Minor Project

On

Bioactivity Prediction against AmpC Beta-lactamase through Machine Learning

Project report submitted in partial fulfilment of the requirement for the degree of

Bachelor of Technology

In

Biotechnology

DEPARTMENT OF BIOTECHNOLOGY

NATIONAL INSTITUTE OF TECHNOLOGY RAIPUR

(November 2021)

Submitted By

Bonala Spandhana **(Roll no. 18112015)**

Padmanabhan M **(Roll no. 18112033)**

Prem Singh Anant **(Roll no. 18112037)**

Shivani Parihar **(Roll no. 18112046)**

Kartik Kumar Singh **(Roll no. 17112024)**

Under the supervision of

Dr. Pratima Gupta

Associate Professor

CERTIFICATE

DEPARTMENT OF BIOTECHNOLOGY ENGINEERING

NATIONAL INSTITUTE OF TECHNOLOGY, RAIPUR

(An Autonomous Institute)

RAIPUR, CHHATTISGARH - 492013

This is to certify that the work contained in the project report titled

“Bioactivity Prediction against AmpC Beta-lactamase through Machine Learning”

submitted by

Bonela Spandana **(Roll no. 18112015)**

Padmanabhan M **(Roll no. 18112033)**

Prem Singh Anant **(Roll no. 18112037)**

Shivani Parihar **(Roll no. 18112046)**

Kartik Kumar Singh **(Roll no. 17112024)**

has been carried out under our supervision and this work has not been submitted elsewhere for a degree.

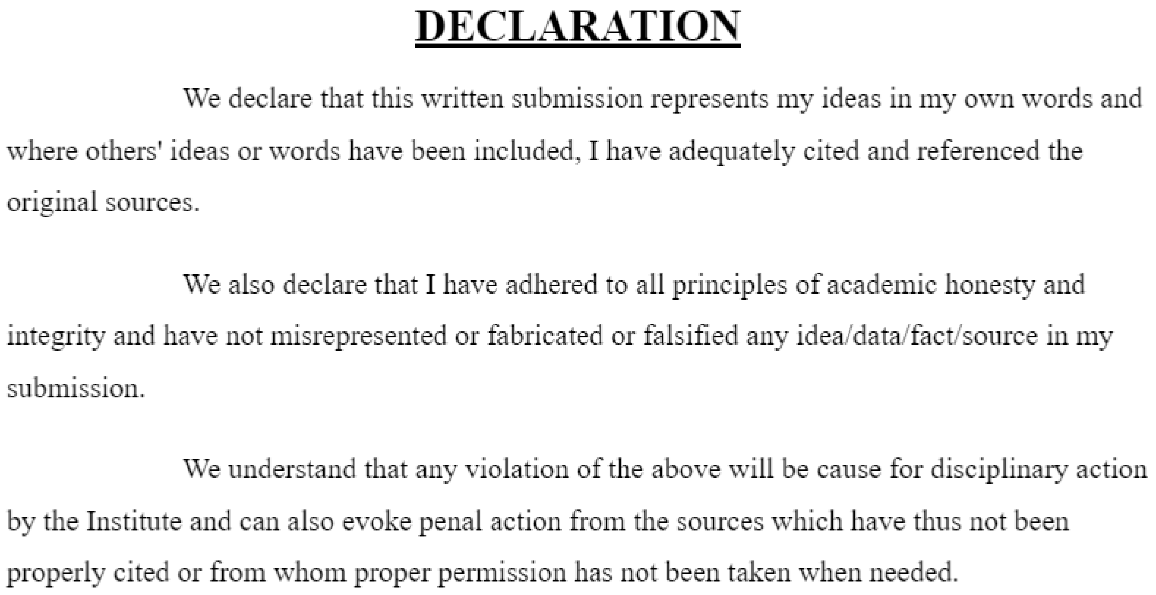
…………………………….

Dr. Pratima Gupta

Associate professor

Department of biotechnology

National Institute of Technology, Raipur



DECLARATION

We declare that this written submission represents my ideas in my own words and where others' ideas or words have been included, I have adequately cited and referenced the original sources.

We also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission.

We understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

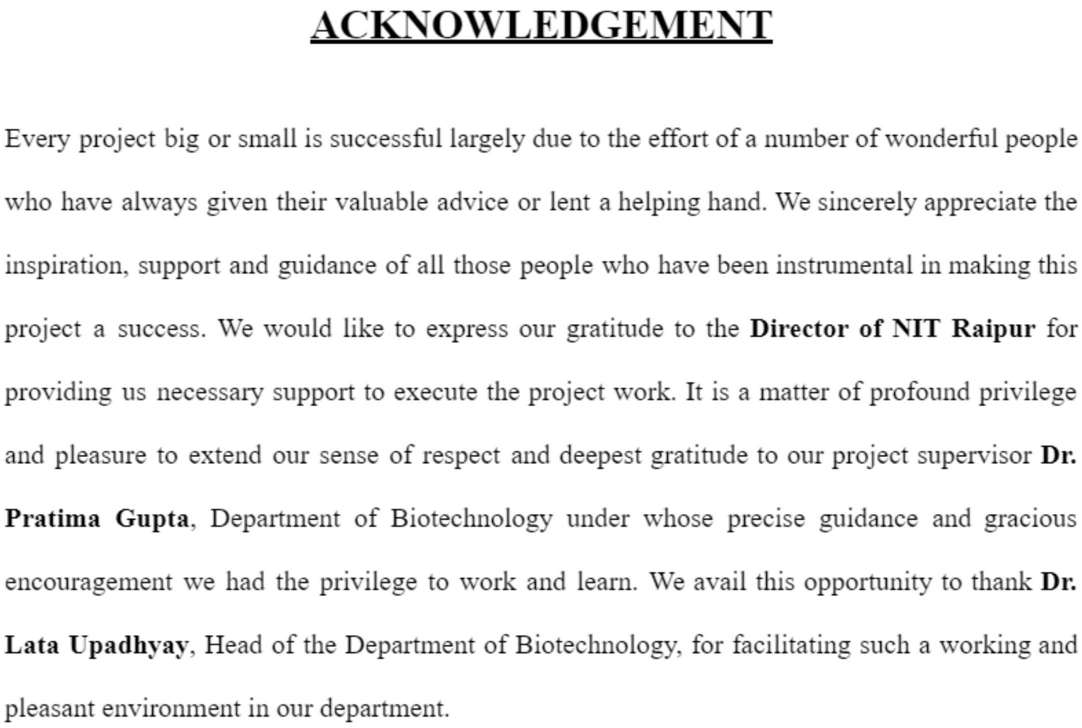
Bonela Spandana **(Roll no. 18112015)**

Padmanabhan M **(Roll no. 18112033)**

Prem Singh Anant **(Roll no. 18112037)**

Shivani Parihar **(Roll no. 18112046)**

Kartik Kumar Singh **(Roll no. 17112024)**



ACKNOWLEDGEMENT

Every project big or small is successful largely due to the effort of a number of wonderful people who have always given their valuable advice or lent a helping hand. We sincerely appreciate the inspiration, support and guidance of all those people who have been instrumental in making this project a success. We would like to express our gratitude to the **Director of NIT Raipur** for providing us necessary support to execute the project work. It is a matter of profound privilege and pleasure to extend our sense of respect and deepest gratitude to our project supervisor **Dr. Pratima Gupta**, Department of Biotechnology under whose precise guidance and gracious encouragement we had the privilege to work and learn. We avail this opportunity to thank **Dr. Lata Upadhyay**, Head of the Department of Biotechnology, for facilitating such a working and pleasant environment in our department.

Bonela Spandana **(Roll no. 18112015)**

Padmanabhan M **(Roll no. 18112033)**

Prem Singh Anant **(Roll no. 18112037)**

Shivani Parihar **(Roll no. 18112046)**

Kartik Kumar Singh **(Roll no. 17112024)**

Table of Contents

INTRODUCTION

The task of treating diseases caused by an organism are becoming more cumbersome in the modern day due to the organism's ability to develop resistance mechanisms against the particular drug or antimicrobial agent used. This ability of microorganisms like bacteria to develop mechanisms against the treatment is becoming a concern globally. This topic of concern is called **Antimicrobial Resistance aka AMR.** Antimicrobial agents have played a key role in fighting against infectious diseases but their frequent and misleading usage has led organisms to develop a myriad of resistance mechanisms that enable them to resist the effects.

WHY IS ANTIMICROBIAL RESISTANCE AN IMPORTANT TOPIC OF CONCERN?

AMR represents one of the biggest challenges to global public health, it was estimated to have claimed more than 7 lakh lives in the year of 2014 and this number is estimated to increase if appropriate measures are not taken to tackle it. Hence it is essential to protect the integrity of the antimicrobials that are currently in use as the discovery of novel antibiotics have taken a step back in the past few decades (1).

ANTIMICROBIAL RESISTANCE IN BACTERIA

With the ever-increasing usage of antimicrobial agents, microorganisms specifically bacteria have developed resistance mechanisms through mutations that occur in the genes present in chromosomes or by the integration of extrinsic genetic material such as plasmid which paves way for the expression for a variety of resistance genes.

Microorganisms use a myriad of mechanisms to resist the effects of antimicrobial agents. First, Some bacteria may acquire the genes which encode enzymes such as *β-lactamases* which can nullify the effects of the antimicrobial agents used. Second, Bacteria at times have a pump which can efflux out the antibacterial agent from the cell before it exerts its effect on the target site. Third, resistant bacteria are capable of producing many genes which allow a metabolic pathway that can generate cell walls with alternating structure which inhibits the binding of antibacterial agents to the binding site and fourth mechanism is the downregulation of porin genes by mutations which does not allow antibacterial agent to reach the target site (2).

As described above *ß-lactamases* represent a class of enzymes which are essential for antimicrobial resistance and it stands as one of the key points of interest in this project.The enzyme ***ß-lactamase*** comes under a large group of microbial enzymes which has the unique ability to hydrolyze the cyclic amide bond present in the beta-lactam antibiotics. With the improvement of technology in the molecular biology and protein chemistry fields it was found that they actually belong to a diverse family of proteases (serine) in which PBPs aka penicillin-binding proteins are also present; these are also a target of ß-Lactam drugs (3).

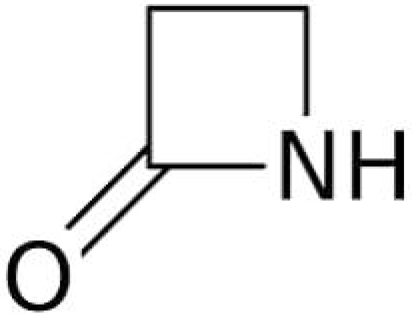


Fig: A beta-lactam ring

Above is a pictorial representation of a lactam which is infact a cyclic amide, the nitrogen atom present in the ring is attached to the beta-carbon relative to the carbonyl group, therefore this structure is called a beta-lactam. The functioning of the enzyme is by the hydrolysis of peptide bond which is unique to the 4-membered beta-lactam ring, this renders the beta-lactam antibiotic ineffective. So far, almost one hundred ninety beta-lactamase types are known, the method of functioning of most of these enzymes is by a serine ester hydrolysis mechanism.

