iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 9 **Issue:** VII **Month of publication:** July 2021

DOI: https://doi.org/10.22214/ijraset.2021.37221

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue VII July 2021- Available at www.ijraset.com

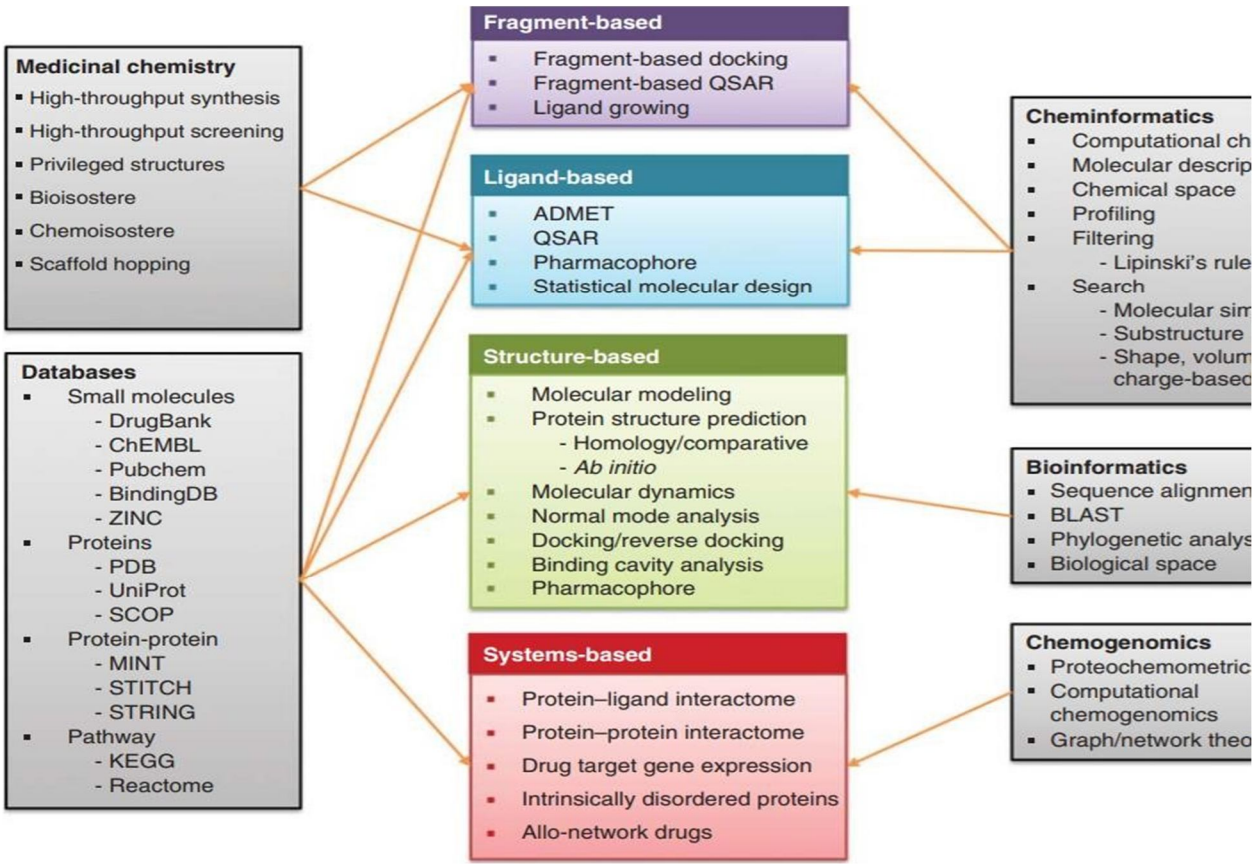
A Data Science Approach to Bioinformatics

Palepu Narasimha Rakesh[1], Palepu Mounika[2], Maka Keli Ratna Sai Sailaja[3], P. Alhena Minhaz[4]
[1, 2, 3, 4]*Department of Biotechnology, K L University, Andhra Pradesh, India*

Abstract: ***Computer aided drug design (CADD) which uses the computational advance towards to develop, discover and scrutinize and examine drugs and alike biologically agile molecules. CADD is a specialized stream which uses the computational techniques to mimic drug-receptor interactions. CADD procedures are so much dependent on the tools of bioinformatics, databases & applications. There are so many advantages of computer aided drug discovery; it saves lot of time which is one of the main advantages followed by low cost and more accuracy. CADD required less manpower to work. There are different types of CADD such as ligand and structure based design. Objectives of the Computer aided drug design are to boost up the screening process, to test the rational of drug design, to efficiently screen and to remove hopeless ones as early as possible. In Drug designing the selected molecule should be organic small molecule, complementary in shape to the target and oppositely charged to the biomolecular target. The molecule will interacts and binds with the target which activates or inhibits the function of a biomolecule such as a protein or lipid. The main basic goal in the drug design is to forecast whether a given molecule will bind to target and if thus how strongly. Molecular mechanics techniques also used to provide the semi quantitative prediction of the binding affinity. These techniques use machine learning, linear regression, neural nets or other statistical methods to derive predictive binding affinity equations. Preferably, the computational technique will be able to forecast the affinity prior to a compound is synthesized, saving huge time and cost. Computational techniques have quickened the discovery by decreasing the number of iterations required and have often produced the novel structures.***

Keywords: Computer aided drug design, Target molecule, Binding affinity and Receptor

I. INTRODUCTION



A. An Overview Of Computational Drug Discovery

Drug discovery has various domains in Biology. We look for an approach where there is least human involvement, cost-efficient, and less time-consuming. This approach is considered as Computational Drug discovery.

