemoinformatics for drug discovery. Wiley, Hoboken, pp 343–399, 2014.
10. Ulf Norinder, Ola Spjuth, Fredrik Svensson. Using Predicted Bioactivity Profiles to Improve Predictive Modeling. Journal of Chemical Information and Modeling, 60 (6) ,2830-2837, 2020.
11. <https://peerj.com/articles/2322>
12. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5465813/>