Classification of beta-lactamases was done on the basis of its extent of action against certain antibiotics. Richmond & Sykes proposed the first scheme that was widely accepted, this was done in 1964, this scheme was based on whether an enzyme hydrolysed penicillin more or less rapidly than cephaloridine and whether its activity was inhibited by cloxacillin and/or p-chloromercuribenzoate . Inefficiencies in this scheme led to its revision by Bush in 1989 which was based on relative activity against cephaloridine, extended spectrum cephalosporins, benzylpenicillin, carbenicillin, oxacillin and imipenem, and susceptibility to inhibition by cloxacillin, aztreonam and clavulanate, recognizes four major beta-lactamase classes, among which one is further divided into 8 subgroups. Technological advancements in the field of molecular biology allowed the classification which were based on schemes of sequences. Distortion due to mutations and kinetic activity would not take place in such a classification.The first classification was proposed by Ambler, This revealed four new classes of *β*-lactamases, these included Class A to Class D. Class A and C includes the most clinically important *β*-lactamases, Class C comprises the AmpC *β*-lactamases which are the point of interest in this project (4).

The class C beta-lactamases (AmpC) are clinically significant because Gram-negative-bacteria which confer AmpC are also resistant to penicillin, cephalosporin, cephamycin and monobactam groups. AmpC Beta lactamase mediated by plasmid as well as chromosomal are known. Serratia, Pseudomonas, Acinetobacter, Citrobacter and Enterobacter spp are some examples for chromosomal mediated AmpC beta-lactamase enzymes. Enterobacteriaceae including *E. coli, K. pneumoniae, Salmonella spp, Citrobacter freundii, Morganella morganii, Serratia marcescens, Pseudomonas aeruginosa, Providencia* and *Proteus mirabilis* has shown presence of AmpC beta-lactamase mediated by plasmids (5).

Studying the effects and functioning of beta-lactamase enzymes can provide better insights into the antimicrobial resistance mechanisms adopted by microorganisms. Hence it is essential to know how a compound affects the functioning of such enzymes. 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 l to predict the bioactivity of certain compounds like flavonoids and terpenoids against AmpC beta-lactamase.

METHODOLOGY

A brief description of technologies applied in this project;

PaDEL-descriptor

This is a software which can be used to find fingerprints as well as molecular descriptors.

Computation of these descriptors and fingerprints are done by using the chemistry development kit. Additional descriptors and fingerprints were added, including atom type electrotopological state descriptors, McGowan volume, molecular linear free energy relation descriptors, ring counts, Laggner chemical substructures counts, and binary fingerprints and count of chemical substructures identified by Klekota and Roth.

A molecular descriptor is defined as "the end result of a logic and mathematical technique that converts chemical information recorded within a symbolic representation of a molecule into a useful number or the result of some standardised experiment"[6]. For chemical substances, molecular descriptors are calculated and utilised to construct quantitative structure activity relationship (QSAR) models for predicting biological activity.

PaDEL-Descriptor was written in Java and consists of two parts: an interface component and a library component [7].The workhorse of the system is the library component and it is contained by itself. This indicates that it can work without the interface component and can be linked to another QSAR software inorder to get the computation feature in the descriptor. Wrapper classes are provided for the 43 molecular descriptor algorithms and seven fingerprint algorithms built in CDK by the library component. The library component additionally included four molecular descriptors and three fingerprint methods not included in CDK.

These include atom type electro topological state descriptors, McGowan volume, molecular linear free energy relation descriptors, ring counts, Laggner's count of chemical substructures, and Klekota and Roth's binary fingerprints and count of chemical substructures[8]. These algorithms were created by extending important CDK classes in order to be compatible with existing CDK descriptor algorithms and to be able to be introduced to CDK in the future.

PubChem fingerprints

For chemical structures, the PubChem System creates a binary substructure fingerprint. PubChem uses these fingerprints for similarity and neighbouring searching.

A substructure is a piece of a chemical structure that has been broken apart. An ordered set of binaries (1/0) bits constitutes a fingerprint. Each bit indicates a Boolean determination or test for the presence of anything in a chemical structure, such as an element count, a type of ring system, atom pairing, atom environment (nearest neighbors), and so on(9).

Molecular fingerprints are commonly utilized in cheminformatics applications like as diversity analysis and similarity searches. The molecular representation of each component in a chemical library is frequently required for fingerprint-based analysis of chemical libraries, especially big collections, which can lead to storage space concerns and duplicate computations. In reality, data redundancy is built into the data, resulting in binary digit places in the fingerprint that are devoid of important information. By defining molecules as chemical fingerprints and recording their structural features as a vector, cheminformatics enables the exploitation and interpretation of this variety. These fingerprints may be utilized for quick similarity comparisons, which can be employed for research of structure–activity relationships, virtual screening, and the creation of chemical space maps. Numerous molecular fingerprints were developed, confirmed, and utilized to characterize small molecule medications within the Lipinski limits , and are not well adapted to bigger compounds(10).

ChEMBL DATABASE

This is a large-scale bioactivity database which is developed to provide vital data during drug discovery. This database is open and contains information on binding, functional and ADMET information for a large number of drug-like bioactive compounds. The data present in this database is abstracted manually from literature which is already published, this is done on a regular basis. The literature is then curated and standardized so that the quality and utility remains top notch in order to solve a plethora of drug discovery related queries. At present, the database consists of 5.4 million measurements of bioactivity for more than 1 million compounds and five thousand two hundred protein targets. A web based interface is used for access to the database, https://www.ebi.ac.uk/chembldb .