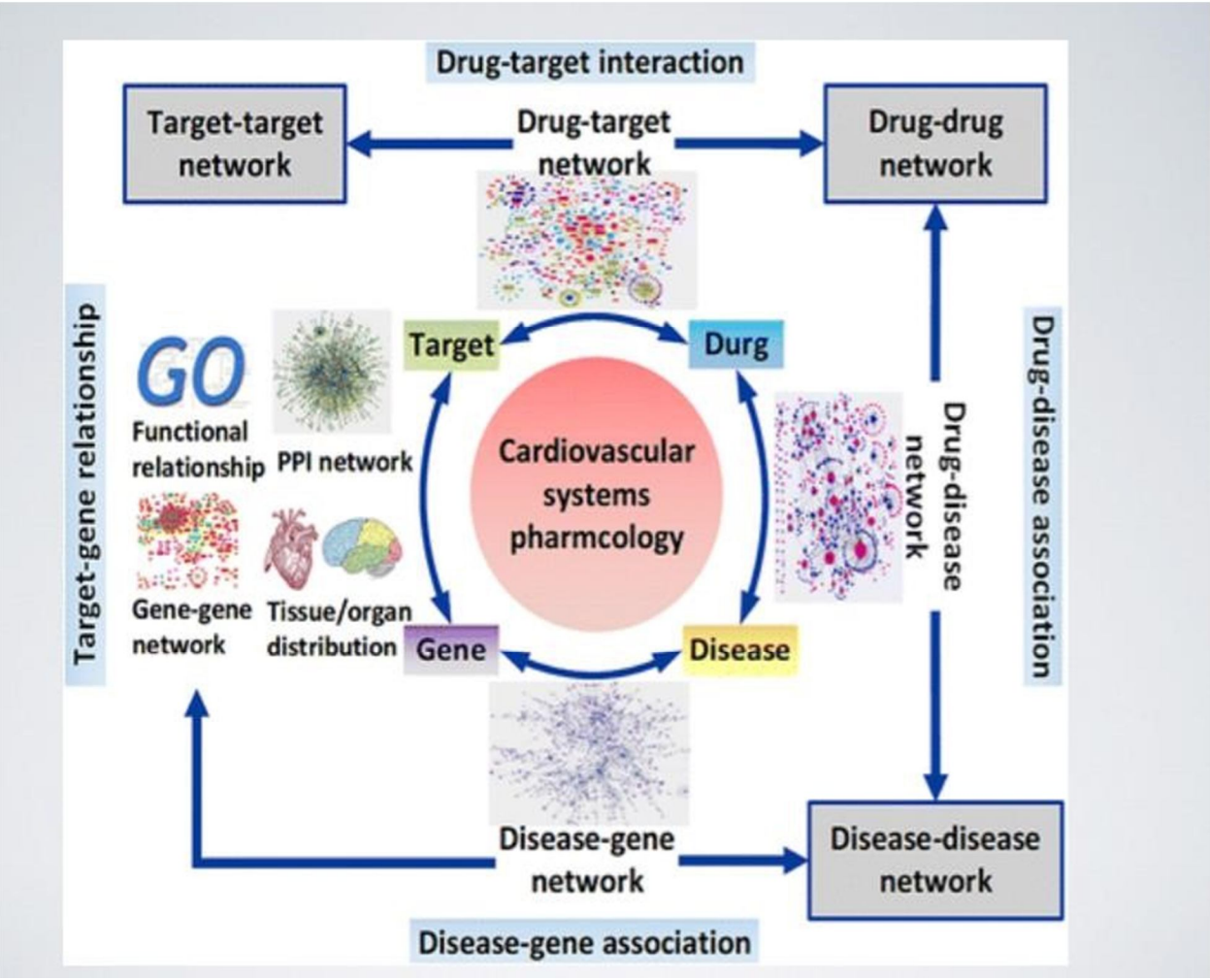
Computational Aided Drug Discovery (CADD) tools are supposed to be a shortcut for assisting the above features. In therapeutic development, CADD has become an efficient and indispensable tool in recent times. The amount of sequence data that the Human Genome Project has made, can be utilized for many Drug Discovery projects. Scientists are finding out novel computational methods and algorithms to come over huge Biological problems that have become difficult to solve. In advanced pharmaceutical research, the usage of computational methods and in silico tools is greater than ever before. Computer-aided drug design (CADD) helps the chemists or physicists in making the new drug design in an advanced way. Computer-aided drug design lies in the hand of computational scientists who are able to manipulate the molecule on the screen. The CADD approach we are using is a ligand-based.

The project is destined for Computational Drug Discovery and especially into a domain called QSAR (QUANTITATIVE STRUCTURE ACTIVITY RELATIONSHIP) techniques.

Machine Learning modules and Python libraries are efficiently utilized.

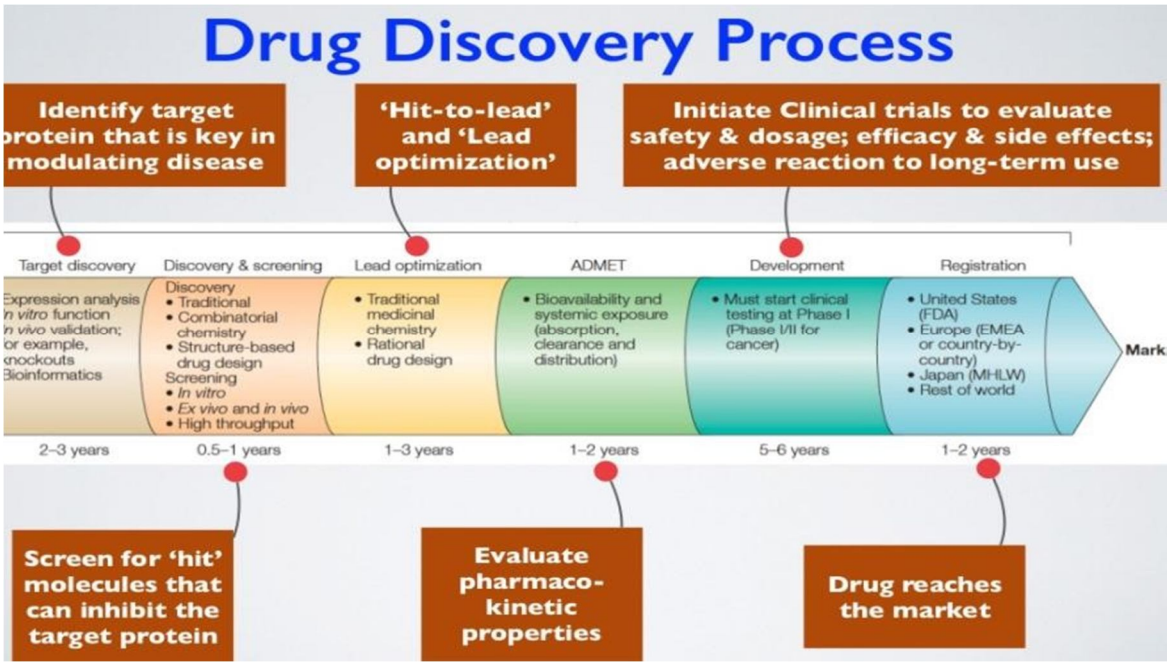
When bound to each other, to form a stable complex, a favored orientation is forecasted from one molecule to another.

Huge biological data used to make sense back in time. But as Machine learning and Convolutional Neural Networking techniques have emerged, scientists are able to understand the patterns and make sense of such big biological data, understand and compute good algorithms for branching the fields of biology deeper into the roots.