Data extraction

The data used in this project was taken from the ChEMBL database. Specifically, the bioactivity of 64280 compounds against beta-lactamase AmpC. This data contains chembl id of the compunds along with the chemical structure in form of canonical_smiles, standard_relation, standard_value, standard_units, standard type, pchembl_value, target_pref_name and, bao_label.

Environment used

The project was mostly done on google colab and jupyter notebook and python programming language was used.

Important Packages and libraries

Several libraries are used here

Pandas -

Pandas is a python library that provides useful data structure and operations for the analysis of data. It is one of the most widely used libraries for data cleaning and analysis operations. Pandas provide the easiest task and are pretty fast in their operations.

One of the main reasons for the popularity of pandas is the data frame data structure. Pandas can load data from CSV, Excel and JSON files.

It is best suited for working with tabular or different kinds of data. As database-like join/merge procedures are available to connect several tables of data, multiple tables can be merged both column and row wise .

Pandas includes a lot of capabilities for working with dates, times, and time-indexed data, and it has a lot of support for time series.

NumPy -

Numpy means numerical python and it is the short form of numerical python. If we want to perform any array-based or matrix-based calculation and data analysis then, it is better to use the NumPy library. We can do complex calculations very easily with the help of NumPy. So we can perform any element-wise operations without using any loop with the help of Numpy.

It is best suited for working with the same kind of data.It's a table containing the same datatype elements or numbers, indexed by a tuple of positive integers. Dimensions are referred to as axes in NumPy, and the number of axes is referred to as rank.

NumPy's strong n-dimensional array increases data processing. NumPy includes capabilities for connecting with other programming languages such as C, C++, and others, and may easily interact with other Python packages.

For constructing, manipulating, and modifying NumPy arrays, the NumPy library includes a vast number of built-in functions. The array data structure in Python is also available, although it is not as adaptable, efficient, or helpful as the NumPy array. The formal name for the NumPy array is ndarray, however it is most generally referred to as array.

Scikit Learn -

Scikit-learn offers a standard Python interface for a variety of supervised and unsupervised learning techniques.It is provided under several Linux distributions and is published under a liberal simplified BSD licence, enabling academic and commercial use.

The library's goal is to achieve the degree of consistency and support necessary for usage in production systems. This involves a concentrated effort on issues like usability, code quality, collaboration, documentation, and performance.

The library focuses on data modelling. It isn't focused on data loading, manipulation, or summarization.

Following things can be done with this: -

Classification: Spam detection, image recognition.

Clustering: Drug response, Stock price.

Regression: Customer segmentation, Grouping experiment outcomes.

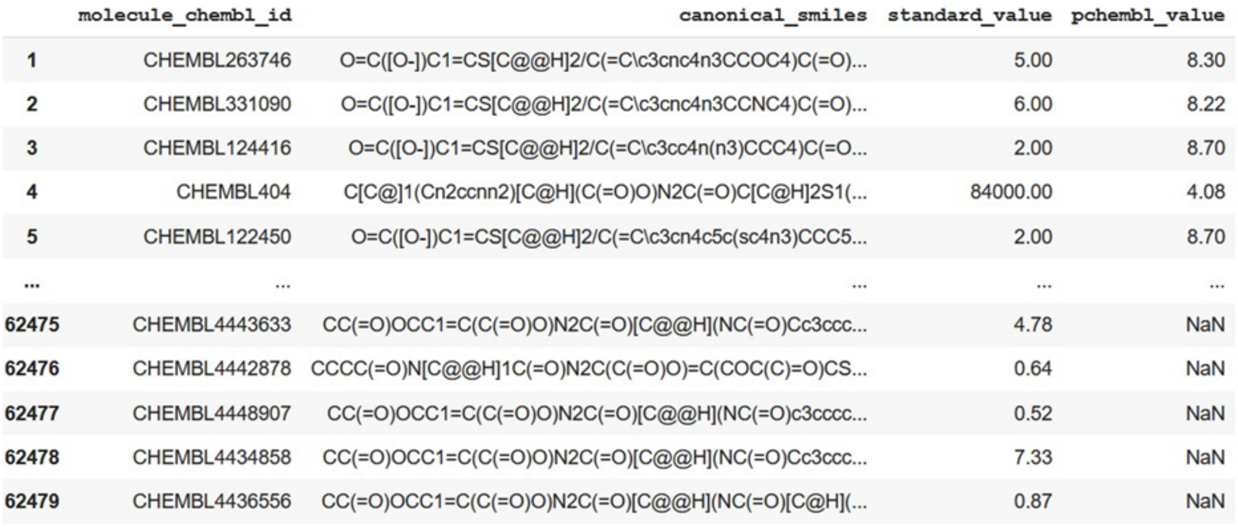
Dimensionality reduction: Visualization, Increased efficiency.

Model selection: Improved precision due to parameter adjustments

Pre-processing: Preparing input data as a text for machine learning algorithms to process.

Data cleaning

Data cleaning is the process of fixing or removing erroneous, corrupted, badly formatted, duplicate, or incomplete data from a dataset. When combining various data sources, there are numerous possibilities for data to be duplicated or mislabeled. The conclusions and techniques are untrustworthy if the data is incorrect, even if they appear to be right. There is no one-size-fits-all way for prescribing the precise terms in the data cleaning process because the methodologies vary from dataset to dataset. Creating a model for your data cleaning technique, on the other hand, is crucial so that you know you're doing it right every time.

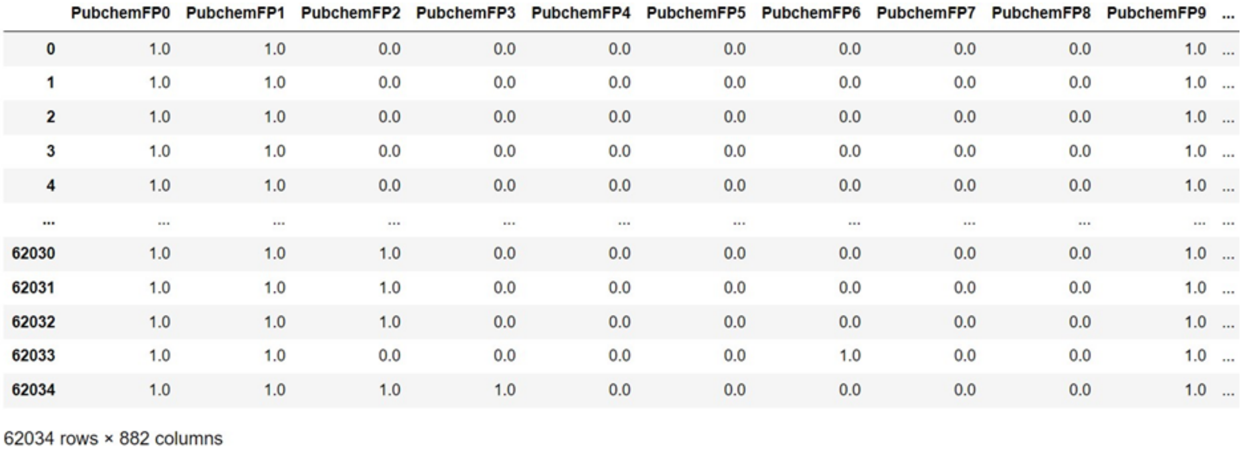


	molecule_chembl_id	canonical_smiles	standard_value	pchembl_value
1	CHEMBL263746	O=C([O-])C1=CS[C@@H]2/C(=C\c3cnc4n3CCOC4)C(=O)...	5.00	8.30
2	CHEMBL331090	O=C([O-])C1=CS[C@@H]2/C(=C\c3cnc4n3CCNC4)C(=O)...	6.00	8.22
3	CHEMBL124416	O=C([O-])C1=CS[C@@H]2/C(=C\c3cc4n(n3)CCC4)C(=O...	2.00	8.70
4	CHEMBL404	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

Important columns were selected with relevant data for further data processing. The column with standard_value was also removed as we needed only pchembl_values column. Our data had many NaN and infinity values which were needed to be removed, after cleaning the data out of 64280 compounds we were left with 62034 which had all the necessary values for our further processes.

Generating pubchem fingerprints

Then we calculated pubchem fingerprints of all those 62034 compounds. The fingerprints are calculated with the help of molecules structural components that is the smiles notation. Now our data had pubchem fingerprints and pchembl_value.



	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	...

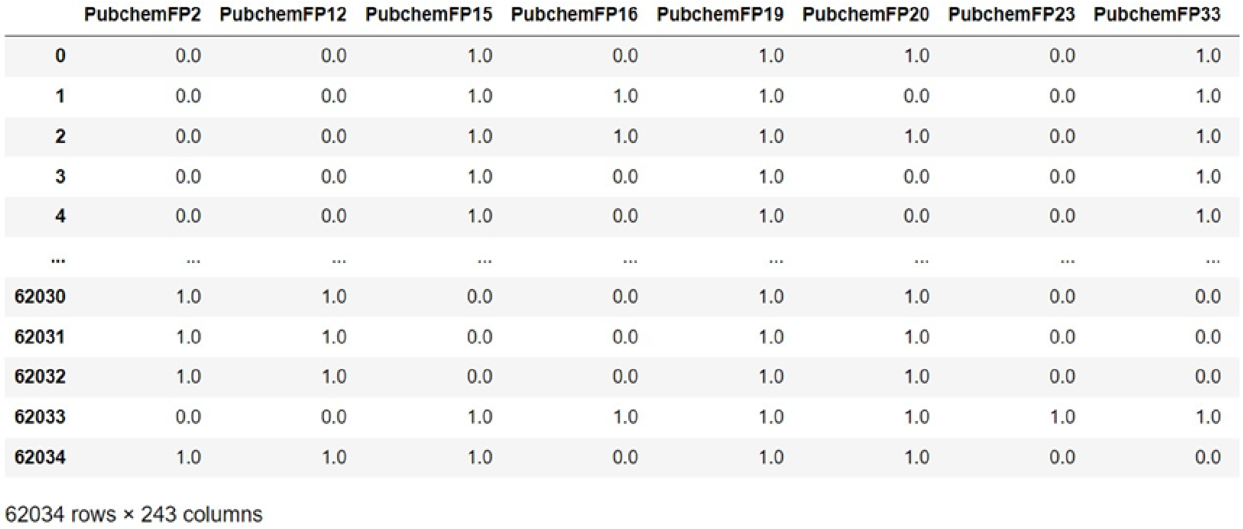
62034 rows × 882 columns

Preparing the X and Y Data Matrices

The data needs to be prepared for model building. Therefore, it is necessary to remove the redundant values and so the pubchem fingerprints and pchembl_value were individually preprocessed.

X data matrix

We loaded the pubchem fingerprints in the X variable and removed the low variance features, after that out of 881 columns of pubchem fingerprints we were left with only 243 columns of fingerprints.