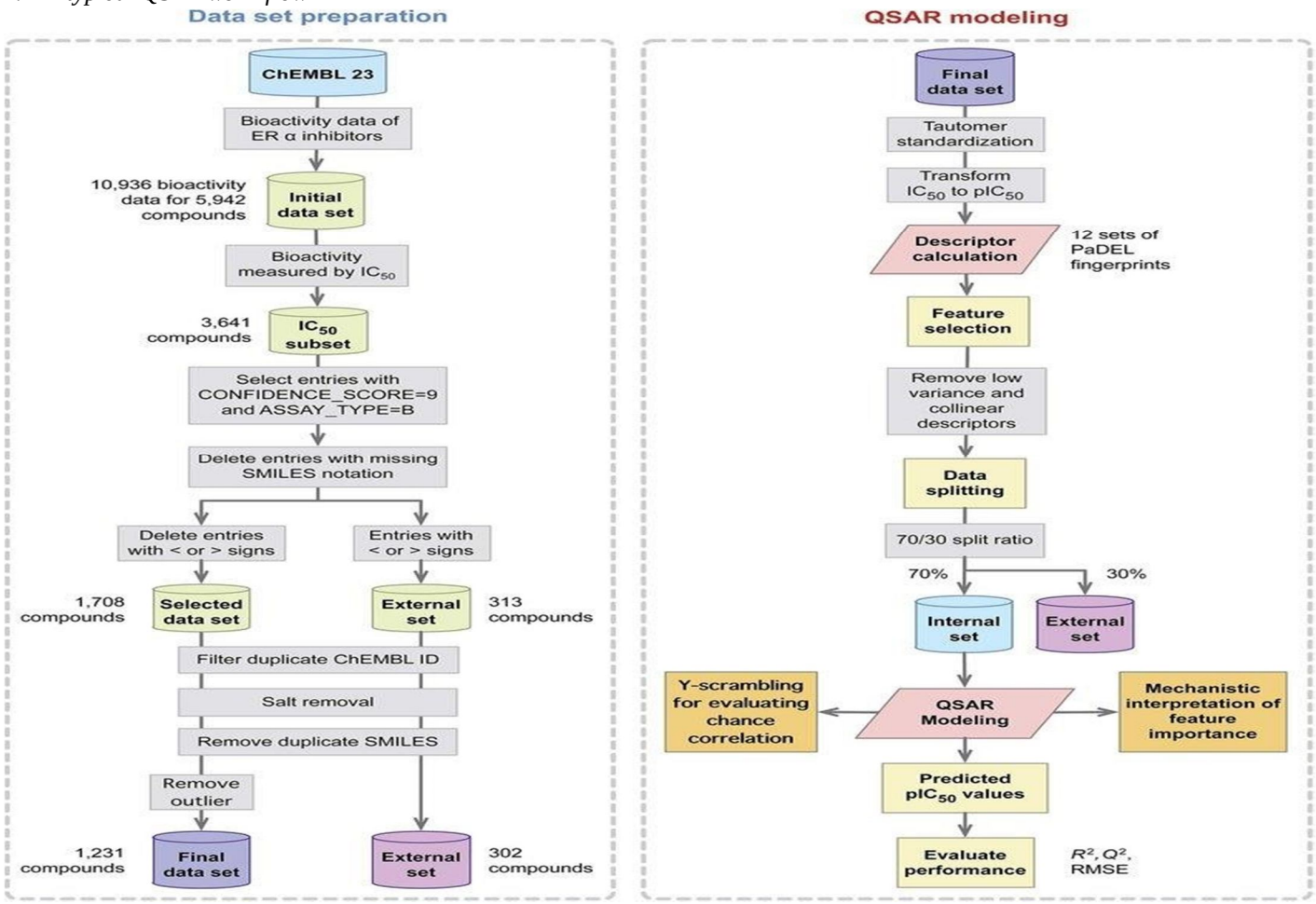
Target-Gene Relationship

The accustomed Drug Discovery Process in a nutshell _



The novel drug computer aided drug discovery process

B. A typical QSAR work-flow



II. METHODOLOGY

In the chEMBL database there will be curated bioactivity data of more than 1.5 million compounds and there will be 13000 targets. The chEMBL database contains around 78000 documents. As the first step, we must install the chEMBL web service package so that we can retrieve the bioactivity data from this database i.e chEMBL. After the installation of the chEMBL database successfully we will import the python libraries. There is a python library named pandas, it is one of the most used python libraries.

III. EXPLORATORY DATA ANALYSIS

Exploratory data analysis is done using Lipinski descriptors and libraries are imported and there will be a graph of the frequency of 2 bioactivity classes active and inactive will be present. Then there will be a plot between MW versus LogP. Then there will be the presence of a box plot of PIc 50 between inactive and active. Now statistical analysis using pandas will be done and it is known as the Mann-Whitney U test.

There will be continuous plots of inactive and active classes of MW, logP, Number H-Donors, and Number H-Acceptors.