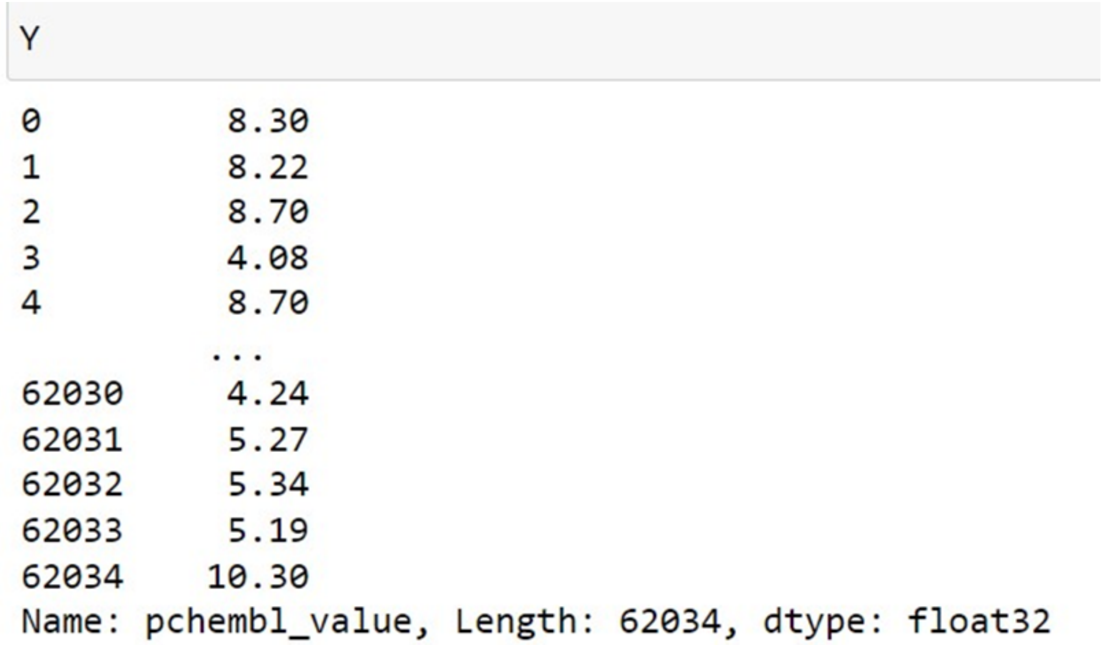


	PubchemFP2	PubchemFP12	PubchemFP15	PubchemFP16	PubchemFP19	PubchemFP20	PubchemFP23	PubchemFP33
0	0.0	0.0	1.0	0.0	1.0	1.0	0.0	1.0
1	0.0	0.0	1.0	1.0	1.0	0.0	0.0	1.0
2	0.0	0.0	1.0	1.0	1.0	1.0	0.0	1.0
3	0.0	0.0	1.0	0.0	1.0	0.0	0.0	1.0
4	0.0	0.0	1.0	0.0	1.0	0.0	0.0	1.0
...	...	...	...	...	...	...	...	...
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

Y variable

We loaded the pchembl_vlaue in the Y variable. The data type of Y variable was in float64 but our model accepts only float 32 therefore, we converted it into float32.

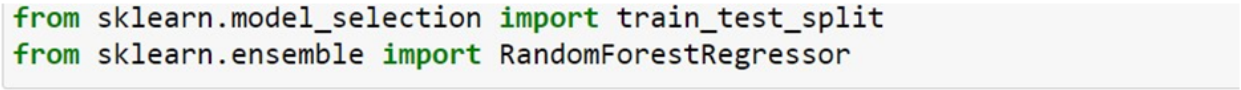


```
Y
```

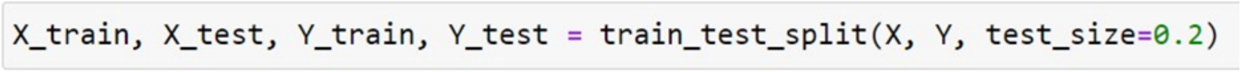
```
0         8.30
1         8.22
2         8.70
3         4.08
4         8.70
          ...
62030     4.24
62031     5.27
62032     5.34
62033     5.19
62034    10.30
Name: pchembl_value, Length: 62034, dtype: float32
```

Data split (80/20 ratio)

The data was splitted into 80/20. 80% goes to the training of the model and the rest 20% for testing.



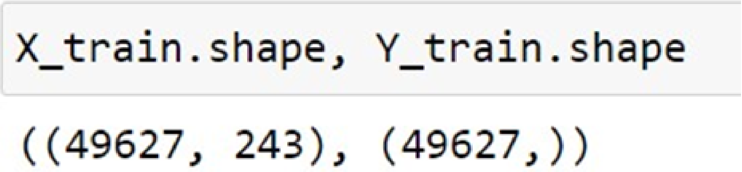
```
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestRegressor
```



```
X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size=0.2)
```

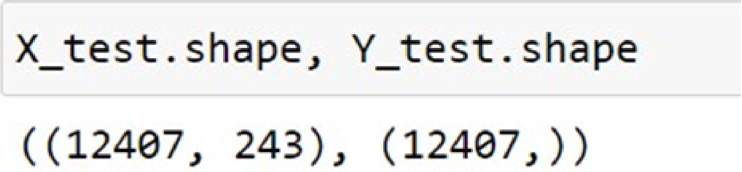
Let's examine the data dimension

The dimension of both the training set and testing set is examined to see if they are correctly splitted.