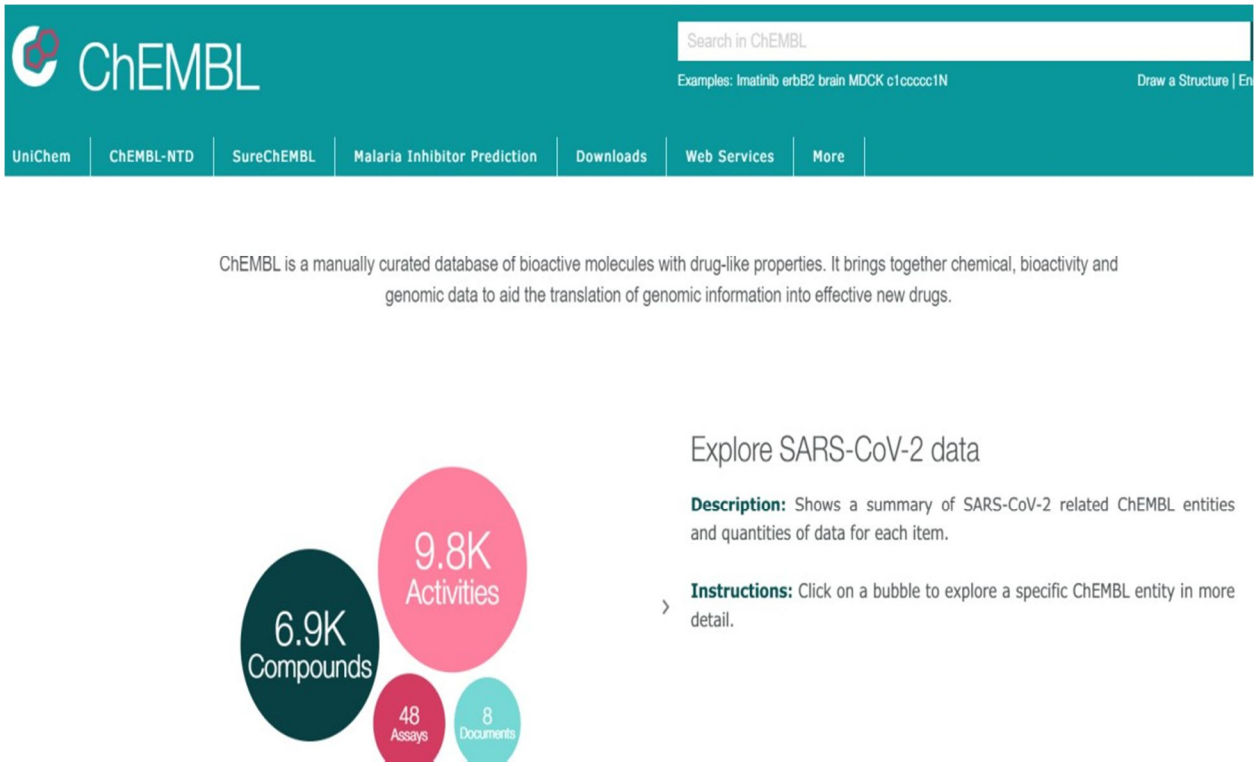
A good molecule would have a very high pIC50 value (normally higher than 6 is considered to be good). It should be noted that "good" in this context refers to strong inhibition of the target protein of interest. In a practical setting, a good molecule in addition to strong inhibition of the target protein of interest it should also possess favorable "pharmacokinetic profiles" (the various molecular properties pertaining to the absorption, distribution, metabolism, excretion and toxicity of a drug). Thus, ideally, an ideal drug would have strong inhibition and good pharmacokinetic profiles. In practice, this is essentially an optimization problem and it is rather challenging to find such a molecule. Thus, in the real world setting, we have drugs that are potent but also lead to side effects in patients

A. Part 1

1) Installation of Chembl web Resource Clients - ChEMBL Database

ChEMBL (Curated Chemical Database of bioactive Molecules with drug-like properties) which is a PubChem data resource, is installed so as to retrieve bioactivity data. Bioactivity is reported in Kd (Dissociation constant), KI (Inhibition constant), IC50 (Half-maximal inhibitory constant), &EC50 (Half maximal effective concentration).

During Drug discovery, data can be filtered and analysed for lead identification in developing compounds, screening libraries.



ChEMBL Database

2) Importing Python Libraries

Panel Data - The word from which the name Pandas is derived which is an Econometrics from any dimensional data. Pandas tools provide useful and powerful data structures along with high performance data manipulation and analysis.

Prior to Pandas, Python was majorly used for data preparation and munging. Contribution to data analysis was very less. Pandas solved this problem.

Regardless of origin of data, we can basically with the use of Pandas, can accomplish five typical steps which include - loading, preparing, manipulating, modelling, and analyzing.

3) Searching And Selecting And Retrieving A Target For The Respective Organism

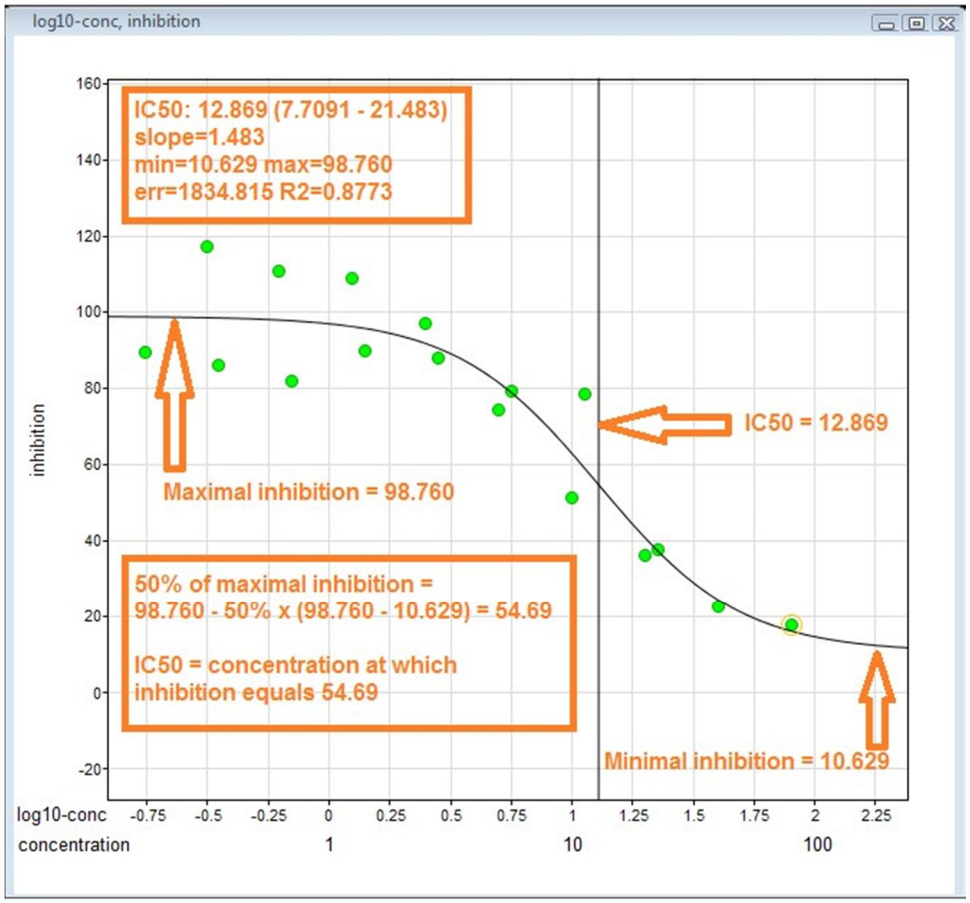
From the pool of Targets, we search for respective targets from ChEMBL Database which we already imported. So the targets here refer to the target proteins or target organism that the drug will act on. So biologically, these compounds will come into contact with the protein or the organism and induce a modulatory activity towards it

It could either activate the protein or the organism or inhibit the same.

Project using single Protein for our further investigation. target only ic50 values.

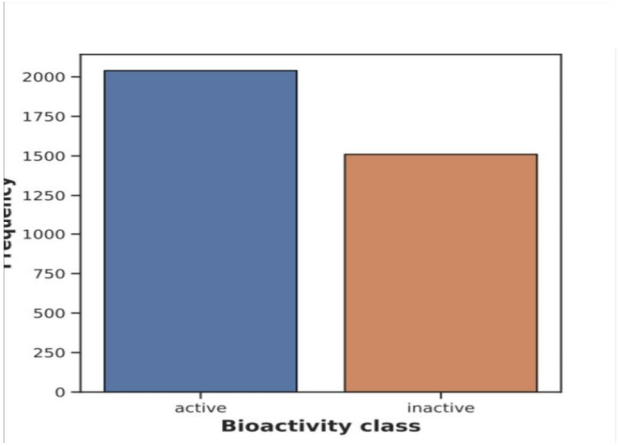
a) IC50: Half maximal inhibitory concentration. The effectiveness measure of a substance for which it inhibits a specific biochemical or biological function is called (IC50) Half-maximal concentration.

b) EC50: Half maximal effective concentration. It Measures the activation. pIC50 = (negative logarithm)-log10 (IC50).

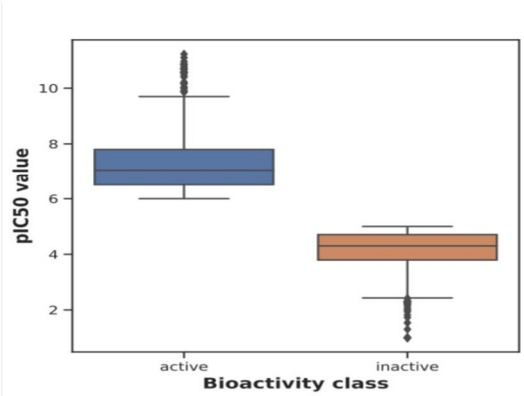


It's the same type for this but it might be IC50, EC50 or percentage of activity. The more the number the higher the potency becomes and vice versa. The number here represents the concentration of the drug. the lower the concentration of the drugs the better it is.

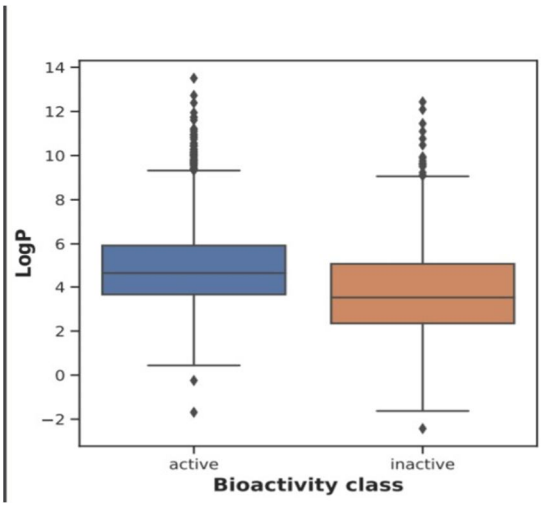
The below are Exploratory Data Analysis (Chemical Space Analysis) via Lipinski descriptors



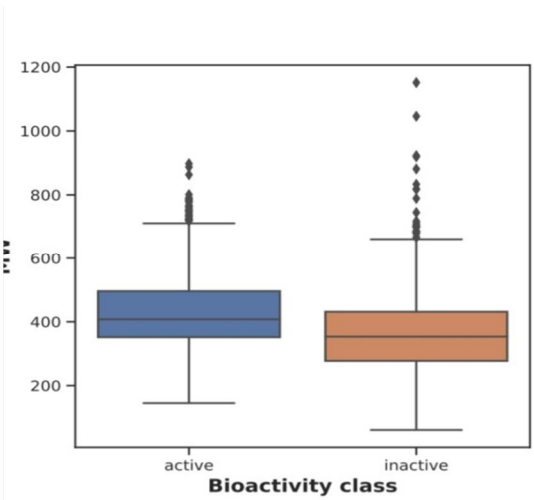
Frequency plot of the 2 bioactivity classes



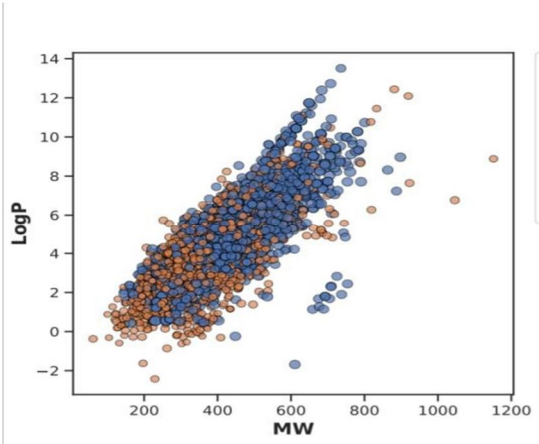
pIC50 value



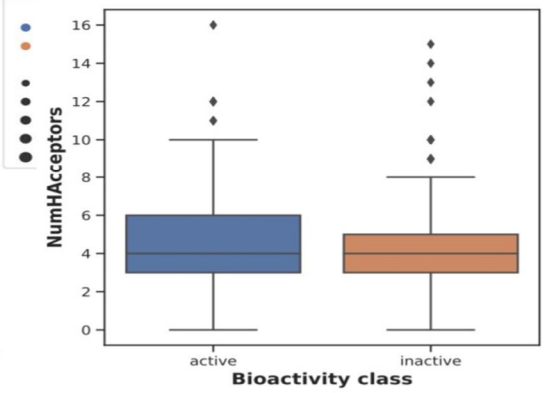
MW



LogP



Scatter plot of MW versus LogP



NumHAcceptors

B. PART 2

The downloaded dataset from Chembl shows the biological activity. The dataset comprises the molecule names and the corresponding smiles notation which is the information of chemical structure.

Now we use the same in this part to compute the molecular descriptors. The data also contains the IC50 values where we performed binding to the bioactivity class.

In this, we select only active and inactive so that we can easily compare between them.

This takes in the smiles notation which contains the chemical information. Meaning, the atomic details of the molecules and the information is used as the input to calculate molecular descriptors.

In the table: Molecular weight, LogP - solubility, H-donors, H-acceptors.

Now combine the df data frame and Lipinski data frame together. Because we want the standard value and bioactivity of class. The log values and Lipinski columns are added. Values greater than 100,000,000 will be fixed at 100,000,000 otherwise the negative logarithmic value will become negative.

Now we need to convert IC50 to pIC50 values. The reason for the same is essentially the negative logarithmic transformation from the IC50 value is that the original IC50 value has an uneven distribution of df. To make it even, we convert.