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

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

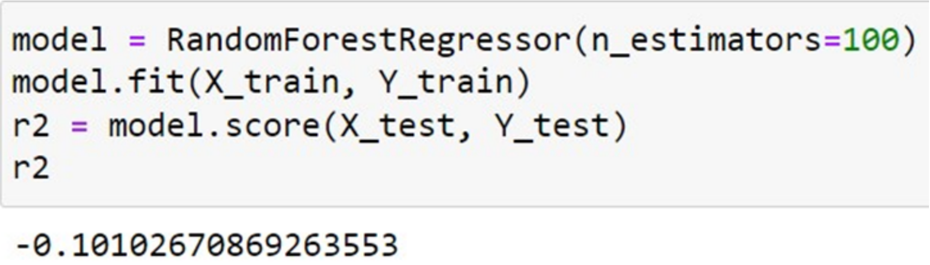


```
X_test.shape, Y_test.shape
```

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

Building a Regression Model using Random Forest

We built a Random Forest Regression model for our project.

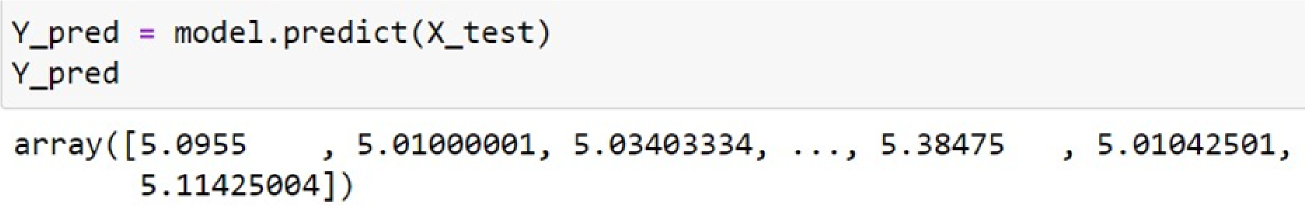


```
model = RandomForestRegressor(n_estimators=100)
model.fit(X_train, Y_train)
r2 = model.score(X_test, Y_test)
r2
```

```
-0.10102670869263553
```

Model prediction

We predicted the pchembl_values using the model.predict function and we got predictions for the compounds we stored in the X_test variable. The result was in the form of arrays.



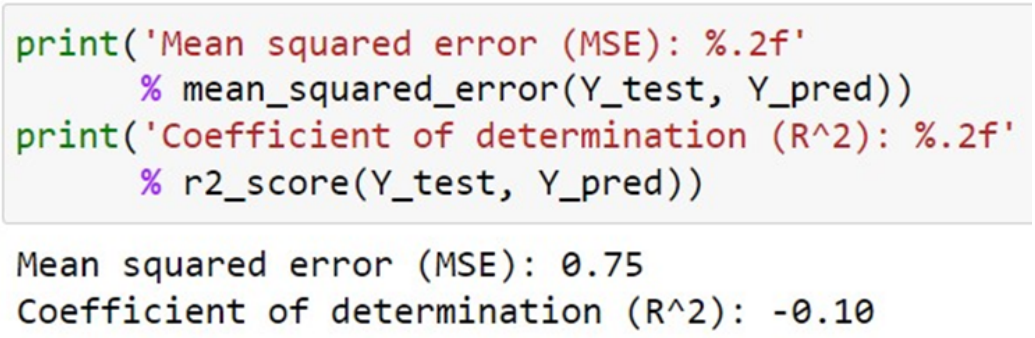
```
Y_pred = model.predict(X_test)
Y_pred
```

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

Model performance

Mean Square Error: the average squared difference between the estimated values and true value, is measured by the Mean Squared Error of an estimator. It's a risk function that corresponds to the squared error loss's predicted value.

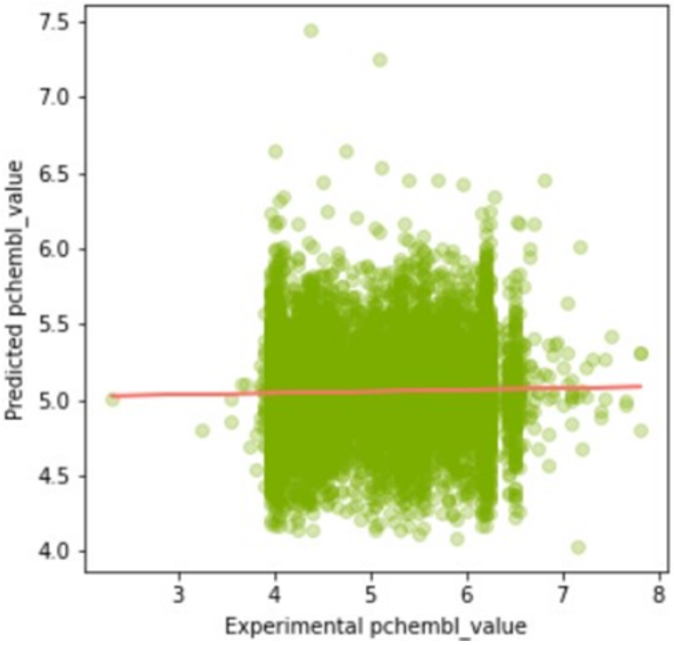
Coefficient of determination: The proportion of the variation in the dependent variable predicted by the independent variable is known as the R squared method. It shows how much variety there is in a given data collection. The R2 value indicates if the model is a good match for the data set. For each given percent of variation in the context of analysis.



```
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
Coefficient of determination (R^2): -0.10
```

Scatter Plot of Experimental vs Predicted pChembl Values



RESULT

Testing with new data: We Tested few classes of flavonoids and terpenoids with our machine learning model and results are mentioned below:

Flavones: We tested 4 Flavones molecules out of them 0 had the pchembl value greater than 5.

molecule_name	Predicted pchembl_value

Apigenin	4.8575
Tangeretin	4.655
Baicalein	4.669667
Rhoifolin	4.84925

Table 1 Flavones

Flavanones: - We tested 13 Flavanones molecules out of them 5 showed the pchembl value greater than 5.

molecule_name	Predicted pchembl_value
Butin	4.98210832
Eriodictyol	5.108991691
Hesperetin	4.849249971
Hesperidin	4.950999987
Homoeriodictyol	5.095708322
Isosakuranetin	4.699016674
Naringenin	4.502500031
Naringin	5.100025016
Pinocembrin	4.699016674
Poncirin	4.816466735
Sakuranetin	5.100025016
Sakuranin	5.108383364
Pinostrobin	4.936469046

Table 2 Flavanones

Flavonols: We tested 27 flavonols molecules out of them 19 showed the pchembl value greater than 5.

molecule_name	Predicted pchembl_value

Azaleatin	5.161483351
Fisetin	4.96933335
Galangin	4.950133357
Gossypetin	5.058433348
Kaempferide	4.887450027
Quercetin	5.058433348
Myricetin	4.950133357
Rutin	4.950133357
Morin	5.114225028
Kaempferol	5.139400009
Isorhamnetin	5.02122503
Natsudaidain	5.114225028
Rhamnazin	5.294499977
Rhamnetin	5.04120002
Astragalin	5.128999994
Azalein	5.058433348
Hyperoside	4.918499997
Isoquercitin	4.918499997
Kaempferitrin	5.003491699
Myricitrin	5.128999994
Quercitrin	5.198741679
Robinin	4.898249998
Spiraeoside	5.198741679
Xanthorhamnin	5.19177502
Amurensin	5.28986668

Icariin	5.28986668
Troxerutin	5.182025011

Table 3 Flavonols

Flavononols: - We tested 8 Flavononols molecules out of them 5 showed the pchembl value greater than 5.

molecule_name	Predicted pchembl_value
Dihydrogossypetin	4.96933335
Dihydromorin	5.100025016
Garbanzol	4.96933335
Ampelopsin	5.095780013
Aromadendrin	5.02273001
Fustin	4.895730007
Taxifolin	5.02122503
Pinobanksin	5.153663349

Table 4 Flavononols

Flavonoids: - We tested 22 Flavonoids molecules out of them 10 showed the pchembl value greater than 5

molecule_name	Predicted pchembl_value
Quercetin	4.97325
Rutin	5.058433
Macluraxanthone	4.918583
Genistein	4.918583
Scopoletin	4.89573
Daidzein	5.072358
Taxifolin	5.100025
Naringenin	5.161483
Abyssinones	4.655
Eriodictyol	5.2945
Fisetin	4.809155
Theaflavin	4.967333
Peonidin	4.967333
Diosmetin	5.233458
Tricin	4.97825
Biochanin	5.125383
Hesperidin	4.950133
Epicatechin	5.058433
Myricetin	4.839392
Kaempferol	4.669667
Luteolin	5.095708
Apigenin	5.146

Table 5 Flavonoids

Isoflavons: - We tested 12 isoflavons molecules out of them 1 showed the pchembl value greater than 5.

molecule_name	Predicted pchembl_value
Daidzin	4.918583
Genistein	4.918583
Glycitin	5.009475
Ononin	4.96475
Sissotrin	4.853116
Acetyldaidzin	4.901424
Acetylgenistin	4.901424
Acetylglycitin	4.901424
Malonyldaidzin	4.826661
Malonylgenistin	4.853116
Malonylglycitin	4.96475
Malonylononin	4.826661

Table 6 Isoflavons

Anthocyanin: - We tested 9 Anthocyanin molecules out of them 9 showed the pchembl value greater than 5.

molecule_name	Predicted pchembl_value
Aurantinidin	5.294499977
Cyanidin	5.233458342
Delphinidin	5.424450026
Europinidin	5.294499977
Pelargonidin	5.233458342
Malvidin	5.28986668
Peonidin	5.233458342
Petunidin	5.233458342
Rosinidin	5.233458342

Table 7 Anthocyanin

Terpenoids: - We tested 10 Terpenoids molecules out of them 9 showed the pchembl value greater than 5.

molecule_name	Predicted pchembl_value
Geraniol	5.022950003
Phytol	5.287075006
Farnesol	5.022950003
Geranylgeraniol	4.745499988
PhytanicAcid	5.114350026
Auraptene	5.341850019
Bixin	5.04216666
AbieticAcid	5.336199937
DehydroabieticAcid	5.190946686
Norbixin	5.247449962

Table 8 Terpenoids

Our model has mean square error (MSE) : **0.75**

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.**

FUTURE PERSPECTIVE AND CONCLUSION

In this project we have worked with single fingerprints, looking into the future, this project has the scope of working with multiple fingerprints. Once we get the desired results after multiple fingerprints are used, there is the possibility of applying these results into renowned softwares like Schrodinger. Insights that will be available after working with such a software will enable us 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 such as the ones used in this project helps one to explore a wider biochemical space and predict the bioactivity and molecular behaviour of a plethora of compounds with high degree of accuracy.

In this modern age of genomics more researchers are trying to understand the phenomena of antimicrobial resistance as the evolution and evolutionary drivers of AMR in populations of bacteria are gaining more attraction. Therefore it is necessary 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 alogorithm an immense amount of data which 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.

<u>REFERENCES</u>

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. B 289:321–331.

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. Carhart, 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.