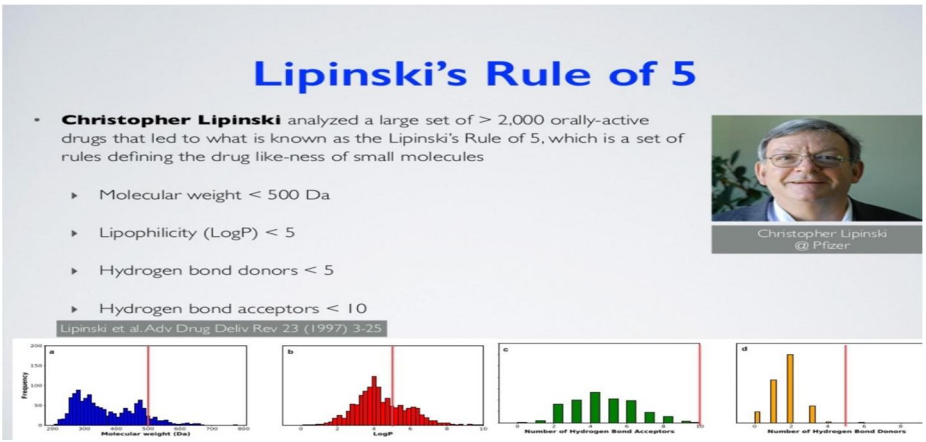
Now we perform exploratory data analysis using the Lipinski rule of 5 - We're gonna call this chemical space analysis because it is what it essentially does is it allows us to look at the chemical space. and the chemical space is kind of like the chemical universe. which chemical compound could be thought of as like stars active molecules could be compared to the constellation.

So, the atom molecule has the largest size compared to the smallest molecule. We are going to apply a similar concept here. To develop a new drug, scientists always talk about the Lipinski rule of 5.

It states the following:

Molecular weight < 500 Dalton

- Octanol-water partition coefficient (LogP) < 5
- Hydrogen bond donors < 5
- Hydrogen bond acceptors < 10



C. Part 3

In Part 3, we will be calculating molecular descriptors that are essentially quantitative descriptions of the compounds in the dataset. Finally, we will be preparing this into a dataset.

As a first step, we will download padel descriptor software for calculating the molecular descriptors. Molecular descriptors are nothing but the mathematical representation of molecules having properties that are produced by algorithms. The molecular descriptor is the ultimate and key result of mathematical and logic technique which transforms chemical data encoded with the symbolic depiction of a molecule into a needful number. Download the curated ChEMBL bioactivity data that has been pre-processed from Parts 1 and 2 of this project. Here we will be using the bioactivity_data_3class_pIC50.csv file that essentially contains the pIC50 values that we will be using for building a regression model.

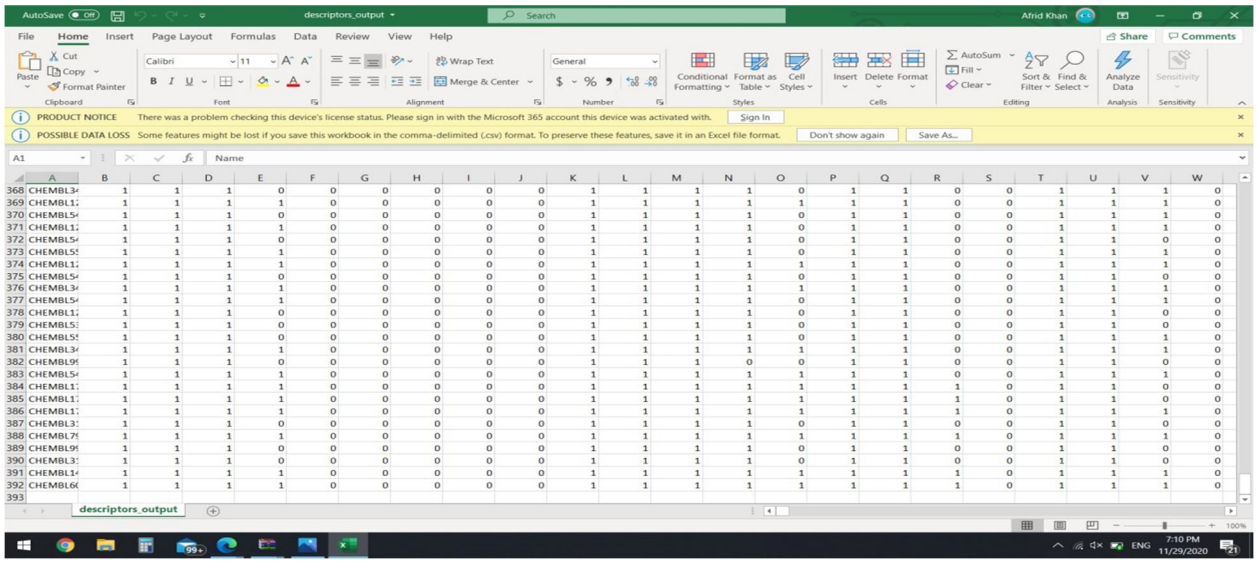
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue VII July 2021- Available at www.ijraset.com

Calculation of padel descriptor values using pandas and preparing the X and Y data matrices. Therefore we will combine these data matrices into a single dataset and in that there will be a column of PubChem Fp values. Saving the CSV file as dataset 3, this will be used further for model building.

Each descriptor type is used for different purposes although both are used for describing the molecular properties of a molecule. Firstly, the Lipinski descriptors are describing the general properties of a molecule at the macro level since it accounts for the molecule's entire molecular weight, the total count of hydrogen bond donors/acceptors and the molecule's solubility (LogP). Secondly, the PubChem descriptors from PaDEL are accounting for the local properties of a molecule at the micro level since it breaks the molecule down into its atomic constituents since each of the 881 descriptors are binary fingerprints of whether the molecule have or not have a particular functional group (a small collection of bonded atoms, which if compared to Lego building blocks, 1 Lego would correspond to 1 atom more would correspond to a functional group (the order and connectivity of these Lego blocks matter and influences the molecular property, thus allow researchers to design new drugs. This is the primary expertise of medicinal chemists).

Load the bioactivity class that was already installed, import pandas, read the csv file Import pandas, read the csv file. The smiles notation represents the chemical information that tells us about the chemical structure. We have 4695 molecules. padel descriptors and bash clears the salt and other impurities internally.

Type of fingerprint - pubchem read the output of fingerprint descriptors. prepare the x and y data matrices. x - pubchem fingerprint, y - IC50 to pIC50 The prepared dataset is then again used for model building.



D. Part 4

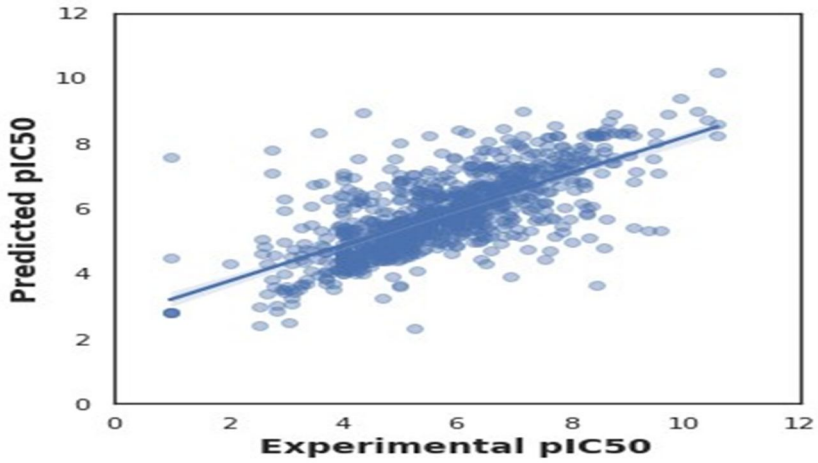
Randomforest algorithm: Random forest is a supervised learning algorithm which is used for both classification as well as regression. But however, it is mainly used for classification problems. As we know that a forest is made up of trees and more trees means more robust forest. Similarly, random forest algorithm creates decision trees on data samples and then gets the prediction from each of them and finally selects the best solution by means of voting. It is an ensemble method which is better than a single decision tree because it reduces the over-fitting by averaging the result.

Working of Random Forest Algorithm

We can understand the working of Random Forest algorithm with the help of following steps −

1) *Step 1* − First, start with the selection of random samples from a given dataset.
2) *Step 2* − Next, this algorithm will construct a decision tree for every sample. Then it will get the prediction result from every decision tree.
3) *Step 3* − In this step, voting will be performed for every predicted result.
4) *Step 4* − At last, select the most voted prediction result as the final prediction result. we have to figure out the most essential way of the (lego building blocks) pubchem fingerprints in a way that the molecule provides the most potency towards the target drug molecule, being safe and not toxic.

The ml algorithm will learn from unique properties in terms of molecular properties of the compound and then create a model that will be able to distinguish between compounds that are active and those are inactive.
functional groups that are essential for designing a good and potent drug split the data, build a regression model using random forest, make the prediction, scatter plot of experimental vs predicted pIC50 value.
In Part 4, we will be building a regression model of acetylcholinesterase inhibitors using the random forest algorithm. Again using Pandas importing libraries and loading the dataset and the Acetylcholinesterase data set contains 881 input features and 1 output variable (pIC50 values). The input contains PubChem Fp values and now we will examine the data dimensions X and Y and write a code to remove low vacancy features. Now data split is done in an 80/20 ratio. Now we will write a code for building the regression model using the random forest and finally, we will plot a graph between the experimental and predicted PIC 50 values.



Scatter plot between experimented vs predicted pIC50 value

The scatter plot between experimental and predicted pIC50 will allow us to visually see the correlation between the 2 variables. Ideally, a perfect prediction with 100% Accuracy would show all data points to fall on the trend line. In a practical setting, the residuals or errors (taking the predicted values and subtracting it from the experimental or actual values give us the residual or error values) would cause data points to fall either above or below the trend line (essentially the variance).

IV. CONCLUSION

It is without a doubt that the QSAR paradigm boasts many benefits for the rationa5.h of robust compounds.
Nevertheless, there are certain shortcomings that may limit the widespread application of QSAR.

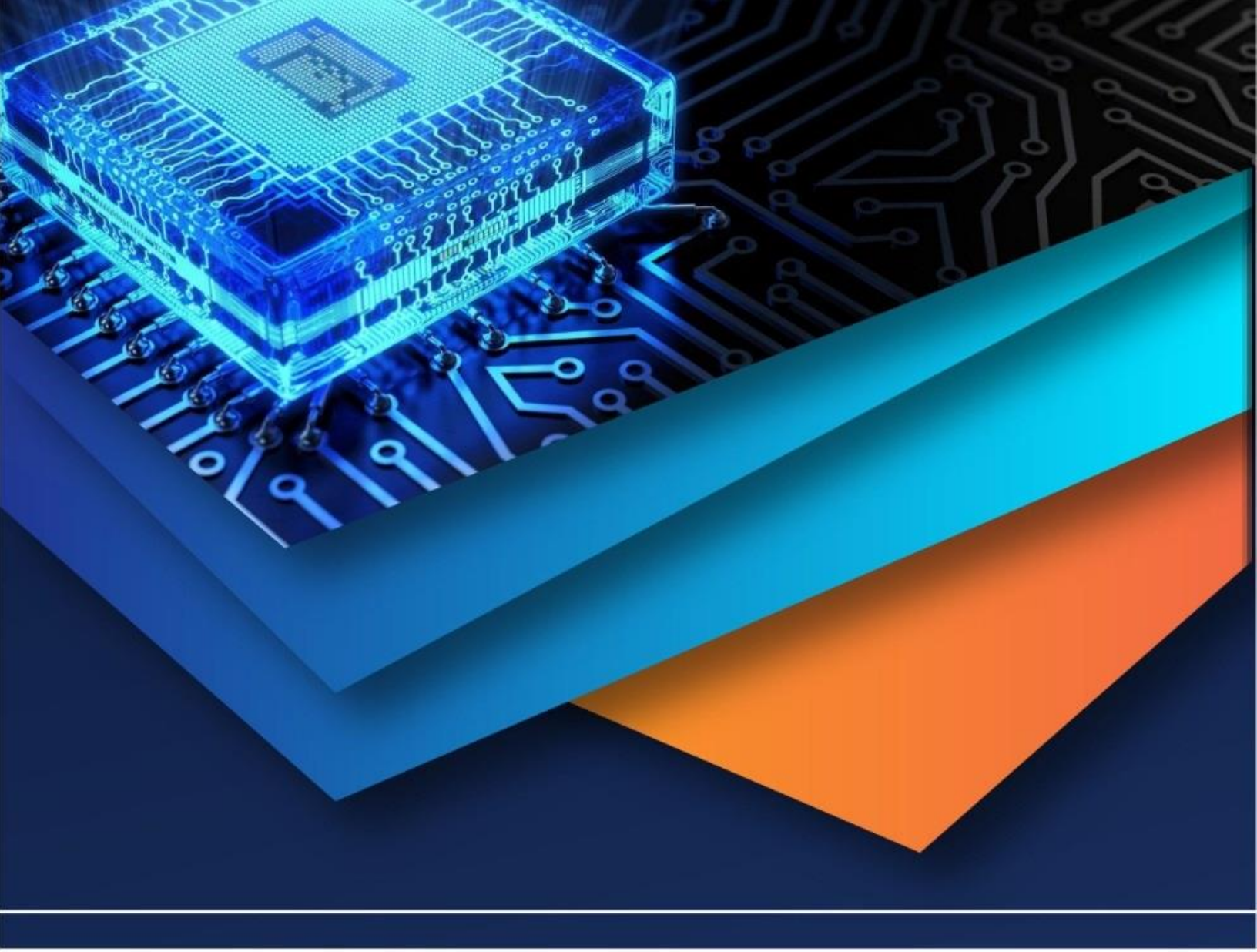
The workflow of QSAR model development.

1) The high dimensionality of the input space.
2) Representation of the molecular structure.
3) Developed QSAR models meaning and interpretability.
4) Presence of outliers or activity cliffs.
5) Validation of QSAR model performance.
6) Applicability in a real-world setting.

The QSAR paradigms, In spite of certain inherent flaws, inevitably one of the most useful forces contributing to the rapid design and development of drug discovery.
Compared to all technologies, QSAR is not perfect: however, the weaknesses and flaws are continuously being identified, solved, and reformed to help shape a new improved and robust approach that is approaching minimal predictive error.
The regression model will allow us to predict the pIC50 value which is the degree at which a molecule can inhibit or not inhibit the target protein of interest (if it can inhibit with potently then it can be a good drug candidate. Afterwards, they will need to be subjected to further scrutiny such as their pharmacokinetic profiles (ADMET properties encompassing the properties of molecules pertaining to Absorption, Distribution, Metabolism, Excretion and Toxicity).

REFERENCES

[1] Cohen ML. Changing patterns of infectious disease. Nature. 2000;406:762– 767. [PubMed] [Google Scholar]
[2] Walsh C. Where will new antibiotics come from? Nat Rev Micro. 2003;1:65–70. [PubMed] [Google Scholar]
[3] Schneider G, Fechner U. Computer-based de novo design of drug-like molecules. Nat Rev Drug Discov. 2005;4:649–663. [PubMed] [Google Scholar]
[4] Yu W, Guvench O, MacKerell AD. Computational approaches for the design of protein– protein interaction inhibitors. In: Zinzalla G, editor. Understanding and exploiting protein– protein interactions as drug targets. London, UK: Future Science Ltd; 2013. pp. 99–102. [Google Scholar]
[5] Panecka J, Mura C, Trylska J. Interplay of the Bacterial Ribosomal A-Site, S12 Protein Mutations and Paromomycin Binding: A Molecular Dynamics Study. PLoS ONE. 2014;9:e111811. [PMC free article] [PubMed] [Google Scholar]
[6] Resat H, Mezei M. Grand Canonical Monte Carlo Simulation of Water Positions in Crystal Hydrates. J Am Chem Soc. 1994;116:7451–7452. [Google Scholar]
[7] Deng Y, Roux B. Computation of binding free energy with molecular dynamics and grand canonical Monte Carlo simulations. J Chem Phys. 2008;128:115103. [PubMed] [Google Scholar]
[8] Small MC, Lopes P, Andrade RB, MacKerell AD., Jr Impact of Ribosomal Modification on the Binding of the Antibiotic Telithromycin Using a Combined Grand Canonical Monte Carlo/Molecular Dynamics Simulation Approach. PLoS Comput Biol. 2013;9:e1003113. [PMC free article] [PubMed] [Google Scholar]
[9] Bento A.P., Gaulton A., Hersey A., Bellis L.J., Chambers J., Davies M., Kruger F.A., Light Y., Mak L., McGlinchey S., et al. The ChEMBL bioactivity database: an update. Nucleic Acids Res. 2014;42:D1083–D1090. [PMC free article] [PubMed] [Google Scholar]
[10] Gaulton A., Bellis L.J., Bento A.P., Chambers J., Davies M., Hersey A., Light Y., McGlinchey S., Michalovich D., Al-Lazikani B., et al. ChEMBL: a large-scale bioactivity database for drug discovery. Nucleic Acids Res. 2012;40:D1100–D1107. [PMC free article] [PubMed] [Google Scholar]
[11] Arrowsmith J., Miller P. Trial watch: phase II and phase III attrition rates 2011-2012. Nat. Rev. Drug Discov. 2013;12:569. [PubMed] [Google Scholar]
[12] Bunnage M.E. Getting pharmaceutical R&D back on target. Nat. Chem. Biol. 2011;7:335– 339. [PubMed] [Google Scholar]
[13] Hay M., Thomas D.W., Craighead J.L., Economides C., Rosenthal J. Clinical development success rates for investigational drugs. Nat. Biotechnol. 2014;32:40–51. [PubMed] [Google Scholar]
[14] Kola I., Landis J. Can the pharmaceutical industry reduce attrition rates? Nat. Rev. Drug Discov. 2004;3:711–715. [PubMed] [Google Scholar]
[15] Cook D., Brown D., Alexander R., March R., Morgan P., Satterthwaite G., Pangalos M.N. Lessons learned from the fate of AstraZeneca's drug pipeline: a five-dimensional framework. Nat. Rev. Drug Discov. 2014;13:419–431. [PubMed] [Google Scholar]
[16] Waring M.J., Arrowsmith J., Leach A.R., Leeson P.D., Mandrell S., Owen R.M., Pairaudeau G., Pennie W.D., Pickett S.D., Wang J., et al. An analysis of the attrition of drug candidates from four major pharmaceutical companies. Nat. Rev. Drug Discov. 2015;14:475– 486. [PubMed] [Google Scholar]
[17] Morgan P., Van Der Graaf P.H., Arrowsmith J., Feltner D.E., Drummond K.S., Wegner C.D., Street S.D. Can the flow of medicines be improved? Fundamental pharmacokinetic and pharmacological principles toward improving Phase II survival. Drug Discov. Today. 2012;17:419–424. [PubMed] [Google Scholar]
[18] Papadatos G., Davies M., Dedman N., Chambers J., Gaulton A., Siddle J., Koks R., Irvine S.A., Pettersson J., Goncharoff N., et al. SureChEMBL: a large-scale, chemically annotated patent document database. Nucleic Acids Res. 2016;44:D1220–D1228. [PMC free article] [PubMed] [Google Scholar]
[19] Gilson M.K., Liu T., Baitaluk M., Nicola G., Hwang L., Chong J. BindingDB in 2015: A public database for medicinal chemistry, computational chemistry and systems pharmacology. Nucleic Acids Res. 2016;44:D1045–D1053. [PMC free article] [PubMed] [Google Scholar]



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY